



**Race to the Top Assessment Program
Written Input
Submitted November 30 to December 2, 2009**



Author	Title	Date Submitted
Eugene E. Garcia	Assessment of English Language Learners in Education Settings	11/30/2009
Barry Topol	Statement for Race to the Top Assessment Program Cost Affordability of Future Assessments	11/30/2009
Christopher Camacho	Race to the Top: General Assessment Public Meeting, Denver, Colorado	12/1/2009
Lindy Crawford	Race to the Top Assessment Program: General Assessment	12/1/2009
John D. Forester	Race to the Top Assessment Input	12/1/2009
Matt Gianneschi	Written Transcript of Testimony Provided by Dr. Matt Gianneschi, Senior Policy Analyst for Education for Bill Ritter, Jr., Governor of Colorado, for the Race to the Top Assessment Program Public and Expert Meeting in Denver, Colorado, December 1, 2008	12/1/2009
Stuart Kahl	Feedback Regarding the Race to the Top Assessment Program	12/1/2009
Clifford W. Lazar	Fair Educational Assessment in 24 Days	12/1/2009
Jody Papini	Talking Points for Jody Papini, Douglas County Federation, CO, On Behalf of the American Federation of Teachers, To the U.S. Department of Education, Dec. 1, 2009	12/1/2009
Julie Woestehoff	Assessment and Accountability under NCLB	12/1/2009
Jim Ysseldyke	Denver-General Assessment Meeting, Tuesday, December 1, 2009, Public Speaker Testimony	12/1/2009
Susan Zelman	Role of Public Service Media in a National Assessment System	12/1/2009
Damian W. Betebenner and Robert Linn	Growth in Student Achievement: Issues of Measurement, Longitudinal Analyses & Accountability	12/2/2009
Henry Braun	Issues in Measuring Student and Conducting Productivity Analyses;	12/2/2009
Kay Brilliant	NEA Response to RTTT Assessment Program	12/2/2009
Christopher Camacho	Computer Dynamic Assessment for Early Childhood	12/2/2009
Christopher Camacho	Additional Information for "Computer-Dynamic Assessment for Early Childhood" Statement	12/2/2009
Wayne Camara and Kevin Sweeney	Comments on the Race to the Top Assessment Program	12/2/2009

Julie Conde	No Title	12/2/2009
Linda Darling-Hammond	Developing Assessment Systems that Support High-Quality Learning	12/2/2009
Jay Diskey and Alan J. Thiemann	Comments by the Association of Test Publishers and the Association of American Publishers	12/2/2009
Richard Dobbs	Measured Progress Race to the Top Assessment Program Input	12/2/2009
Barbara Flores	Race to the Top Assessment Program Notice of Public Meetings and Request for Input	12/2/2009
Ellen Forte	Comments: Race to the Top Assessment Pubic and Expert Input Meeting: ELL Assessment	12/2/2009
Catalina Fortino	Talking Points: Catalina Fortino, United Federation of Teachers Teacher Center Staff, New York, NY, On Behalf of the American Federation of Teachers, To the U.S. Department of Education, Dec. 2, 2009	12/2/2009
Nancy S. Grasmick	State Department of Education Input Regarding Race to the Top Assessment Program	12/2/2009
Nancy Green	Race to the Top Assessment Program	12/2/2009
Valerie Greenhill	Race to the Top Assessment Program Comments	12/2/2009
Edward H. Haertel	Student Growth Data for Productivity Indicator Systems	12/2/2009
Ellen Haley	No Title and Assessment in the 21st Century: Preparing Students for the Global Workplace	12/2/2009
Margaret Heritage	Assessment for Teaching and Learning	12/2/2009
Terry Holliday	Race to the Top Assessment - Written Input, Kentucky Department of Education, Commissioner Terry Holliday, December 2, 2009	12/2/2009
Laura Kaloi	Consortium for Citizens with Disabilities Input on Assessment of Students with Disabilities	12/2/2009
Daniel Koretz	Implications of Current Policy for Educational Measurement	12/2/2009
Sheri Krause	No Title	12/2/2009
Magaly Lavadenz	Race to the Top Assessment Program Notice of Public Meetings and Request for Input	12/2/2009
Robert Linn	Comments on Presentation on Implications for Policy by Dan Koretz	12/2/2009
Scott Marion	A Comprehensive Assessment System: Tough Choices for the RTTT Assessment Program	12/2/2009
Luis-Gustavo Martinez	Race to the Top Assessment Program	12/2/2009
Karen Mulattieri	ELL Trend Data and Race To The Top-Comments	12/2/2009
Anna Nicotera	Race to the Top Assessment Input	12/2/2009
John H. Oswald	Educational Testing Service Response to Request for Input on the Race to the Top Assessment Program	12/2/2009

Lisette Partelow	Race to the Top Fund Assessmet Program	12/2/2009
Stephanie Petska	Input on Assessment of Students with Disabilities	12/2/2009
Michael A. Resnick	National School Boards Association Response to Notice of Public Meetings and Request for Input to Gather Technical Expertise Pertaining to a Possible Race to the Top Program; published in the Federal Register on October 23, 2009	12/2/2009
Kristina Robertson	No Title	12/2/2009
EJ Rodriguez	Assessment of ELLs: Assessment and Instruction Design for Our Clients	12/2/2009
Michael Russell	Race to the Top Written Comment	12/2/2009
Ricki Sabia	Universal Design for Learning and the Race to the Top Assessment Program	12/2/2009
Ricki Sabia	Race to the Top Assessment Program Comments	12/2/2009
Edynn Sato	Race to the Top Assessment Program: Assessment of English Language Learners	12/2/2009
Jerome Shaw	Testimony to the U.S. Department of Education Race to the Top Assessment Program	12/2/2009
Teri Siskind	Race to the Top Assessment Input	12/2/2009
Guillermo Solano-Flores	Assessment of English Language Learners	12/2/2009
David Stevenson	Feedback on RTTT Assessment RFP	12/2/2009
Jennifer Thayer	Comments from Wisconsin on Race to the Top Assesment Program	12/2/2009
Marc Tucker	A Design for an American Examination and Testing System	12/2/2009
John S. Twing	No Title	12/2/2009
Randi Weingarten	Recommendations to the U.S. Department of Education's Race to the Top Assessments Program from the AFT	12/2/2009
James Wendorf	Race to the Top Assessment Program	12/2/2009
Joe Willhoft	Assessment of English Language Learners - Denver, CO	12/2/2009
Mark Wilson	Assessment for Learning AND for Accountability	12/2/2009
Marci Young and Kathy Patterson	Pre-K Now Comment on Race to the Top Assessment Program	12/2/2009
Gerald L. Zahorchak	Testimony of Gerald L. Zahorchak	12/2/2009
Deborah A. Zeigler	Race to the Top Assessment Program	12/2/2009
Susan Zeiman	The Role of Public Service Media in a National Assessment System	12/2/2009

Assessment of English Language Learners in Education Settings

(Draft, Invited Testimony for Race to the Top, December 2, 2009)

Eugene E. Garcia, Ph.D.

Arizona State University

The Challenge of Assessing ELLs in Education Settings

The increasing demand for evaluation, assessment, and accountability at all levels of education comes at a time when the fast growing student population in the country is children whose home language is not English. This presents several challenges to practitioners and school systems generally who may be unfamiliar with important concepts such as multilingual development, second language acquisition, acculturation, and the role of socioeconomic background as they relate to test development, administration, and interpretation. Because assessment is critical in developing and implementing effective curricular and instructional strategies that promote student learning, English language learner (ELL) children have the right to be assessed. Through individual assessments, teachers can personalize instruction, make adjustments to classroom activities, assign children to appropriate program placements, and have more informed communication with parents. And systems need to know how ELLs are performing in order to make proper adjustments and policy changes. However, there is a lack of adequate instruments to use with ELLs, especially considering the hundreds of languages represented in the United States. Some tests exist in Spanish, but most lack the technical qualities of a high-quality assessment tools. Additionally, there is a shortage of bilingual professionals with the skills necessary to evaluate these children, and conceptual and empirical work systematically linking context with student learning. The intent of this testimony is to deal with these challenges/practices, and to review important principles associated with high-quality assessments for ELLs. **This testimony attempts to sound a very critical tone for the use of any “high stakes” assessment in the Race to the Top efforts, but, instead recommends using this effort to develop,**

enhance and expand needed reliable and valid assessments and systems of assessment for this important population of US students that are aligned with the purposes of the assessments.

English Language Learners: Who Are They?

Assessing the development of ELLs demands an understanding of who these children are in terms of their linguistic and cognitive development, as well as the social and cultural contexts in which they are raised. The key distinguishing feature of these children is their non-English language background. In addition to linguistic background, other important attributes of ELL children include their ethnic, immigrant, and socioeconomic histories (Abedi, Hofstetter, & Lord, 2004; Capps et al., 2005; Figueroa & Hernandez, 2000; Hernandez, 2006). Though diverse in their origins, ELL students, on average, are more likely than their native English-speaking peers to have an immigrant parent, to live in low-income families, and to be raised in cultural contexts that do not reflect mainstream norms in the US (Capps et al., 2005; Hernandez, 2006).

English language learners represent diverse ethnic backgrounds. In the 2000-2001 school year, approximately four in five ELLs were from Spanish-speaking homes, followed by Vietnamese (2%), Hmong (1.6%), Cantonese (1%), Korean (1%), and many more native and foreign languages. While a majority of Hispanic ELLs are of Mexican origin (approximately 7 in 10), substantial proportions have origins in Puerto Rico, Central America, South America, Cuba, and the Dominican Republic (Hernandez, 2006). Within and among these groups, ELL children represent diverse social and cultural customs and histories, which are essential to consider thoroughly when assessing the

child's linguistic, cognitive, social, and emotional development within home and school contexts.

Finally, it is important to consider the socioeconomic status of English language learners, including family income as well as the amount of educational capital (i.e., parental education) in the home. In 2000, 68 percent of ELLs in PK to grade 5 were in low-income families (defined as family income below 185 percent of the federal poverty level), compared to 36 percent of English proficient children in the same grades (Capps et al., 2005). Moreover, nearly half of ELL children in elementary school had parents with less than high school educations in 2000, compared to 9 percent of parents of English proficient children. A quarter of ELL elementary school students had parents less than 9th grade educations, compared to 2 percent of parents of English proficient students (Capps et al., 2005). Parent education levels are important indices as they influence language and educational practices in the home, and, therefore, the development of skills valued in US schools.

Assessment Issues

ELLs have the right to benefit from the potential advantages of assessment. The current empirical knowledge-base and the legal and ethical standards are limited yet sufficient to improve ways in which ELLs are assessed. Improvements will require commitments from policymakers and practitioners to implement appropriate assessment tools and procedures, to link assessment results to improved practices, and to utilize trained staff capable of carrying out these tasks. This is the substantive challenge in Race to the Top efforts. Assessments of contextual processes will be necessary if current

assessment strategies, which largely focus on the individual, are to improve classroom instruction, curricular content, and, therefore, student learning (Rueda, 2007; Rueda & Yaden, 2006).

Purpose of Assessment

Sensing an increase in demands for greater accountability and enhanced educational performance of children, the National Education Goals Panel developed a list of principles to guide early educators through appropriate and scientifically-sound assessment practices (Shepard, Kagan, & Wurtz, 1998). Moreover, the panel presented four purposes for assessing children. Pertinent as well to the assessment of ELL children, the purposes were a) to promote children's learning and development, b) to identify children for health and special services, c) to monitor trends and evaluate programs and services, and d) to assess academic achievement to hold individual students, teachers, and schools accountable (i.e., high stakes testing) (Shepard, Kagan, & Wurtz, 1998). Embedded within each of these purposes are important considerations for practice so as to preserve assessment accuracy and support interpretations of results that lead to increased educational opportunity for the student. The foundation for educational assessment set for by this effort as paramount for ELL student assessment.

Legal and ethical precedent

The impetus for appropriate and responsive assessment practices of ELLs is supported by a number of legal requirements and ethical guidelines, which have developed over time. Case law, public law, and ethical codes from professional

organizations support the use of sound assessment tools, practices, and test interpretations. A widely cited set of testing standards are found in a recent publication from the American Psychological Association (APA), the American Educational Research Association (AERA), and National Council on Measurement in Education (NCME) entitled *Standards for Educational and Psychological Testing* (1999). Revised from the 1985 version, in its fourth edition, this volume offers a number of ethical standards for assessing the psychological and educational development of children in schools, including guidelines on test development and application. Included is a chapter on testing children from diverse linguistic backgrounds, which discusses the irrelevance of many psychoeducational tests developed for and normed with monolingual, English-speaking children. Caution is given to parties involved in translating such tests without evaluating construct and content validity and developing norms with new and relevant samples. It also discusses accommodation recommendations, linguistic and cultural factors important in testing, and important attributes of the tester. Similar, though less detailed provisions exist in the *Professional Conduct Manual* published by the National Association of School Psychologists (2000).

It has been argued that the standards presented by APA, AERA and NCME have outpaced present policy, practice, and test development (Figueroa & Hernandez, 2000). However, the federal Individuals with Disabilities Education Act (IDEA 2004) does provide particular requirements related to the assessment of ELLs. It requires, for example, the involvement of parents/guardians in the assessment process as well as a consideration of the child's native language in assessment. Unlike ethical guidelines, which often represent professional aspirations and are not necessarily enforceable, public

law requires compliance. The Office of Civil Right (OCR) is given the charge to evaluate compliance to federal law and, where necessary, audit public programs engaged in assessment practices and interpretations of ELLs and other minority children.

Assessment practice: use and misuse

In addition to the concerns that afflict the assessment of all children, there are central issues inherent in the assessment of children from non-English language backgrounds. Implementation research suggests that assessment practices with ELLs continue to lag behind established legal requirement and ethical standards set forth by APA, AERA and NCME. In part, this is because of a lack of available instruments normed on representative samples of English language learners, because of inadequate professional development and training, and partly because of insufficient research to inform best practice.

The academic achievement (or performance) of ELLs in Race to the Top may be assessed for several reasons. Assessments for accountability purposes tend to rely on criterion-references tests developed by state departments of education (Abedi, Hofstetter, & Lord, 2004; Abedi, Lord, Hofstetter, & Baker, 2000; Hakuta & Beatty, 2000). Debates have continued over the past decades regarding the inclusion if ELLs in large-scale student assessment programs. Due to antidiscrimination laws, court cases, and standards-based legislation, there has been a push to include all students in state assessments, including ELLs. This has led to the appropriation of accommodations—changes in the test process, in the test itself, and/or in the test response format—to more accurately portray the performance of ELLs and not discriminate against language background

(Abedi, Hofstetter, & Lord, 2004). Currently, however, decisions about which accommodations to use, for whom, and under what conditions are based on little empirical evidence.

Assessments of academic achievement are also used to improve student learning and for special service identification. For children in early education, these tend to assess early literacy (e.g., sound and letter recognition, sight words) and numeracy (e.g., numbers, shapes, relative size, ordinality) skills. A larger variety of tools and practices are used for these purposes, which can be categorized by two general types of performance assessment. First, commercial (mostly norm-referenced) tests are used. Some of the same concerns with regard to normative cognitive assessment are relevant to normative academic assessment. That is, many of the tests have been developed essentially as back translations or adaptations of existing English language measures, without evaluating their construct and content validity. Moreover, the normative samples often do not reflect the ethnic, socioeconomic, and linguistic backgrounds of ELL students.

Even when these obstacles are overcome, and where bilingual achievement tests have been produced with representative samples, the argument is made that the content of standardized tests does not necessarily predict success in the curriculum. The base case for this argument is that test content often does not reflect classroom content, and that academic outcomes do not inform, per se, instructional and/or curricular interventions. For these reasons, a second option for the achievement assessment to improve student learning and to determine special service identification, known as curriculum-based measurement (CBM), has accumulated evidence and attention over the

past few decades (Fuchs, 2004; Rhodes, Ochoa & Ortiz, 2005). Conceptualized initially as an approach to student progress monitoring (Deno, 1985), CBM tasks are used to assess student performance in the curriculum on a weekly basis. Results are used simultaneously to monitor student progress and to inform instructional and/or curricular interventions. The slope of scores over time is used to monitor progress and the rate of growth toward a determined goal or standard. IDEA 2004 allows CBM approaches to replace traditional testing approaches (i.e., normative testing) of academic achievement to determine special education eligibility for learning disabilities.

Professional development and training. A number of problems arise when school personnel are engaged in the assessment of English language learners without the necessary competence, tools, and, therefore, practices. The literature on disproportional representation of language minority children in special education programs, for example, has pointed to culturally and linguistically unresponsive referral, assessment, and eligibility determination practices in schools as causes of disproportionality (Coutinho, & Oswald, 2000; Rhodes, Ochoa & Ortiz, 2005). Moreover, though the research and legal and ethical declarations mandate responsive practice, several studies have documented referral, assessment, and interpretation practices that are below standard. These studies have highlighted language barriers and low expectations of teachers (McCardle, Mele-McCarthy, & Leos, 2005), questionable intellectual assessment practices (Bainter, & Tollefson, 2003), questionable language assessment practices (Ochoa, Galarza & Amado, 1996; Yzquierdo, Blalock & Torres-Velasquez, 2004), invalid and/or irrelevant interpretations (Harry & Klingler, 2006), and inappropriate translation and interpretation

practices (Hakuta & Beatty, 2000; Ochoa, Gonzalez, Galarza & Guillemard, 1996; Paredes Scribner, 2002; Santos, Lee, Valdivia & Zhang, 2001).

This has several implications for ongoing implementation research in the area professional development and training for assessing ELLs. This research will need to focus on strategies to improve staff competencies necessary to work as a part of a professional team, to work with interpreters, and to choose and administer appropriate assessment batteries. Moreover, implementation research should highlight strategies to train practitioners to develop their competence in second language acquisition, acculturation, and the evaluation of educational interventions.

Principles in the Assessment of ELLs in Early Education Settings, Pre/K-4

Hence, the gap between current *practice* in the assessment of English language learners in the US and the *standards* set forth through research, policy, and ethics is largely a function of the gap between practical and optimal realities. Due to the many demands and constraints placed on teachers and schools from local, state, and federal governments, including budgeting responsibilities and the many programs implemented each school year, it can be extremely challenging to keep pace with best practices and ethical standards. However, given the large and increasing size of the young ELL child population in the US, the current focus on testing and accountability, and the documented deficits in current assessment practices, improvements are critical. These improvements are necessary at all phases of the assessment process, including pre-assessment and assessment planning, conducting the assessment, analyzing interpreting

the results, reporting the results (in written and oral format), and determining eligibility and monitoring.

Researchers and organizational bodies have offered principles for practitioners engaged in the assessment of young ELLs (Clifford et al., 2005). Clifford et al. present seven detailed recommendations “to increase the probability that all young English language learners will have the benefit of appropriate, effective assessment of their learning and development” (p.1). Because these recommendations—presented here as *principles*—materialized as a collaborative effort from a committee comprised of over a dozen researchers in the field, they are quite representative of recommendations found in the literature.

First, *screening and assessment instruments and procedures are used for appropriate purposes*. Screening tools should result in needed supports and services and, if necessary, further assessment. Assessments should be used fundamentally to support learning, including language and academic learning. For evaluation and accountability purposes, young ELLs should be included in assessments and provided with appropriate tests and accommodations.

Second, *screenings and assessments should be linguistically and culturally appropriate*. This means assessment tools and procedures should be aligned with cultural and linguistic characteristics of the child. When tests are translated from its original language to that of the native language of the ELL child, they should be culturally and linguistically validated to verify the relevance of the content (i.e., content validity) and the construct purported to be measured (i.e., construct validity). Moreover, in the case of

normed-based tests, the characteristics of children included in the normative sample should reflect the linguistic, ethnic, and socioeconomic characteristics of the child.

Third, *the primary purpose of assessment should be to improve instruction*. The assessment of student outcomes using appropriate tools and procedures should be linked closely to classroom processes. This means relying on multiple methods and measures, evaluating outcomes over time, and using collaborative assessment teams, including the teacher, who is a critical agent for improved learning and development. Assessment that systematically informs improved curriculum and instruction is the most useful.

Fourth, *caution ought to be used when developing and interpreting standardized formal assessments*. As discussed, standardized assessments are used for at least three purposes—to identify disabilities and determine program eligibility, to monitor and improve learning, and for accountability purposes. It is important young ELLs are included in large-scale assessments, and that these instruments continue to be used to improve educational practices and placements. However, those administering and interpreting these tests ought to use caution. Test development issues—including equivalence, translation, and norming—must be scrutinized, and evidence-based accommodations ought to be provided during accountability assessments.

Fifth, *those administering assessments should have cultural and linguistic competence*. This may be the most challenging of the recommendations. Professional development and training of teachers, school psychologists, speech pathologists, and school administrators constitutes a long-term goal which will demand ongoing funding and implementation research. Those assessing young ELLs should be bicultural, bilingual, and be knowledgeable about second language acquisition. In many cases,

consultants and interpreters are used where the supply of school personnel possessing these qualifications is limited. Implementation research is needed to understand best practices in working with consultants and interpreters through the pre-assessment and assessment planning, conducting the assessment, analyzing interpreting the results, reporting the results (in written and oral format), and determining eligibility and monitoring.

Finally, *families should play critical roles in the assessment process*. Under federal law, parents have the right to be included in the decision making process regarding the educational placement for their child. Moreover, the educational benefit of the assessment process for a given child is optimal when parents' wishes are voiced and considered throughout. Although family members should not administer formal assessments, they are encouraged to be involved in selecting, conducting, and interpreting assessments. The process and results of assessment should be explained to parents in a way that is meaningful and easily understandable.

Directions for Practice within the Context of Race to the Top

As mentioned, there is a gap between current assessment practice of ELLs and what the research and the legal and ethical standards suggest is best practice. It is important, therefore, that new practices are developed to improve this scenario.

First, the field needs more assessments developed and normed especially for young English language learners. This will require a bottom-up approach, meaning assessment tools, procedures, and factor analytic structures are aligned with cultural and linguistic characteristics of ELL children, as opposed to top-down approaches where, for

example, assessment tools and practices are simply translated from their original language to the native languages of ELLs. Assessments must also take into account important characteristics of the child, including their linguistic, ethnic, and socioeconomic histories.

Second, it is time conceptual and empirical work on student assessment move beyond the individual level. That is, the majority of the discussion in this testimony reflects the extent literature which has focused heavily on the assessment of processes and outcomes within the individual—assessing language, cognitive development, academic learning, and so forth. With this knowledge-base teachers and schools are expected to adjust aspects of the environment to improve learning. I has become clear that processes outside the individual—including within the classroom (e.g., teacher-student interactions, peer to peer interactions), the home (e.g., frequency of words spoken, amount of books), and within the school (e.g., language instruction policies)—affect learning, the field presently lacks conceptual frameworks and the measures necessary to move this research forward to systematically improve student learning. Preliminary research on the role of context in learning suggests that variations environmental factors can increase student engagement and participation (Christenson, 2004; Goldenberg, Rueda, & August, 2006), which, in turn can lead to increased learning—and that the influence of contextual contingencies on learning outcomes is mediated by children’s motivation to learn (Rueda, 2007; Rueda, MacGillivray, Monzó & Arzubiaga, 2001; Rueda & Yaden, 2006). Conceptual frameworks should account for the multilevel nature of contexts, including the nesting of individuals within classrooms and families, classrooms within schools, and schools within school districts,

communities, and institutions. Moreover, the role of culture and the feasibility of cultural congruence across within- and out-of-school contexts will be important to this work.

Meaningful empirical work in this area will require the convergence of research methods (e.g., multi-level statistics and the mixing of qualitative approaches with quasi-experimental designs) and social science disciplines (e.g., cognitive psychology, educational anthropology, sociology of education).

Finally, more efforts documenting the current scenario of the assessment of young ELLs across the country is needed. As the population of ELLs continues to grow and disperse to states with historically low representations of ELL students, more work is needed to evaluate assessment practices in their localities. Observational approaches will be needed to document practices in pre-assessment and assessment planning, conducting the assessment, analyzing interpreting the results, reporting the results (in written and oral format), and determining eligibility and monitoring. This work will aid the development of strategies to train professionals with the skills necessary to serve young ELL children.

Selected References

Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English-language learners: Implications for policy-based empirical research. *Review of Educational Research, 74*(1), 1–28.

Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice, 19*(3), 16-26.

AERA (American Educational Research Association), APA (American Psychological Association), & NCME (National Council on Measurement in Education). (1999). *Testing and assessment: The standards for educational and psychological testing*. Washington, DC: AERA. Online: www.apa.org/science/standards.html.

August, D. (2006). Demographic overview. In D. August & T. Shanahan (Eds.), *Report of the national literacy panel on language minority youth and children*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bainter, T. R., & Tollefson, N. (2003). Intellectual assessment of language minority students: What do school psychologists believe are acceptable practices? *Psychology in the Schools, 40*(6), 899-903.

Borghese, P., & Gronau, R. C. (2005). Convergent and discriminant validity of the Universal Nonverbal Intelligence Test with limited English proficient Mexican-American elementary students. *Journal of Psychoeducational Assessment, 23*, 128-139.

Bracken, B., & McCallum, R. S. (1998). *The Universal Nonverbal Intelligence Test*. Chicago, IL: Riverside.

Bradley-Johnson, S. (2001). Cognitive assessment for the youngest children: A critical review of tests. *Journal of Psychoeducational Assessment, 19*, 19-44.

Capps, R., Fix, M., Murray, J., Ost, J., Passel, J. S., & Herwantoro, S. (2005). *The new demography of America's schools: Immigration and the No Child Left Behind Act*. Washington, DC: The Urban Institute.

Carter, A. S., Briggs-Gowan, M. J., Ornstein Davis, N. (2004). Assessment of young children's social-emotional development and psychopathology: Recent advances and recommendations for practice. *Journal of Child Psychology and Psychiatry, 45*(1), 109-134.

- Christenson, S. L. (2004). The family-school partnership: An opportunity to promote learnign and competence of all students. *School Psychology Review*, 33(1), 83-104.
- Clifford, D. et al. (2005). *Screening and assessment of young English-language learners*. Washington, DC: National Association for the Education of Young Children. Available online at http://www.naeyc.org/about/positions/ELL_Supplement.asp
- Coutinho, M. J., & Oswald, D. P. (2000). Disproportionate representation in special education : A synthesis and recommendations. *Journal of Child and Family Studies*, 9, 135-156.
- De Avila, E. & Duncan, S. (1990). *Language assessment scales—oral*. Monterrey, CA: CTB McGraw-Hill.
- Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, 52, 219-232.
- Duncan, S., & De Avila, E. (1988). *Language assessment scales—reading and writing*. Monterrey, CA: CTB McGraw-Hill.
- Duncan, S. E., & DeAvila, E. (1998). *Pre-language assessment scale 2000*. Monterey, CA: CTB/McGraw-Hill.
- Dunn. L. M., & Dunn, L. M. (1981). *Peabody picture vocabulary test*. Circle Pines, MN: American Guidance Service.
- Dunn, L. M., Padilla, E. R., Lugo, D. E., & Dunn L. M. (1986). *Test de vocabulario en imágenes Peabody*. Circle Pines, MN: American Guidance Service.
- Figuroa, R. A., & Hernandez, S. (2000). *Testing Hispanic students in the United States: Technical and policy issues*. Washington, DC: President’s Advisory Commission on Educational Excellence for Hispanic Americans
- Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review*, 33(2), 188-192.
- Fugate, M. H. (1993). Review of the Bayley Scales of Infant Development, Second Edition. *Mental Measurements Yearbook*, 13.
- García, E. E. (2005). *Teaching and learning in two languages: Bilingualism and schooling in the United States*. New York: Teachers College Press
- Garcia, G. E., McKoon, G., & August, D. (2006). Synthesis: Language and literacy assessment. In D. August & T. Shanahan (Eds.), *Developing literacy in second language learners*. Mahwah, NJ: Lawrence Erlbaum Associates.

Genesee, F., Geva, E., Dressler, C., & Kamil, M. (2006). Synthesis: Cross-linguistic relationships. In D. August & T. Shanahan (Eds.), *Report of the national literacy panel on language minority youth and children*. Mahwah, NJ: Lawrence Erlbaum Associates.

Goldenberg, C., Rueda, R., & August, D. (2006). Synthesis: Sociocultural contexts and literacy development. In D. August & T. Shanahan (Eds.), *Report of the national literacy panel on language minority youth and children*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hakuta, K. & Beatty, A. (2000). *Testing English language learners in US schools*. Washington, DC: National Academy Press.

Harry, B. & Klingler, J. (2006). *Why are so many minority students in special education? Understanding race and disability in schools*. New York: Teachers College Press.

Hernandez, D. (2006). *Young Hispanic children in the US: A demographics portrait based on Census 2000*. Report to the National Task Force on Early Childhood Education for Hispanics. Tempe, AZ: Arizona State University.

McCardle, P., Mele-McCarthy, J., & Leos, K. (2005). English language learners and learning disabilities: Research agenda and implications for practice. *Learning Disabilities Research & Practice*, 20(1), 69-78.

National Association of School Psychologists. (2000). *Professional conduct manual*. Bethesda, MD: Author.

National Clearinghouse for English Language Acquisition (2006). *The growing numbers of limited English proficient students: 1993-94-2003/04*. Office of English Language Acquisition (OELA): US Department of Education.

Ochoa, S. H., Galarza, S. & Amado, A. (1996). An investigation of school psychologists' assessment practices of language proficiency with bilingual and limited-English-proficient students. *Diagnostique*, 21(4), 17-36.

Ochoa, S. H., Gonzalez, D., Galarza, A., & Guillemard, L. (1996). The training and use of interpreters in bilingual psychoeducational assessment: An alternative in need of study. *Diagnostique*, 21(3), 19-40.

Paredes Scribner, A. (2002). Best assessment and intervention practices with second language learners. In A. Thomas & J. Grimes (Eds.), *Best Practices in School Psychology IV*. Bethesda, MD: National Association of School Psychologists.

Reynolds, C. R., & Kamphaus, R. W. (2003). *Behavior Assessment System for Children, Second Edition*. Minneapolis, MN: Pearson.

- Rhodes, R., Ochoa, S. H., & Ortiz, S. (2005). *Assessing culturally and linguistically diverse students: A practical guide*. New York: Guilford.
- Rueda, R., MacGillivray, L., Monzó, L., & Arzubiaga, A. (2001). Engaged reading: A multi-level approach to considering sociocultural features with diverse learners. In D. McNerny & S. Van Etten (Eds.), *Research on sociocultural influences on motivation and learning* (pp. 233-264). Greenwich, CT: Information Age.
- Rueda, R. (2007). *Motivation, learning, and assessment of English learners*. Presented at the School of Education, California State University Northridge, Northridge, CA, April.
- Rueda, R., & Yaden, D. (2006). The literacy education of linguistically and culturally diverse young children: An overview of outcomes, assessment, and large-scale interventions. In B. Spodek & O.N. Saracho (Eds.), *Handbook of Research on the Education of Young Children, 2nd Ed.* (pp. 167-186). Mahwah, NJ: Lawrence Erlbaum Assoc., Pub.
- Santos, R.M., S. Lee, R. Valdivia, & C. Zhang. 2001. Translating translations: Selecting and using translated early childhood materials. *Teaching Exceptional Children* 34 (2): 26–31.
- Shepard, L., Kagan, S. L. & Wurtz, L (Eds.) (1998). *Principles and recommendations for early childhood assessments*. Goal 1 Early Childhood Assessments Resource Group. Washington, DC: National Education Goals Panel. Retrieved online at http://eric.ed.gov/ERICDocs/data/ericdocs2/content_storage_01/0000000b/80/24/51/e6.pdf
- Wechsler, D. (2003). *The Wechsler Intelligence Scale for Children*, 4th ed. San Antonio, TX: Psychological Corporation.
- Wechsler, D. (2004). *The Wechsler Intelligence Scale for Children—Spanish*, 4th ed. San Antonio, TX: Psychological Corporation.
- Yzquierdo, Z., Blalock, G., & Torres-Velasquez, D. (2004). Language-appropriate assessments for determining eligibility of English language learners for special education services. *Assessment for Effective Intervention*, 29(2), 17-30.

Additional Related Readings

Alvarado, C. G. (1999). *A Broad Cognitive Ability—Bilingual Scale for the WJ-R Tests of Cognitive Ability and the Batería Woodcock-Muñoz Pruebas de Habilidad Cognitiva—Revisada* (Research Report Number 2). Itasca, IL: Riverside.

Artiles, A. J. (1998). The dilemma of difference: Enriching the disproportionality discourse with theory and context. *Journal of Special Education, 32*, 32-36.

Butler, F. A., & Stevens, R. (2001). Standardized assessment of the content knowledge of English language learners in K-12: Current trends and old dilemmas. *Language Testing, 18*(4), 409-427.

CLAS (Culturally and Linguistically Appropriate Services) Early Childhood Research Institute. 2000. *Review guidelines for material selection: Child assessment*.
Online: <http://clas.uiuc.edu/review/RG-ChildAssessment.html>

Chun, K. M., Organista, P. B., & Marín, G. (2003). *Acculturation: Advances in theory, measurement, and applied research*. Washington, DC: American Psychological Association.

Collins, R., & R. Ribeiro. 2004. Toward an early care and education agenda for Hispanic children. *Early Childhood Research and Practice 6* (2).
Online: <http://ecrp.uiuc.edu/v6n2/collins.html>

Espinosa, L. 2005 Curriculum and assessment considerations for young children from culturally, linguistically, and economically diverse backgrounds. Special issue, *Psychology in the Schools, 42*(8), 837-853.

Fives, C. J., Flanagan, R. (2002). A review of the Universal Nonverbal Intelligence Test (UNIT): An advance for evaluating youngsters with diverse needs. *School Psychology International, 23*(4), 425-448.

Gonzalez, V., P. Bauerle, & M. Felix-Holt. 1996. Theoretical and practical implications of assessing cognitive and language development in bilingual children with qualitative methods. *The Bilingual Research Journal* 20 (1): 93–131.

Lopez, E. C. (2002). Best practices in working with school interpreters to deliver psychological services to children and families. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV*, (pp. 1419-1432). Washington, DC: National Association of School Psychologists.

McLean, M. (2002). Assessing young children for whom English is a second language. In *Young Exceptional Children Monograph Series*, no. 4. *Assessment: Gathering meaningful information*, 73–82. Longmont, CO: Sopris West.

Ortiz, S. (2002). Best practices in nondiscriminatory assessment. In A. Thomas & J. Grimes (Eds.), *Best Practices in School Psychology IV*. Bethesda, MD: National Association of School Psychologists.

Raven, J. C. (1995). *Raven's coloured progressive matrices*. San Antonio, TX: Psychological Corporation.

Stevens, R. A., Butler, F. A., & Castellón-Wellington, M. (2000). *Academic language and content assessment: Measuring the progress and English-language learners* (CSE Technical Report No. 552). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Valdés, G., & Figueroa, R. (1994). *Bilingualism and testing: A special case of bias*. Norwood, NJ: Albex.



Statement for Race to the Top Assessment Program

Cost Affordability of Future Assessments

Barry Topol, CPA

Managing Partner, Assessment Solutions Group – December 2, 2009

General Assessments – Cost and Implementation Issues

The current slate of initiatives in progress to reform and upgrade the educational system in the United States represents a generational opportunity to improve public education in this country. The Race to the Top Assessment program and development of common core standards are two of the more important initiatives in this effort. It is critical, therefore, that these initiatives are formulated and implemented in a way that maximizes their chances for successful adoption by states and educators. Doing so will require policy makers to keep several things in mind throughout the process. Some are obvious (and are being done) and relatively easy to accomplish such as maintaining transparency throughout the process and involving constituents in both the formulation and implementation aspects of policies. Others such as designing the most effective and efficient assessment systems, determining the proper type of technology initiatives to implement, and ensuring that proper funding is available to sustain new initiatives into the future are more difficult. We will briefly address the latter areas in this paper.

We believe that new assessment systems should be balanced in their nature and include not only summative tests but interim/benchmark and formative assessments as well. A more comprehensive and unified assessment system will provide teachers with important information on how students are doing and allow for intervention strategies to be developed for those students needing additional help during the school year. Additional constructed response items and new item types including performance events, performance tasks and, possibly, portfolio and other types of performance exhibitions should be part of a new assessment system. Constructed response items and new item types will enable evaluation of the higher level critical thinking skills required of our students in the 21st century. The development of a new assessment system must be inclusive of all students and address the needs of students with disabilities and English language learners. The assessments need to be universally designed. Accommodations that improve access to the test for both SWD and ELL, as well as alternate assessments for students with disabilities and English language proficiency tests for ELLs that don't yet understand English should be an important part of a new system. New, online, methods for providing accommodations and increasing accessibility to the test material are being developed and will soon be affordable for all states. Finally, any new assessment system must also include

Assessment Solutions Group - "your assessment experts"

215 Loch Lomond Way

Danville, CA 94526

210-859-9920

btopol@assessmentgroup.org

www.assessmentgroup.org



professional development opportunities for teachers, not only in assessment, but as importantly, in the underlying curricula and instructional design and use of materials.

There is probably not a tremendous amount of controversy around the ideas mentioned above. However, we must be mindful of how any new assessment system will be funded ***and be sustainable*** into the future. Past efforts at developing new, common standards and assessments have failed because states did not have the money and staff resources to implement the innovative approaches to assessment on an ongoing basis. With the current financial situation in states likely to persist for several more years, it is *critical* that the costs for the ongoing administration of any new assessment system be no greater than that currently incurred by a state. In fact, with the many budget cuts states are experiencing and still more planned for next year, costs may need to be less than in current state budgets. The current funding issues would indicate that new assessment systems should be developed with the idea, and be accommodative to the fact, that some states may need to implement changes to their current systems over time.

Given the funding constraints, several key questions arise. How does one best develop and implement an improved assessment system under these conditions? What are appropriate costs for developing new assessments based on common standards across a large number of states? How can this work be done most efficiently and at the lowest cost possible, without sacrificing quality? How can effective and efficient assessment services be delivered to states by testing vendors? Given that vendors will “bid” on consortium work more or less “sole source,” what controls will the consortia have to avoid uncompetitive pricing? What will the costs be to states for sustaining the new assessments in future years? How will states know if the ongoing costs will be affordable?

We believe that it is imperative to have a solid understanding of costs for both development and implementation/administration as new assessment system requirements are being developed and proposed. The USED and states must have access to good cost models (not based on NAEP) to understand state assessment features, benefits and costs so that trade-offs can be made and costs evaluated before the development process unfolds. State consortia looking to implement a common assessment must understand all their future ongoing administration and maintenance costs *prior to* submitting a proposal for potential award. Expected costs, per state, versus current costs should be compared and included in any proposal, as well as plans to address any shortfalls. Furthermore, the USED and states must be able to objectively evaluate the cost quotes from vendors to ensure they are competitively priced, based on established benchmarks for fair pricing.

Assessment costs will rise with the inclusion of improved or additional constructed response items, new item types and new assessment components. There are several important strategies to

Assessment Solutions Group - “*your* assessment experts”

215 Loch Lomond Way

Danville, CA 94526

210-859-9920

[**btopol@assessmentgroup.org**](mailto:btopol@assessmentgroup.org)

[**www.assessmentgroup.org**](http://www.assessmentgroup.org)



consider in order to hold down the cost of the new assessments. Consortia of states will be able to spread out the overhead for item development, project management, IT, QA/QC etc. over the consortia. This will result in a decrease in costs which will likely be offset, somewhat, by the requirements of increased security resulting from using a consortia. Teacher scoring of constructed response items may also be significant in driving down costs, although the extent of this benefit is related to the number of such items used, the extent to which all local educators can be called on to do the scoring and the amount, if any, to be paid to them.

The move to state consortia should also bring about an environment where greater standardization and use of best practices in assessment development and administration is possible. Today, while assessment functions are similar in all states, the operational manner in which these functions are carried out varies tremendously across states, driving inefficiency and higher cost. A group of states should be able to implement a set of standard development and administration activities that will reduce costs and improve quality. The combination of these factors will bring down the cost of new assessments, but by how much? It is important to be able to evaluate the impact of these strategies on assessment cost in order to design the most efficient and effective system.

Technology and Innovation Input

The use of technology in assessment administration will also be a key factor in the affordability of any new assessment system. Using the appropriate technologies and testing systems from the right vendors will result in dramatic reductions in assessment cost. Therefore, the move to state online testing is of critical importance. A key factor in the slow implementation of state online testing has been the high ratio of students to PCs and the resultant impact on the required length of testing windows. Therefore, we feel that an excellent use of federal money and efforts is to assist states in procuring additional PCs and to convene industry and expert groups to develop and define interoperability standards, features and functionality for testing systems.

Once states have enough PCs and the testing standards, features and functionality are mutually defined, the market will enable the innovation of new software systems, methods and technologies to bring testing into the 21st century. Efforts to develop a single platform will likely not be successful and/or not result in the best product as innovation is generally stifled in such situations. Ultimately, the ability to create and manage items electronically, administer tests on PCs and score constructed responses online will enable better and less expensive assessments. Competition among online test vendors should be *encouraged* so this happens as soon as practical and at the lowest cost.

Assessment Solutions Group - “*your* assessment experts”

215 Loch Lomond Way

Danville, CA 94526

210-859-9920

btopol@assessmentgroup.org

www.assessmentgroup.org



General Assessment Input

In light of the above, we believe it will be important for the USED, individual states, and state consortia to have access to assessment cost models that can determine the development and administration costs of new assessment systems, including online assessments. Scenarios need to be run to estimate the impact of the strategies mentioned above on state assessment cost. States cannot afford to go blindly into the process of developing new assessments.

The development of state consortia to implement common assessment based on the common core standards should be encouraged. It is, however, our belief that multiple consortia of multiple sizes should be encouraged. Differences in state testing calendars, budgets, online capabilities, designs, instruments, etc. may make it difficult to form large (> 25 states) consortia. Instead, the goal should be the number of states able to adopt the common core standards and not the size of state consortia.

In conclusion, we would recommend that the USED and states do the following:

- Develop new assessment systems that are cost-effective and flexible enough so as not to require states to find new funding in order to begin implementation of the new systems and to maintain the systems in the future
- Design assessment systems that include summative, interim benchmark, and formative assessments and include a variety of performance-based test items
- Create assessments that are universally designed and inclusive of all students, including students with disabilities and English language learners
- Allow states to implement new assessment systems over time
- Encourage teacher professional development as part of this effort to measure the common standards
- Gain access to assessment cost models that yield comprehensive cost data so both the USED and states can understand the cost and feature/functionality trade-offs of potential new assessment systems as they are being developed
- Conduct a detailed study of the costs for all types of assessment components among consortia of different sizes to not only determine the cost of the assessment but to also identify ways to improve the cost effectiveness and efficiency of different state assessment designs. The data from this type of study should be compared to “fair and

Assessment Solutions Group - “*your* assessment experts”

215 Loch Lomond Way

Danville, CA 94526

210-859-9920

[**btopol@assessmentgroup.org**](mailto:btopol@assessmentgroup.org)

[**www.assessmentgroup.org**](http://www.assessmentgroup.org)



reasonable” costs for each assessment element/function and this information can be used as reference points for the USED. Experts in determining benchmarks on what are fair and reasonable assessment costs can assist the USED with this.

- In their bids, all vendors should use a common, standardized cost sheet template that will allow for detailed cost data to be captured, analyzed in a cost model, and fairly compared across all proposals, so the USED can objectively evaluate the bids better and negotiate for more cost-effective approaches to be used with the state consortia. Cost input worksheets should consist not only of the dollars estimated to perform a specific activity but the key metrics involved in the activity, for example, number of items developed, number of pages composed, number of testbooks printed, etc. This will allow the consortia to make sure that the vendor understands the program and is bidding enough resources to do the job. It will also allow for apples to apples comparisons across vendors and/or consortia.
- Stimulate the development of online testing technology by helping states improve their student to PC ratios and form standard setting committees to help define testing system requirements.
- Take steps to encourage the market to develop next generation testing systems

Statement of Involvement in the State Assessment Process

The Assessment Solutions Group (ASG) is a consulting organization with a mission of assisting state departments of education in adding value throughout assessment costing, procurement and management functions. ASG senior consultants and technical advisors have more than 100 years combined experience in the assessment industry and expertise in all areas of the assessment function, making ASG unique in the industry in being able to provide states with services in test development, psychometrics, IT, production and manufacturing, quality assurance, scoring operations, and logistics. ASG makes extensive use of its proprietary costing model in providing services to its customers in the areas of cost-effective and efficient assessment program design. The company’s other product offerings include RFP preparation and analysis, technical and cost proposal reviews, ongoing assessment program evaluation, and program management services.

Assessment Solutions Group - “*your* assessment experts”

215 Loch Lomond Way

Danville, CA 94526

210-859-9920

btopol@assessmentgroup.org

www.assessmentgroup.org

Race to the Top

General Assessment
Public Meeting
Denver, CO

Christopher Camacho, PhD

Director of Research
Children's Progress

December 1, 2009



**Children's
Progress**

Improving Education Through Computer-Dynamic Assessment

(Response to General Assessment Questions 1 and 2)

Good afternoon. My name is Christopher Camacho and I am the Director of Research at Children's Progress. Children's Progress is an educational technology company devoted to helping schools foster learning for students in the early grades. Through over ten years of research in educational psychology and computer-dynamic assessments, Children's Progress has developed insights into how assessments can be designed and implemented to improve student learning. Based on our work, we believe there are several core principles of learning and assessment that the Department should consider as fundamental to achieving progress in your initiatives, some of which I'd like to share with you today.

1) A dynamic approach to assessment (providing scaffolding after incorrect responses) allows educators to better understand students' learning potential.

It is important to distinguish a dynamic approach to assessment from more traditional, static methods of assessment. Static assessments primarily only gauge a child's state of pre-existing knowledge; they are able to reveal two polar states of understanding: unaided success and unaided failure. However, dynamic assessments are identified by the objective to quantify a child's learning potential by presenting students with scaffolding after incorrect responses to dissociate what they can do independently from what they are able to do with guided assistance. This approach allows dynamic assessments to provide more valuable information for individualizing instruction to build upon a child's strengths and correct weaknesses.

2) Assessment must have formative value with content built upon a developmental model.

As we look toward a common set of learning standards, it is essential that educators know where children's skills fall within the developmental sequence of skills required for attaining proficiency. With this knowledge, educators are able to provide the most effective instruction. Further, assessment should not be a conclusion to a school year, but an integral component of the educational process. Frequently administered assessments - as often as three times a year - can be used to evaluate whether instruction is adequately addressing students' needs, particularly students who are identified as "at-risk." For younger children, assessment should take place with even greater frequency during the critical periods of development when measurement error and developmental lag have the greatest potential to impact instructional decisions.

3) Innovative technology should be used to create interactive assessment environments to engage students and to provide teachers with immediately available and actionable data.

As we explore new types of assessments, the manner in which these assessments are delivered must also be reconsidered. Assessment material specifically designed for and enabled by multimedia allows for the creation of engaging interactive environments that are capable of delivering much richer content and collecting more information within an assessment in a much shorter amount of time. More capable technology platforms also allow for immediate results, providing teachers and administrators with immediately interpretable and pedagogically useable information. An assessment taken this morning should impact instruction in the classroom this afternoon.

The Children's Progress Academic Assessment: Computer-Dynamic Assessment for Early Childhood

The kind of innovation that I have described here is not a far-off prospect. These principles are currently being implemented in districts and states across the country through their use of the Children's Progress Academic Assessment (CPAA) - a technology-driven, low-cost scalable assessment solution. The CPAA is a language arts and mathematics computer-dynamic assessment for children in pre-kindergarten through third grade. Assessment items feature encouraging audio feedback and interactive features to accommodate all young learners and is independently completed by a child in a typical class period. The content contained in the assessment is built upon a developmental model and designed to be used at least three times throughout the school year (developed with three discrete banks of content). The CPAA provides immediately generated graphical, narrative, and progress reports for teachers, administrators, and parents to help all educators individualized instruction. Moreover, the CPAA addresses early identification of potential academic problems. As the Department considers new approaches to assessment, special attention needs to be paid in the younger grades where early identification and intervention can have significant impact for the future success of children in school and life.

Additional information about the Children's Progress Academic Assessment can be found online at www.childrensprogress.com and in the included appendix.

Appendix A: Sample Assessment Item

The Children’s Progress has developed an assessment approach whereby incorrect responses are followed-up with scaffolded questions. The type of scaffolding presented to the child depends upon the child’s incorrect response. The example presented in Figure 1 is a screenshot from a sample rhyming question.

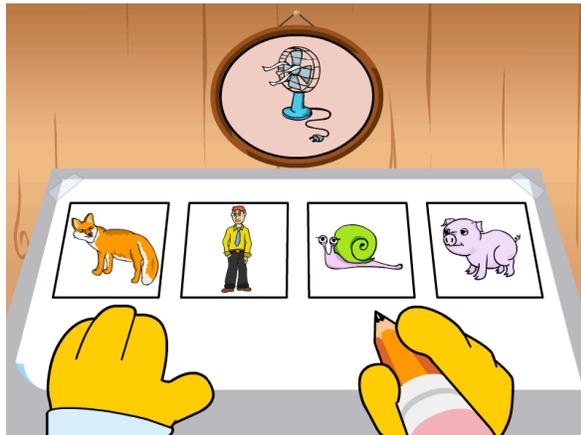


Figure 1. A rhyming question from the CPAA. In these questions, the child is asked to identify a word that rhymes with a target word. If the child answers the Independent Question incorrectly, then the child is presented with the Scaffolded Question.

Independent Question Audio Script: “Click on the picture that rhymes with the word ‘fan.’”

Scaffolded Question Audio Script: [presented when the child incorrectly clicks on “fox”]. “Fox. Fan. They sound the same at the beginning, but not at the end. Fan rhymes with can and pan. Click on the picture that rhymes with the word ‘fan.’”

Questions like this one were presented to children in kindergarten in the fall. All these questions began with the Independent presentation of the question and followed up by the Scaffolded presentation of the question only if the child answered the question incorrectly. The data from these rhyming questions is presented below.

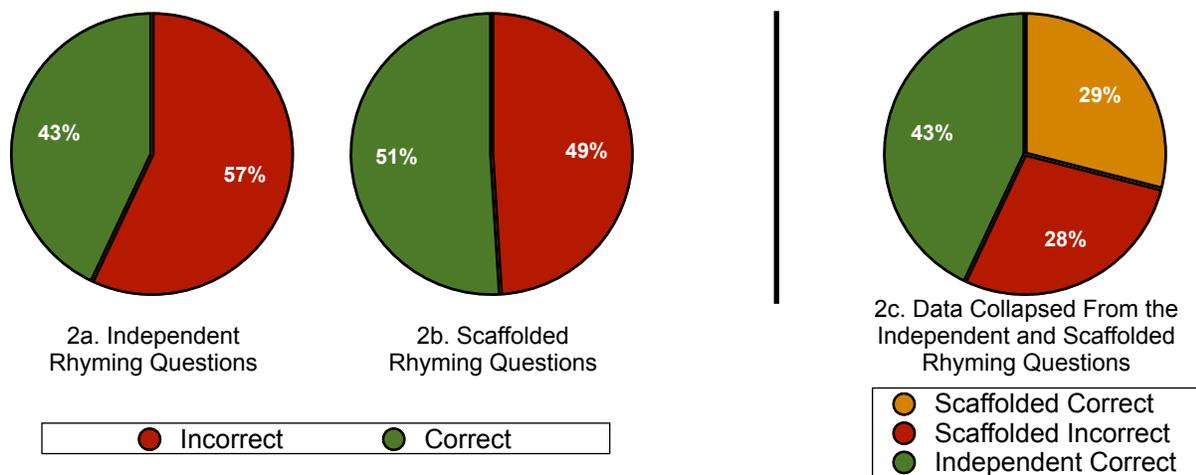


Figure 2a-c. Data collected from Independent and Scaffolded questions on rhyming from children in kindergarten. Figure 2a (left) presents the data collected from all the Independent rhyming questions. Figure 2b (center) presents the data collected from all the Scaffolded rhyming questions (children only see the rhyming questions after an incorrect response to the Independent question). Figure 2c presents the data collapsed from the Independent and Scaffolded questions. By presenting data with three different outcomes (as in Figure 2c), we can gain a deeper insight into the children’s understanding of the content. Certainly, children who answer correctly independently are different from children who answer scaffolded question correctly from children who answer scaffolded question incorrectly.

Appendix B: Sample Children's Progress Online Teacher Reports

The CPAA generates user-friendly reports for teachers, administrators and instructional specialists. All reports are available instantly (as soon as students complete the assessment). Below are a few examples of the reports available to teachers.

Fig 1. Class Summary Report

An overview of a classroom's latest assessment, with colorful charts representing performance levels by concept

- Teacher Tools
- Edit My Profile
- Download Software
- View Help
- Manage Roster [Add]
- This Report
- Print All Full Reports
- My Class
- Grades / Classes
- Alexander Johnson
- Students
- A - L
- [Abati, Trinity](#)
- [Axon, Yoshiko](#)
- [Bennick, Rosario](#)
- [Bernacchi, Oliver](#)
- [Brown, Samantha](#)
- [Copeland, Velma](#)
- [Dahlberg, Buffy](#)
- [Debraga, Lizeth](#)
- [Enix, Jed](#)
- [Greenleaf, Fred](#)
- [Locsin, Ulysses](#)
- M - Z
- [Niwa, Genia](#)
- [Schellhase, Leda](#)
- [Schrantz, Damian](#)
- [Storto, Frederic](#)
- [Streicek, Shalanda](#)
- [Trumbull, Gavin](#)
- [Wesner, Sherell](#)
- [Zike, Hilma](#)

Fall Gr2 | Winter Gr1 '07-'08 | Show All
Recent Assessments

Alexander Johnson's Class Report

Winter Gr2

Print version

Proctor:	Alexander Johnson	Legend: Above expectation (3.5 - 4.0) At expectation (2.5 - 3.5) Approaching expectation (1.5 - 2.5) Below expectation (1 - 1.5)
Date:	01/18/09	
Assessment:	CPAA Grade 2 Winter	

Report Areas

Report Card
Class Roll
Activities
Progress

Click on concept to see details.

Language Arts - Class's Concept Scores Score scales 1 to 4

Concept	Graph	Level	Class Avg.	School Avg.
Phonemic Awareness	<div style="width: 100%; height: 15px; background-color: #76b82a;"></div>	At Expectation	2.6	2.2
Reading	<div style="width: 100%; height: 15px; background-color: #76b82a;"></div>	At Expectation	2.5	2.7
Writing	<div style="width: 100%; height: 15px; background-color: #76b82a;"></div>	At Expectation	2.9	2.5

Mathematics - Class's Concept Scores Score scales 1 to 4

Concept	Graph	Level	Class Avg.	School Avg.
Measurement	<div style="width: 100%; height: 15px; background-color: #f1c40f;"></div>	Approaching Expectation	2.3	2.4
Numeracy	<div style="width: 100%; height: 15px; background-color: #f1c40f;"></div>	Approaching Expectation	2.4	2.4
Operations	<div style="width: 100%; height: 15px; background-color: #76b82a;"></div>	At Expectation	2.7	2.6
Patterns and Functions	<div style="width: 100%; height: 15px; background-color: #f1c40f;"></div>	Approaching Expectation	2.1	2.2

Fig 2. Class Roster

An interactive roster, sortable and printable by performance in any concept.

Teacher Tools

- [Edit My Profile](#)
- [Download Software](#)
- [View Help](#)
- [Manage Roster \[Add\]](#)

This Report

- [Print All Full Reports](#)

My Class

- [Grades / Classes](#)
- Alexander Johnson

Students

A - L

- [Abati, Trinity](#)
- [Axon, Yoshiko](#)
- [Bennick, Rosario](#)
- [Bernacchi, Oliver](#)
- [Brown, Samantha](#)
- [Copeland, Velma](#)
- [Dahlberg, Buffy](#)
- [Debraga, Lizeth](#)
- [Enix, Jed](#)
- [Greenleaf, Fred](#)
- [Locsin, Ulysses](#)

M - Z

- [Niwa, Genia](#)
- [Schellhase, Leda](#)
- [Schantz, Damian](#)
- [Storto, Frederic](#)
- [Strejcek, Shalanda](#)
- [Trumbull, Gavin](#)
- [Wesner, Sherell](#)
- [Zike, Hilma](#)

Fall Gr2 | Winter Gr1 '07-'08 | Show All
Recent Assessments

Alexander Johnson's Class Report

Winter Gr2

 Print version

Proctor:	Alexander Johnson	Legend: Above expectation (3.5 - 4.0) At expectation (2.5 - 3.5) Approaching expectation (1.5 - 2.5) Below expectation (1 - 1.5)
Date:	01/18/09	
Assessment:	CPAA Grade 2 Winter	

Report Areas

[Report Card](#) |
 [Class Roll](#) |
 [Activities](#) |
 [Progress](#)

View: Language Arts Mathematics

Click on the concept headers to sort by that concept.
Click on the student name to see that student's individual report.

Language Arts - Concept Scores Per Student Score scales 1 to 4

Students	Phonemic Awareness	Reading	Writing
Strejcek, Shalanda	1	2	3
Zike, Hilma	1	2	3
Abati, Trinity	2	3	3
Copeland, Velma	2	1	3
Enix, Jed	2	3	3
Niwa, Genia	2	4	3
Schantz, Damian	2	1	4
Trumbull, Gavin	2	4	2
Wesner, Sherell	2	4	4
Axon, Yoshiko	3	2	2
Bennick, Rosario	3	4	3
Bernacchi, Oliver	3	2	2
Brown, Samantha	3	1	2
Greenleaf, Fred	3	3	4
Schellhase, Leda	3	2	2
Storto, Frederic	3	4	3
Dahlberg, Buffy	4	2	4
Debraga, Lizeth	4	2	3
Locsin, Ulysses	4	2	2

Fig 3. Class Activity List

A list of recommended activities for a classroom (generated based on assessment performance). Each activity can be opened and printed, complete with a list of suggested participants. Activity lists can also be viewed for individual students.

Teacher Tools

- [Edit My Profile](#)
- [Download Software](#)
- [View Help](#)
- [Manage Roster \[Add\]](#)

This Report

- [Print All Full Reports](#)

My Class

Grades / Classes

- Alexander Johnson

Students

A - L

- [Abati, Trinity](#)
- [Axon, Yoshiko](#)
- [Bennick, Rosario](#)
- [Bernacchi, Oliver](#)
- [Brown, Samantha](#)
- [Copeland, Velma](#)
- [Dahlberg, Buffy](#)
- [Debraga, Lizeth](#)
- [Enix, Jed](#)
- [Greenleaf, Fred](#)
- [Locsin, Ulysses](#)

M - Z

- [Niwa, Genia](#)
- [Schellhase, Leda](#)
- [Schrantz, Damian](#)
- [Storto, Frederic](#)
- [Streicek, Shalanda](#)
- [Trumbull, Gavin](#)
- [Wesner, Sherell](#)
- [Zike, Hilma](#)

Fall Gr2
Winter Gr1 '07-'08
Show All

Recent Assessments

Alexander Johnson's Class Report

Winter Gr2

 Print version

Proctor: Alexander Johnson

Date: 01/18/09

Assessment: CPAA Grade 2 Winter

Legend:

- Above expectation (3.5 - 4.0)
- At expectation (2.5 - 3.5)
- Approaching expectation (1.5 - 2.5)
- Below expectation (1 - 1.5)

Report Areas

[Report Card](#) [Class Roll](#) [Activities](#) [Progress](#)

Click on each of the activities to see recommended participants.

Language Arts - Recommended Activities

Phonemic Awareness	Reading	Writing
Building Words	Active Reading	Dear Pen Pal
Dissecting and Creating Words	Beginning, Middle and End	Editing Scavenger Hunt
Singing Words	Boring Word Pit	Fill-in-the-Blank
Syllable Steps	Can You Arque With That?	Capitalization
Vowel Changes	Complete the Sentence	Fix the Mistakes
Word Jumble	Dice Roll	Grammar Reinforcer
Word Shuffle	Does It Belong?	Guess My Word
	How Are We Alike?	Invent-a-Poem
	I Wonder?	Missing Vowels
	Let's Make a Poem/Song	Mistakes Galore
	Listen To This	Period Hunt
	One Sentence	Sentence Stumpers
	Summaries	Spelling Bee
	Personal Dictionaries	Who's At Bat?

Fig 4. Student Detailed Report

A detailed, state standards-referencing narrative, outlining an individual student's assessment experience and highlighting specific strengths and weaknesses. Recommended activities are included based on concept-specific performance

- Teacher Tools
- Edit My Profile
- Download Software
- View Help
- Manage Roster [Add]
- This Report
- Print All Full Reports
- My Class
- Grades / Classes
- * Alexander Johnson
- Students
- A - L
- [Abati, Trinity](#)
- [Axon, Yoshiko](#)
- [Bennick, Rosario](#)
- [Bernacchi, Oliver](#)
- [Brown, Samantha](#)
- [Copeland, Velma](#)
- [Dahlberg, Buffy](#)
- [Debraqa, Lizeth](#)
- [Enix, Jed](#)
- * Greenleaf, Fred
- [Locsin, Ulisses](#)
- M - Z
- [Niwa, Genia](#)
- [Schellhase, Leda](#)
- [Schrantz, Damian](#)
- [Storto, Frederic](#)
- [Streicek, Shalanda](#)
- [Trumbull, Gavin](#)
- [Wesner, Sherell](#)
- [Zike, Hilma](#)

Winter Gr2 | Fall Gr2 | Winter Gr1 '07-'08 | Show All

Recent Assessments

Fred Greenleaf's Report

Winter Gr2

Proctor: Alexander Johnson

Assessment: CPAA Grade 2 Winter

Date/Time: 01/18/09 7:11pm

Legend:

- Above expectation (3.5 - 4.0)
- At expectation (2.5 - 3.5)
- Approaching expectation (1.5 - 2.5)
- Below expectation (1 - 1.5)

Print version

Go to class report

Report Areas

Report Card
Full Report
Activities
Progress

View: Language Arts Mathematics

Language Arts

Phonemic Awareness
Reading
Writing

Phonemic Awareness

At Expectation

Open all sub concepts (details view) | Close all sub concepts

[1]Fred added a phoneme to an existing word to create a new word containing a blend. [3]In the following section, Fred decoded a nonsense word containing a complex rime and a digraph without assistance.

Correct answer
 Correct answer with hint
 Incorrect answer

Phonemic Addition

Fred Greenleaf was able to:	Fred Greenleaf should be able to:	Recommended Activities:
Fred added a phoneme to an existing word to create a new word containing a blend. (PA.10.1.a PA.10.1.b PA.10.3.a)	Fred should blend sounds using knowledge of letter-sound correspondences in order to decode unfamiliar, but decodable, multisyllabic grade-level words (NY Learning Standard Reading 1-4).	Building Words
<input checked="" type="checkbox"/> If you add the sound /b/ to the beginning of "ring", what new word do you get?		
<input checked="" type="checkbox"/> If you add the sound /s/ to the beginning of "pot", what new word do you get?		

Fig 5. Student Progress

An individual student's progress in literacy and mathematics, sortable by concept and time period

Teacher Tools

- [Edit My Profile](#)
- [Download Software](#)
- [View Help](#)
- [Manage Roster \[Add\]](#)

This Report

- [Print All Full Reports](#)

My Class

Grades / Classes

- * Alexander Johnson

Students

A - L

- [Abati, Trinity](#)
- [Axon, Yoshiko](#)
- [Bennick, Rosario](#)
- [Bernacchi, Oliver](#)
- [Brown, Samantha](#)
- [Copeland, Velma](#)
- [Dahlberg, Buffy](#)
- [Debraqa, Lizeth](#)
- [Enx, Jed](#)
- * [Greenleaf, Fred](#)
- [Locsin, Ulysses](#)

M - Z

- [Niwa, Genia](#)
- [Schellhase, Leda](#)
- [Schrantz, Damian](#)
- [Storto, Frederic](#)
- [Streicek, Shalanda](#)
- [Trumbull, Gavin](#)
- [Wesner, Sherell](#)
- [Zike, Hilma](#)

Winter Gr2 | Fall Gr2 | Winter Gr1 '07-'08 | Show All

Recent Assessments

Fred Greenleaf's Report

Winter Gr2

[Print version](#)
[Go to class report](#)

Proctor: Alexander Johnson

Assessment: CPAA Grade 2 Winter

Date/Time: 01/18/09 7:11pm

Legend:

- Above expectation (3.5 - 4.0)
- At expectation (2.5 - 3.5)
- Approaching expectation (1.5 - 2.5)
- Below expectation (1 - 1.5)

Report Areas

[Report Card](#) |
 [Full Report](#) |
 [Activities](#) |
 [Progress](#)

View: Selected Year Over The Years

Subject scores for the selected year

Month	Language Arts	Mathematics
Oct	2.3	2.8
Nov	2.3	1.8
Dec	2.3	1.8
Jan	3.4	2.5

Concept scores for the selected year

Subject: Language Arts Mathematics

Month	Phonemic Awareness
Oct	3.0
Nov	2.0
Dec	2.0
Jan	3.0

Appendix C: Sample Children's Progress Print Reports

Any Children's Progress report can be printed. Below are some examples of commonly printed reports

Fig 1. Student Detailed (Narrative) Report

A list of the concepts the student was tested in, the corresponding state learning standard, and how the student responded (correctly, correctly after seeing a hint, or incorrectly)

Return to normal view | Print this page

Amelia Bedelia's Report

CPAA-K-Fall

Proctor: Teacher Preview
 Assessment: CPAA Kindergarten Fall
 Date and Time: 10/21/08 5:28am

View: Summary Full Details

Student Narrative Report: Full Details View

This report provides teachers with an additional level of detail. They can identify exactly which questions each student saw, the corresponding **CALIFORNIA CONTENT STANDARD**, and how the student responded (correctly, correctly after seeing a hint, or incorrectly).

Language Arts

Writing

Amelia was asked to find some letters. She identified 2 of 3 letters on the first try. She moved on to the letter-sound section. On the first try, Amelia was not able to match either of two presented letters to its respective sound.

Approaching Expectation

- ✔ Correct answer
- ✔ Correct answer with hint
- ✘ Incorrect answer

Letter ID

Amelia Bedelia was able to:	Amelia Bedelia should be able to:	Recommended Activities
Amelia identified 2 of 3 letters on the first try.	Amelia should recognize and name all uppercase and lowercase letters of the alphabet (CA ELA Content Standard for K – Reading 1.6)	Letter Hunt Matching Memory
<ul style="list-style-type: none"> ✘ Click on the letter "g". ✔ Click on the letter "S". ✔ Click on the letter "I". 		

Letter-Sound: Single Letter

Amelia Bedelia was able to:	Amelia Bedelia should be able to:	Recommended Activities
On the first try, Amelia was not able to match either of two presented letters to its respective sound.	Amelia should match all consonant and short-vowel sounds to appropriate letters (CA ELA Content Standard for K – Reading 1.14)	Alphabet Taboo Toss a Letter
<ul style="list-style-type: none"> ✔ What letter makes the sound /d/ as in dog? ✔ What letter makes the sound /t/ as in top? 		

Phonemic Awareness

Amelia matched two words with the same initial letter sound without assistance. She had difficulty with the rhyming section and was not able to rhyme a one-syllable word, even with guidance.

Approaching Expectation

- ✔ Correct answer
- ✔ Correct answer with hint
- ✘ Incorrect answer

Fig 2. Student Recommended Activities

A sampling of the activities recommended for a particular student based on his or her assessment performance, organized by subject and concept.

[Return to normal view](#) | [Print this page](#)

 **Amelia Bedella's Report**
CPAA-K-Fall

Proctor: Teacher Preview
Assessment: CPAA Kindergarten Fall
Date and Time: 10/21/08 8:28am

Mathematics > Measurement

Long and Longer
Length Comparison Instructional Activity: Have the class sit in a circle that includes you. Name an object that is not very long, e.g. a paper clip. Then go around the circle and have each person name an object (or distance) once you have exhausted long objects) that is longer than the previous one. In the second round, players have to name an object that is shorter than the last. Make sure you will make it around the circle by drawing attention to and preventing excessively large jumps in size. For example, a school bus is longer than a paper clip, but that may end the game. Challenge students to think of something just a little bit longer or shorter than the previous object.

Making Shapes
Shape ID Instructional Activity: Cut shapes out of felt. Divide the class up into groups. Give each group one of the felt shapes and challenge each group to lie on the floor and use their bodies to make that shape. When the group of students is done with their shape, take a picture of them. Then make a book of shapes with all of the pictures.

More Letters
Quantity Comparison Instructional Activity: Divide the students into pairs. Ask them to write their names on a piece of paper. If they struggle with writing, encourage them to copy their name from an already printed place. Next, have the students count the letters in their name and compare the quantities. Whoever has more proceeds to the child with the longest name in a nearby pair. Continue with the comparisons until the child(ren) with the longest name in the class has been determined.

Shoobox and A Ball
Positions - Reference Instructional Activity: Give one of your students a shoebox and a ball. Tell them to arrange the objects based on what positional term you use (i.e. put the ball inside the shoebox). Make this activity more challenging by asking students to remember the order of positions used. Similarly, you can have two students alternate in thinking of arrangements.

Mathematics > Numeracy

Matching Cards
Number ID Instructional Activity: Create a set of cards. Show the digit on one card and the matching number of dots on the second card. Make a pair of cards for the digits 1 - 10. Children then use the deck to match the digit with the corresponding pictures.

Reorganizing Objects
Subitizing Instructional Activity: Give children a set number of small circular objects such as marbles or checkers pieces. Ask the students to organize the objects into different shapes such as into a triangle, a house, or a rectangle. Ask the students which shape makes it easiest to identify the quantity. Then show them flash cards of different amount of circles and see who can guess the quantity of each one.

Sequential Surprise
Correct Order Instructional Activity: Write the numbers 1-25 on individual pieces of paper. Have the students write four blank lines on a sheet _____. Then pick a number and call it out. The students have to decide where to place the number. For example, if the number 25 was called, the student should place it last; the number four first. Once the number is put down, it cannot be moved. When you are finished calling out all four numbers, see who has the numbers in the correct order from smallest to largest. Take this opportunity to introduce probability concepts if you feel it is appropriate, for instance, by asking if most of the numbers in the 1-25 are above or below 5. How about 10, 20?

Switch Seats
Ordinality Supportive Activity: Ask the group to sit in one long line, and allow one student volunteer to stand outside of the line. Explain that you will be playing a game where the children will have to switch their seats. You will be calling out different ordinal numbers for the children to switch with, one at a time. Tell your volunteer to tap the third person in line. The third child then stands up and gives his seat to the original volunteer. Then you call out another ordinal number, "Tap the tenth person in line" - and the tapping and switching continues. Switch your wording around to reinforce different ways of ordering numbers. Instead of saying twenty seventh, say last. Instead of saying twenty sixth, say second to last. For older children, you could ask them to tap the child exactly in the middle.

Student Narrative Report: Recommended Activities

This report includes a sampling of the activities recommended for a particular student based on his or her assessment performance. Activities are organized by subject and concept.



RACE TO THE TOP ASSESSMENT PROGRAM
Public Hearing, Tuesday, December 1, 2009
Denver, CO

TOPIC AREA: General Assessment

Presenter: Lindy Crawford, PhD
Associate Dean
College of Education
University of Colorado at Colorado Springs
PH: 719-255-4308
Email: mcrawfor@uccs.edu

My name is Dr. Lindy Crawford and I am offering comments today on behalf of the National Center for Learning Disabilities (NCLD). NCLD is a not-for-profit organization founded in 1977 working to ensure that the nation's 15 million children, adolescents and adults with learning disabilities (LD) have every opportunity to succeed in school, work and life.

Currently, two and a half million school-age students receive special education due to learning disabilities. Many of these students are also English language learners. Ensuring that these students can participate in large-scale assessments that produce valid and reliable results is a top priority for NCLD. Our organization supports the accountability components of the current ESEA, particularly the expanded assessment and accountability provisions it contains. To that end, we have produced several reports designed to inform parents, educators, policymakers and other stakeholders of the positive impact of these accountability provisions for students with disabilities. Two of these reports are titled *Rewards and Roadblocks* and *Challenging Change*. Additionally, NCLD produced a detailed report examining the current situation regarding testing accommodations for students with disabilities. This report revealed substantial variance across states, in the area of allowable test accommodations, compromising the validity of what can be inferred from state test data. As author of that report, I am keenly interested in the issue of testing accommodations in the context of a new assessment system. Thank you for the opportunity to provide comments on the Department's proposed assessment initiative.

The development of common, high-quality assessments aligned with a common set of K-12 standards provides an unprecedented opportunity for equity among diverse learners, including students with disabilities and English language learners, the topic of your hearing tomorrow. The next generation of summative assessments must not nibble around the edges of innovation. They must, given our knowledge and expertise and the flexibility provided by technology, facilitate the full and equal participation of all learners.

To that end, on the behalf of NCLD, I offer the following six recommendations to guide the Race to the Top Assessment Program.

1. **Require assessments to be designed within innovative test delivery models, particularly online delivery systems.** Some advantages of online assessment include:
 - immediate score reporting so test results can guide instruction
 - decreased administrative burdens on school personnel
 - increased security of testing materials, and
 - more flexibility in test scheduling.

Additionally, online assessment environments allow maximum flexibility for any additional individual accommodations required by students with disabilities or English language learners.

2. **Require a “Universal Design” (UD) approach to test development.** Test development procedures must employ UD principles from the beginning to provide a more accurate measure of student achievement and eliminate many of the barriers that exist in traditional tests. A UD approach will eliminate the need for many test accommodations required in traditional testing situations, allowing for diverse learners to show what they know.
3. **Require assessments that embed individual student accommodations and allow student control over the test environment.** Researchers have developed systems of online testing environments that provide accommodations that adjust to individual student preferences on demand (such as those developed by Nimble Assessment Systems) as well as online accommodation decision-making tools (such as STELLA developed by Rebecca Kopriva and colleagues at the University of Wisconsin) that increase test validity. Research shows that

accommodations delivered within a computer-based testing environment increase the consistency and integrity of accommodations and result in improved utilization by the student. Students should be provided with an optimal testing environment that allows maximum student engagement and persistence.

- 4. Require states to accept only research-based testing accommodations considered as non-standard.** By non-standard accommodations we mean accommodations that influence the target skill, or measured construct, as opposed to standard accommodations that influence an access skill or non-measured construct. Any accommodation that influences the target skill or the skill measured by the test must be supported by rigorous research evidence. My report, published by NCLD, highlights the fact that many states are currently implementing test accommodation guidelines that are not defensible through research. While universally designed tests delivered within online testing environments are sure to eliminate the need for many test accommodations required in traditional tests, some accommodations will continue to be needed by certain students. Common assessments based on a common set of standards provide for the development of a common set of test accommodations across states. The standardization of test accommodations across states will dramatically improve both the validity and comparability of test results, making test data more useful to educators, parents and policymakers.
- 5. Require that any “adaptive testing” be aligned with grade-level standards.** While online testing environments hold great promise, they also offer opportunity to lower student expectations through “adaptive” approaches that adjust item difficulty based on student responses. Such approaches are not appropriate for summative assessments used for system accountability. While computer adaptive testing might be useful for formative assessment, its use in summative assessment would surely lead to decreased challenge for some students and a lowering of academic expectations for those students. The current ESEA testing requirements do not allow for “out-of-level” testing. This standard has resulted in the demise of a heretofore-widespread practice for students with disabilities. Today, schools are being held accountable for the performance of students with disabilities on general assessments with only limited exceptions. This advancement has resulted in improved access to the general curriculum, expanded learning opportunities and

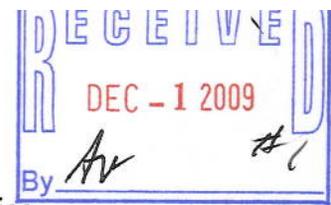
heightened expectations for millions of students. Therefore, any computer adaptive testing developed under this assessment program initiative **for use as a summative assessment** must be aligned to grade-level academic and performance standards. No exceptions for diverse learners such as students with disabilities and English language learners should be permitted.

- 6. Require empirical analyses of test items including the study of interactions between specific items and specific student populations.** Items should be analyzed to ensure that they do not disadvantage certain populations of students in their format and/or linguistic complexity. Research studies, such as cognitive labs, should be designed to investigate the interaction between students and test items. Interactions will differ within one broadly defined population of students (for example students with LD); therefore reviewing items in the absence of their specific interactions with students is insufficient. For assessments to provide useful results, all learners and their specific needs must be included in test development procedures, the field-testing of items, and post-hoc analyses of item by student interactions.

Thank you again for the opportunity to comment on this important initiative. I would be happy to answer any questions.

References

- Cawthon, S., Ketterlin Geller, L., & Carr, T. (2009) *Accommodations Decision Making: What (Online) Tools Can Increase the Validity Of Assessments for Students with Disabilities?* Presentation at the Council of Chief State School Officers Student Assessment Conference. Los Angeles, CA.
- Crawford, L. (2007). *State testing accommodations: A look at their value and validity*. New York, NY: National Center for Learning Disabilities
- Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large-scale assessments* (No. NCEO Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Education Outcomes.
- Universal Design for Assessment: A Multi-State Effort to Enhance State Assessments (2009) Presentation at the Council of Chief State School Officers Student Assessment Conference. Los Angeles, CA. Nimble Assessment Systems, Inc., Newton, MA 02458. www.nimbletools.com
- Winter, P. C., Kopriva, R.J., Chen, C. S., & Emick, J. E. (2006). Exploring individual and item factors that affect assessment validity for diverse learners: Results from a large-scale cognitive lab. *Learning and Individual Differences, 16*, 267-276.



Good morning/afternoon, and welcome to Denver, Colorado. I am Randy DeHoff, Vice Chairman of the State Board of Education. I am grateful for this opportunity to testify regarding the proposed assessment grant.

Colorado is in the midst of updating our standards and assessments to meet the demands of post-secondary and workforce readiness and the 21st Century, an effort that was initiated by legislation in 2008. Over the past twelve months drafts of revised content area standards have been developed, and are scheduled for the State Board to vote on them next week.

Design of a new assessment system aligned with the revised standards is also underway, and is scheduled to be completed by next December. From the very beginning this was envisioned as more than just a replacement for the current state assessment (CSAP). Some of us went so far as to hope that the new system would include components that would allow it to lead to the development and use of effective formative classroom assessments that directly led to improved instruction.

A National Association of State Boards of Education study group spent much of last year studying the issue of 21st Century Assessments, and they expressed that same hope. Their report, *Reform at a Crossroads: A Call for Balanced Systems of Assessment and Accountability*, issued just last month, calls for states to move to a comprehensive assessment system that extends down to the classroom level. I will refer to that report as I address some of the questions you have posed today, and I encourage you to refer to it frequently as you continue to define and refine the Race to the Top Assessment Program.

1) Propose an assessment system (that is, a series of one or more assessments) that you would recommend and that meets the general requirements and required characteristics described in the notice. Describe how this assessment system would address the tensions or tradeoffs in meeting all of the general requirements and required characteristics. Describe the strengths and limitations of your recommended system, including the extent to which it is able to validly meet each of the requirements described in the notice. Where possible, provide specific illustrative examples.

I would like to address your first question by quoting from the NASBE report: *“Assessments should measure applied knowledge and skills with the goal of all students passing rather than constructing measures that describe differences in student’s abilities. The system must ensure scalability so that more students are brought to high levels of performance.”*

An assessment system designed with the goal of all students passing is, in fact, the type of assessment system most compatible with increasingly sophisticated growth models. Such an assessment system must be designed to not just measure student learning, but to improve student learning. Summative assessment may do a reasonable job of the former; they contribute little to the latter. If our goal as education policy makers, and the goal of this assessment program, is to bring the level of assessment up to what is required in the 21st century, to develop assessment systems that complement 21st century standards, then those assessment systems must include more than summative assessments and the resulting data and reports.

The NASBE report outlines what the development of such a system will require: first, shifting investments in research and development from a single-point focus on large-scale assessment toward classroom assessment where teaching occurs; second, incorporating multiple assessments into a system of curriculum, instruction, and educator development that focuses on effective instructional practice; third, defining a clear set of learning goals. (In the Colorado process, defining Post-secondary and workforce readiness was the first step before beginning the revision of the content standards.); fourth, the system design must ensure that the resulting information from the assessment system has maximum utility for guiding instruction in relation to the learning goals.

3) ARRA requires that States award at least 50 percent of their Race to the Top funds to LEAs. The section of the notice entitled Design of Assessment Systems – LEA-Level Activities, describes how LEAs might be required to use these funds. What activities at the LEA level would best advance the transition to and implementation of the consortium's common, college and career ready standards and assessments?

The LEA's will play a key role in the development of this system. The NASBE report points out that local in-school performance assessments serve as the dominant mode of testing in most of the high-achieving countries around the world (e.g. Hong Kong, Singapore, Finland, and Sweden). At the high school level, these countries often use a combination of centralized, national exams (with primarily open-ended and essay items) and locally developed tests. Countries and jurisdictions such as Finland and Hong Kong create banks of tasks that teachers can draw from that include rich assessment tasks for classroom use for formative or benchmark purposes.

The role of LEA's would thus include a voice in the development of the state level assessments, piloting those assessments, and the development and piloting of lower level formative and benchmark assessments at the district, school and classroom level. If the goal is to develop an assessment system that provides an accurate picture of the learning that is taking place, from the individual student in the classroom up to an aggregated picture at the state level, the LEA role in that development is critical to ensure the system is aligned from bottom to top.

4) If a goal is that teachers are involved in the scoring of constructed responses and performance tasks in order to measure effectively students' mastery of higher-order content and skills and to build teacher expertise and understanding of performance expectations, how can such assessments be administered and scored in the most time-efficient and cost-effective ways?

If we accept this premise of the role of the LEA, then the role of the teacher goes well beyond involvement in the scoring of constructed responses. Teachers are central to the process of developing, administering, and scoring school-based classroom assessments as well. The development and deployment of in-class performance measures can serve as robust teacher development that fosters teacher-buy-in and readiness to adopt new instructional practices. Teachers should be trained to administer and evaluate student work using collaboratively determined criteria specified through standardized rubrics and scoring guides, all of which should be vertically aligned with the higher level assessments, content standards, and ultimate learning goals.

3) How would you recommend organizing a consortium to achieve success in developing and implementing the proposed assessment system? What role(s) do you recommend for third parties (e.g., conveners, project managers, assessment developers/partners, intermediaries)? What would you recommend that a consortium demonstrate to show that it has the capacity to implement the proposed plan?

The New England Common Assessment Program (NECAP) provides one of the only models of multi-state collaboration to develop an assessment. As such, there are several issues that must be considered and resolved before a successful consortium is possible. I don't have time to go over them here, but they are covered in the NASBE report.

Finally I would offer a strong admonition. Do not, as federal agencies are wont to do, be overly prescriptive in the requirements for this grant. No one has yet developed and implemented a 21st Century assessment system. I believe Colorado is on the way to doing that, and I believe that this grant program could provide a significant leverage to that effort. But while the general guidelines of such an assessment system may be clearly stated in the NASBE report and in the grant guidelines, the details of that system are still undefined. I encourage you to leave enough flexibility in the grant requirements to encourage proposals for different approaches to solving this problem.

From: John D. Forester [john.forester@wsaa.org]
Sent: Tuesday, December 01, 2009 4:43 PM
To: Race To The Top Assessment Input
Subject: Race to the Top Assessment Program

Office of Elementary and Secondary Education
Attention: **Race to the Top Assessment Program--Public Input Meetings**
U.S. Department of Education
400 Maryland Avenue, SW., Room 3E108, Washington, DC 20202

On October 26, 2009, the Department of Education requested input on a possible Race to the Top program for the development of and implementation of high quality assessments based on common standards.

The Department's notice stated: *If the Secretary determines that it is not feasible to conduct this second program, the \$350 million designated for this program will revert to fund additional grants under the general Race to the Top program.*

On behalf of the Association of Wisconsin School Administrators, the Wisconsin Association of School District Administrators, the Wisconsin Association of School Business Officials, and the Wisconsin Council of Administrators of Special Services **I am emailing to strongly encourage the Department to maintain this second program focused on high quality state assessment systems** and not to allow these funds to revert to the general Race to the Top program.

In Wisconsin, leaders at the school, district and state levels are prepared to transform our current state assessments into a high quality system that builds toward college and career readiness by the time our students' complete high school.

Federal support will be critical for Wisconsin to provide a system of world-class assessments for our students. The goal of developing high quality assessments based upon common standards is worthy of a second distinct program.

Thank you for this important opportunity to provide input.

Sincerely,

John D. Forester

Director of Government Relations
School Administrators Alliance (SAA)
4797 Hayes Road
Madison, WI 53704
608-242-1370
608-242-1290 (fax)
www.wsaa.org

Last Speaker

STATE OF COLORADO

OFFICE OF THE GOVERNOR

136 State Capitol Building
Denver, Colorado 80203
(303) 866 - 2471
(303) 866 - 2003 fax



Bill Ritter, Jr.
Governor

December 1, 2009

Office of Elementary and Secondary Education
Attention: Race to the Top Assessment Program – Public Input Meetings
U.S. Department of Education
400 Maryland Avenue, SW, Room 3E108
Washington, D.C. 20202

RE: Written Transcript of Testimony Provided by Dr. Matt Gianneschi, Senior Policy Analyst for Education for Bill Ritter, Jr., Governor of Colorado, for the Race to the Top Assessment Program Public and Expert Meeting in Denver, Colorado, December 1, 2008.

Submitted by: Matt Gianneschi, Ph.D.
Senior Policy Analyst for Education
Office of Governor Bill Ritter, Jr.
(303) 866-5800
Matt.gianneschi@state.co.us

Begin Testimony:

On behalf of Bill Ritter, Governor of Colorado, I'd like to thank you, Ms. Weiss and the members of the assessment expert panel for providing me with this opportunity to share a few recommendations concerning the direction of Secretary Duncan's assessment program. Governor Ritter had hoped to participate in today's meeting, but was unable to do so as a result of scheduling conflicts.

My comments will largely draw upon the recent experiences in Colorado related to assessment reform and, consequently, focus on ways in which the assessment program can support state-level efforts to align K-12 and higher education policies and ease the implementation of a new assessment program in participating states.

For Colorado, the Race to the Top Assessment Program could not have come at a better time, as the state is now commencing on a yearlong process to overhaul its assessments system. For the past two years, policymakers and educators throughout Colorado have been heavily involved in the process to reform and align the state's standards and assessments.

After years of often contentious debate, in 2008, the Colorado General Assembly passed Senate Bill 08-212, the "Preschool to Postsecondary Alignment Act" (otherwise known as the

“Colorado Achievement Plan for Kids” or “CAP4K”). This bill that required the full vertical alignment of state content standards between the PK, K-12 and higher education systems and, thereafter, the creation of a “new system of assessments” that reflect the content embedded in the new standards and can be used for the purposes of postsecondary admission and placement decisions. I am encouraged that the goal of this legislation is consistent with the stated intentions of the Race to the Top Assessment Program as found in the program’s public notice: to ensure that all students are ready for postsecondary education or the workforce by the time they exit school.

To accomplish the goal of preparing all students for successful transitions into postsecondary education and the workforce, Colorado’s educators and policy leaders have worked together to shift the emphasis of state assessments from one that focused simply on the achievement of annual academic benchmarks primarily for purposes of accountability to a more adaptable system focused on the mastery of specific competencies, alignment of content standards to bona fide college readiness definitions, and the use of growth as a key criterion of progress. In this new system, all students will be held to high standards for performance, but progress is contemplated as the incremental development of specific competencies toward a known postsecondary and workforce readiness definition, and not the amount of time a student sits in a classroom or the completion of a set of courses with certain titles.

With this as a backdrop, I offer the following three suggestions to the Department of Education for its consideration:

1. **Provide Waivers to Certain Aspects of NCLB** It is our sincere hope that the Department will assist participating states by considering incentives in the form of increased administrative flexibility necessary to promote the rapid state-level adoption of new assessment tools. To this end, we encourage the Department to review all existing laws related to state assessments and then consider ways to provide waivers to existing federal policies that could potentially prevent immediate implementation of new systems of assessments. In this way, the Department could help clear the way for states to make the transition from old systems to new.
2. **Focus Energy on the Development of On-line and Interim/Formative Assessments** The Department’s stated goals of providing “timely” assessments and having “the fastest possible turnaround time on scoring” should remain a principal priority. We in Colorado have discussed these same priorities at length and have reached a general understanding that the preferred way to accomplish them is through the on-line delivery of assessments. This idea has broad support from educators and policymakers alike in Colorado, but comes at significant expense and is further complicated by the need to provide universal access to expanded broadband internet bandwidth in schools. This matter is obviously beyond the scope of the Race to the Top Assessment Program. Nonetheless, by providing some waivers to federal requirements on states for the annual administration of state assessments—such as allowing participating states to alter their state assessment

plans quickly and without penalty—the Department could help free otherwise encumbered resources, thereby allowing participating states to make investments in other priority areas. And, while we strongly support the concept of periodic or interim style assessments, we would recommend resisting connecting such assessments to specific courses, as there is no practical way to govern the content of courses.

3. **Require Participating States to Adopt a Rigorous Definition of Postsecondary and Workforce Readiness and Align Assessment Instruments to This Definition** The development of rigorous, internationally benchmarked standards is an important advancement in education policy. However, in the absence of an equally rigorous definition of postsecondary and workforce readiness, the standards—regardless of the thought and care that went into their development—remain somewhat ambiguous. Creating explicit policy alignment between the K-12 and higher education sectors is a necessary prerequisite for the full implementation of assessments intended to ensure that all students are ready for postsecondary education or the workforce by the time they exit school. And while each state’s system of higher education is different and governed by widely varying policies, true cross-system alignment may not be fully realized unless states are able to create functional, standards-based definitions of readiness. Minimally, this definition should address the skills and competencies required to be placed into a credit-bearing college-level course. That is, it should be calibrated to a level that ensures placement above the basic skills or so-called “remedial” levels.

Again, I thank you for the opportunity to share a few recommendations with you. The Race to the Top Assessments program provides a tremendous opportunity for reform-oriented states like Colorado to meaningfully align their systems of education and realize the goal of ensuring that all students are ready for postsecondary and workforce readiness by the time they exit high school.



Feedback Regarding the Race to the Top Assessment Program

**Stuart Kahl
Measured Progress
December 1, 2009**

I'm Stuart Kahl, cofounder and CEO of Measured Progress. I appreciate having the opportunity to speak here today. Measured Progress is a non-profit company based in New Hampshire, with locations in three other states. We've been a contractor for state assessment programs for twenty-six years and currently operate programs in over twenty states, including NECAP (the New England Common Assessment Program) and MCAS (the Massachusetts Comprehensive Assessment Program) for which we were the original contractors. Consistent with our not-for-profit educational mission, our bottom line, recognized by our clients, is teaching and learning.

From the start, we've worked with states on assessments that are customized, inclusive, innovative, non-traditional, and geared toward a variety of student populations – general, special education, and English language learners. Our states' assessments have usually employed a variety of testing approaches, to include not only multiple-choice, but also extended constructed-response, performance tasks, and portfolios – both paper and computer-based. We've created scoring, standard-setting, and analytic techniques for these non-traditional formats that are widely used today.

I've been involved in large-scale educational assessment for thirty-five years and specifically in statewide assessments for almost thirty years. Drawing from these many years in the industry, I, along with my colleagues at Measured Progress, offer comments on four main areas addressed by the Race to the Top assessment program: consortia, multiple measures, teacher scoring, and competency-based testing. Generally, let me say that this program creates a wonderful opportunity. With widespread agreement that state assessment systems need to change and further agreement about the need for these systems to incorporate new components to accommodate multiple measures, including more complex and costly formats, the start-up costs associated with the development and implementation of new assessment systems would present an enormous challenge to states. The first year (or two) of any new program is always significantly more expensive than later, "maintenance" years because of the additional planning, coordination, test development, logistics and analysis programming efforts required. The need for financial assistance in implementing a new program is even greater when the former program continues to operate for data continuity during the early developmental stage of the new program. The Race to the Top assessment program provides states the opportunity to secure funding for the start-up years of their new programs.

State Consortia

Near the beginning of the Federal Register announcement, the support of “one or more consortia” was mentioned. In other documents related to the program, “number of states” in a consortium was identified as a factor in funding decisions. We commend the Department for recognizing the benefits of consortia and encouraging their formation. However, we caution the Department against favoring large consortia for several reasons.

While there have been some relatively large state consortia in the past, they were focused on a limited population of students (English language learners) or a specific, well-defined course domain (algebra). NECAP is the only comprehensive assessment program serving a state consortium. NECAP has been very successful by all standards. However, that success did not come easily and there were a lot of factors contributing to it.

The original NECAP states were three small, like-minded, geographically compact states. A fourth recently joined the group. Their savings were substantial, allowing them to preserve quality, rather than diminish it because of a need to cut back on expenses during economic hard times. For example, they preserved their significant use of constructed-response questions requiring human scoring.

For small states, sharing the fixed costs equally, fixed costs being those for such things as program management, test development, analysis and report programming, was a tremendous benefit since for them, fixed costs were a large part of their overall program budget. For very large states, fixed costs are relatively insignificant compared to variable (per student) costs, thus making consortia-related savings with respect to fixed costs relatively insignificant for them.

The variable costs (printing, materials handling, shipping/receiving, human scoring) are those dependent on the number of students in a state. Savings with respect to variable costs are quite substantial for small states in a consortium because banding together creates economies of scale. For example, going it alone, a small state’s constructed response scorers never get up to speed before they finish a question and start from scratch on the next one – not the case for large states. The large states already have economies of scale, so joining a consortium would offer more limited variable cost savings.

Geographic proximity of the NECAP states offered several advantages also. Management meetings of contractor and state staffs, test development committee meetings, and item and bias review meetings could be as often as needed, face-to-face, and low cost. The success of a consortium is all about relationships – the relationships needed to bear the larger burdens of reaching agreements, coordination, etc. With larger consortia, relationships are strained, with any one state’s influence – and “ownership” – diminished. Also, as mentioned earlier, like-mindedness is critical. The more diverse the states in a larger consortium, the more challenging the task of consensus building. Regarding the tests themselves, geographic proximity allows a regional flavor and greater relevance for reading passages and item contexts.

A letter report to Secretary Duncan from the Board on Testing and Assessment of the National Academies, dated October 5, 2009, makes a good case against the largest possible consortium (50 states). Decisions about federally mandated accountability assessments should not be based on a perceived need for comparability across states. There are too many obstacles to true comparability at both the national and international levels. Besides, NAEP gives us state comparisons that are as good as they're going to get. The problem with the percentages of proficient students being so variable across states and with many seemingly inconsistent with NAEP is that they show that there are some states that have set very low performance standards. All states performance standards should be high, not necessarily comparable. A national test is not needed to fix that.

In summary, we encourage the support of smaller consortia of states, say 3 to 5 states, because of the "diminishing returns" associated with larger numbers of states joining forces, diminishing returns in terms of both cost savings (modest for larger states) and ease of management, consensus building, ownership.

Multiple Measures

We applaud the Department for its emphasis on multiple measures, a hallmark of good assessment practice. No testing expert, company, or user manual has ever failed to warn consumers that major decisions should not be based on the results of a single test. Nevertheless, despite the mention of multiple measures in NCLB, few, if any, states have done justice to the concept. For some, the term meant including two different items types in the same test. As a result of the costs of testing at all the required grades and the timelines associated with meeting NCLB requirements, many states have not even gone that far.

There is considerable discussion across the country of the possibility of additional interim, perhaps local, curriculum-embedded components being added to states' accountability assessment systems. We believe this would be an excellent move, and apparently, so does the Department. However, we believe that the Department should offer guidance about what various components of accountability assessment should and should not be expected to accomplish.

Language in the Federal Register announcement about rapid turnaround and informing instruction can easily be misconstrued, as it often is in campaign rhetoric, to mean having immediate implications for a classroom teacher while teaching a tested topic. We believe an on-demand, combined multiple-choice and constructed-response summative test is a valuable component of an accountability assessment program. However, such a general achievement measure cannot be expected to serve this more immediate formative assessment purpose. It could, however, affect teaching and learning through the use of its results to inform program improvement efforts, a longer term process.

Regarding a curriculum-embedded component of accountability assessment, a component that we would advocate, we believe the Department should make clear certain properties

such a component should and should not have. A common complaint of local educators about end-of-year summative assessments is that they include items addressing content and skills that were taught six months earlier. They argue that tests students take during the course of instruction in a topic should count toward accountability results. We strongly disagree with this position. Schools should be accountable for seeing that students have retained important knowledge and skills. Thus, summative accountability testing should deal with retention, not short-term memory of students.

Taking this a step farther, we believe interim assessments that count toward accountability results should not cover material that can be tested via the more traditional on-demand summative measures. Many states have content standards that are not measured by their more traditional, on-demand summative tests – e.g., oral communication, research skills, media usage. These are the kinds of skills that curriculum-embedded performance assessment components could address effectively. These assessments, to use the words from the Federal Register announcement, would elicit “complex responses and demonstrations of knowledge and skills consistent with the goal of being college and career ready.”

We believe the Department, in its solicitation, should make it clear that for purposes of accountability, interim assessments using traditional measures of knowledge and skills recently taught are not desirable since their results would not reflect what the students ultimately retain. Instead they should tap important skills not readily assessed by the traditional, on-demand tests. (Note: There is a body of literature on how to conduct such performance assessments – i.e., how to assure the quality and rigor of the assessment tasks and how to allow local scoring with centralized auditing to assure scoring accuracy.)

Teacher Scoring

The Federal Register announcement includes a requirement for assessment systems to involve teachers in the “scoring of constructed responses and performance tasks in order to measure effectively students’ mastery of higher-order content and skills and to build teacher expertise and understanding of performance expectations.” There is no question that involvement in such scoring constitutes one of the best professional development activities teachers can experience. Having operated teacher scoring sessions associated with state testing programs, I can tell you that this is a message that is frequently repeated at debriefing sessions. We commend the Department for including this requirement.

Given several of the requirements of high stakes, statewide testing, however, we recommend that teacher scoring not be “overdone.” If, for example, a state’s program includes an end-of-year on-demand assessment component making use of constructed-response questions, we recommend the use of the testing contractors’ proven approaches to scoring – image scoring at contractors’ sites using experienced leadership and temporary scoring staff. Even though scoring of images of student responses can be done on a fully distributed basis allowing anyone to participate in scoring from any location, maintaining scoring accuracy and meeting stringent timelines are more likely

accomplished with the systems testing companies have established and operated for several years. Having the contractors handle constructed-response scoring for summative, on-demand testing does not diminish the potential for very effective use of released items, rubrics, and sample work in professional development activities or local testing.

Occasionally, an article appears in the popular press finding fault with constructed-response scoring. These are written by individuals who are uninformed about what's "under the hood" in these systems and the measurement quality they assure. Oftentimes, the critics attack the qualifications of the scorers/readers. However, the systems, as they exist, apply high levels of expertise where it is needed, at the front end of the scoring process – in the development of the scoring rubrics and the selection of student work corresponding to different score points for use in training and qualifying materials. This reduces the task of scoring to simple encoding or categorizing of responses, which many people can be trained to do effectively. After training, scorers must be qualified to score responses to each question by demonstrating an acceptable level of agreement between the scores they award to selected responses and the scores previously awarded by experts. Of course, scoring accuracy is monitored continuously during a scoring project by various forms of double scoring. The quality of the contractors' scoring systems is well documented in the technical manuals for the assessment programs.

If, on the other hand, an accountability assessment program includes a locally administered interim component, such as a curriculum-embedded performance assessment, then clearly teacher scoring would be desirable. The scorable products of such a component would be scored the same way as on-demand constructed responses, and in fact, products could include responses to follow-up constructed-response questions, along with reports, oral presentations, and other demonstrations of learning. A scoring audit process would also have to be implemented to assure the quality of scoring. There would still be valuable training and generally the same quality of professional development experience. Given the demand for multiple measures, including measures covering the standards not easily assessed by the on-demand tests, such curriculum-embedded components of a program would provide the ideal opportunity for teacher scoring addressing the two goals identified in the Federal Register announcement: measurement of higher-order skills and building of teacher expertise.

We recommend that the guidelines for Race to the Top assessment program funding refer to the "optimal combination of contractor and teacher scoring to complete scoring accurately and in a timely manner and to build teachers' expertise."

Competency-Based Versus Grade-Level-Based Testing

Rather than close with a recommendation regarding the Race to the Top assessment program, I will offer a general comment about competency-based versus grade-level testing, mentioned in the general assessment questions. I am a proponent of the competency-based testing. However, the full benefit of such testing will never be achieved as long as so many other aspects of our educational system are dominated by a grade-level orientation.

We know that in terms of their competencies, kids in a particular grade are far more variable than their grade-level curriculum. The grade-level focus of our instructional programs is inappropriate for the kids at the bottom and top of the competency continuum – too challenging for the first and too limiting for the latter. In this day and age, with the assistance of technology, individualized instruction is much more feasible, but its effectiveness, too, would be limited by grade-level shackles.

While it makes sense to group students by approximate age, we need to work with all kids where they are and get them to where they could be, and neither of those are best accomplished by our grade-level structures. So as far as competency-based testing is concerned, I'm for it; but a lot has to change in our schools and school programs to allow it to play its role in helping us raise achievement levels significantly.

Fair Educational Assessment in 24 Days

presented to

Race to the Top Assessment
Public and Expert Input Meeting
Grand Hyatt Denver
December 1, 2009

by

Clifford W. Lazar
Lazar Developments

Los Angeles

ceo@LazarDev.com

310-838-3885

For slides and text go to
Fair-Ed-Assessment.com



Assessing Principal and Teacher Performance



Very Brief Summary (Tweet):

- Educational Assessment Must Be Done Right
- Demographics-Based Assessment (DBA) is Meaningful and Actionable
- DBA is Superior to Value-Added

Fair Demographics-Based Educational Assessment versus Unfair



Unfair Assessment

Generates Resistance



Sabotage and Gaming the System

Hurts the Students

Fair Assessment

Makes Good Relations



Teachers Rewarded for Working Hard

Benefits the Students

Grading Teachers and Principals Fairly Meaningful and Actionable Results



2008 % Proficient & Advanced on CST tests	Projected	Actual-Projected (negative => underperformer)	Ratio Diff/Projected	School Grade	School Code Name
20	21.0	-1.0	-5%	C	BACH
16	14.8	1.2	8%	B	BEETHOVEN
12	14.6	-2.6	-18%	F	BERLIOZ
21	20.5	0.5	2%	C	BERNSTEIN
31	36.6	-5.6	-15%	F	BIZET
28	24.4	3.6	15%	A	BRAHMS
12	12.0	0.0	0%	C	CHOPIN
13	13.1	-0.1	-1%	C-	COPLAND
35	38.9	-3.9	-10%	D	DEBUSSY
26	28.7	-2.7	-10%	D	DONIZETTI
17	18.7	-1.7	-9%	D	DVORAK
23	22.9	0.1	0%	C	ELLINGTON
41	36.5	4.5	12%	B+	ENESCO
28	25.9	2.1	8%	B	FALLA
39	31.8	7.2	22%	A+	GERSHWIN
50	50.4	-0.4	-1%	C-	GOUNOD
37	32.4	4.6	14%	B	GRIEG
44	44.3	-0.3	-1%	C-	GROFE
26	27.4	-1.4	-5%	D	HANDEL
24	28.2	-4.2	-15%	F	HAYDN
A>15%	B>5%	C<5%,>5%	D>-15%	F<-15%	

The output of the Fair Demographics-based Assessment Methodology is a list of schools graded A, B, C, D, F
2 A's, 4 B's, 6 C's, 5 D's, 3 F's

The A schools out-performed their projected test scores by 15% or more based on the percentage of students who scored Proficient or Advanced on the 2008 CST (California Standards Tests).

The multiple regression projections were based on demographics including % African-American, % Latino, % White (not Latino), % English Learners, % Economically Disadvantaged % Moving in and out during the year
Did not include Census or Crime data

Amazingly, for sociological data, the R-square was .95

Collecting data and the analysis of 20 high schools (no magnets) took 9 days.

Demographics-Based Methodology vs. Value-Added



Issues	Value-Added	Demographics-Based
Definition	Value-added assessment, based on a review of students' test score gains from previous grades, can predict the amount of growth those students are likely to make in a given year	Using demographics based data corrects school achievement scores, allowing inter-school comparisons.
Problems - Time	Data Collection takes years of scores. Every curriculum and test change has a 2+ year lag	Demographic data gets outdated. No lag.
Problems – Student Populations	Inner city schools have 30% to 70% turn-over per year, making longitudinal results unreliable.	Racial and socioeconomic homogeneity can skew results. Implications of self-segregation into magnet schools not known.
Demographics	Value-added ignores demographics. Teachers and principals still can complain that demographics are ignored.	Multiple regression shows demographics has impact on achievement.
Cost	Higher, need to individualize results	Lower, uses whole school results
Time to Complete	Longer, years	Shorter, days
Fair-Ed-Assessment.com		

A Superior Methodology for Assessing Principal and Teacher Performance



-
- Easily Accessible School or Census Demographic Data. Doesn't require individualizing results.
 - Uses Agreed-to Assessment Test Data. Multiple achievement measures can be used, and weighted.
 - Immediate Results – doesn't depend on previous years of student test results
 - Corrects for student population demographics and neighborhood obstacles
 - Fair to Teachers and Principals
 - Can be completed for a middle-sized district in 24 days or less. This study was completed in 9 days.

Benefits of Fairness



- It is fair to the teachers and principals.
- All the counter-arguments posed by the teachers and principals are answered, and still we have a meaningful evaluation with good and bad performers.
- A second benefit is that it is cheap and immediate. Value-added requires multiple years of results, meanwhile the demographics and the teaching staff are changing. Value-added also hurts schools that over-perform in early years.
- This technology was used at Atlantic Richfield Oil Company, to determine the better gas stations. Poorer performers were shut down.

Fairness Methodology

Weaknesses, Strengths



Weaknesses

Demographic data gets out of date.

Not all students in school are from the surrounding area and visa versa.

Strengths

The Fairness Methodology works with any agreed-to achievement measure or multiple measures.

It is fair to the teachers and principals. Demographics and disruptive environments are part of the analysis.

Analyses are quick, easy and understandable. Just 24 Days. This project took nine, including the slides.

Data can be updated using school surveys, newspaper marketing data and census updates.

Fairness Methodology

Problems and Opportunities



□ Problems

It's new to education. Teachers are not statisticians. All the counter-arguments to teacher evaluations will be posed again by the teachers and principals.

□ Opportunities

Objections will be answered and we will have meaningful assessments, with good performers rewarded and under-performers identified and dealt with. Demographics-based approach points to best practices.

Fairness Methodology

Implementation Steps



	Estimated Days to Complete
<input type="checkbox"/> Collect School Test Data	1
<input type="checkbox"/> Collect School Service Boundaries	5
<input type="checkbox"/> Identify Included Census Tracts	5
<input type="checkbox"/> Collect Census Tract Data	2
<input type="checkbox"/> Identify Crime Collection Data Areas	5
<input type="checkbox"/> Collect Crime Data	3
<input type="checkbox"/> Store and Manicure the Data in Excel Spreadsheets	2
<input type="checkbox"/> Perform Multiregression Analysis	1
<input type="checkbox"/> List Schools and Grades Above and Below Expected/Projected Scores	.2
<input type="checkbox"/> Reward Schools Above Expected/Projected Scores	???
<input type="checkbox"/> Deal with Schools Below Expected/Projected Scores	??????

24.2 or less
Many of these tasks can be overlapped.

Fair Educational Assessment Source Data: LAUSD Report Cards



BELMONT SENIOR HIGH
2007-2008 School Year
1575 W 2ND ST
LOS ANGELES, CA 90026

API: **540** (Maximum = 1000; California state target = 800) 1-year change: **16**
Met AYP in 2008?: **No** (Criteria met = **11** ; Criteria possible = **22**)
Program Improvement status: **Year 5**

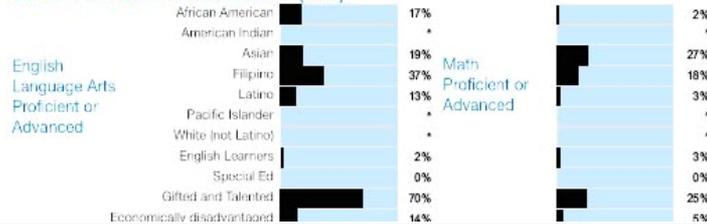


STUDENT OUTCOMES	2007-2008		1-YEAR CHANGE FROM 2006-2007 TO 2007-2008		3-YEAR CHANGE FROM 2004-2005 TO 2007-2008
Students graduating – in 4 years in 5 years	24%	27%	↑ 3%	↓ -6%	Coming December 2009
College / career readiness – students completing A-G requirements in 4 years 5 years	19%	18%	↓ -3%	↑ 6%	Coming December 2009
College / career going – students entering 2-year college 4-year university other	**%	**%	**%	**%	Coming December 2009
Proficient on state tests – students scoring Proficient or Advanced in English Language Arts Math	14%	5%	0%	↑ 1%	Coming December 2009
ACADEMIC PROGRESS					
Progress on state tests (1) – students improving in English Language Arts Math	27%	19%	↓ -9%	↓ -8%	Coming December 2009
Progress on state tests (2) – students declining in English Language Arts Math	-25%	-38%	↓ -8%	↓ -10%	Coming December 2009
English Learner progress – EL students improving on CELDT test	38%		↑ 3%		Coming December 2009
Seniors on track – 12 th graders taking SAT test scoring at least 1400 out of 2400	41%	11%	↓ -8%	↓ -6%	Coming December 2009
Juniors on track – 11 th graders passing EAP college-readiness test in English Language Arts Math	2%	7%	0%	↑ 1%	Coming December 2009
Sophomores on track – 10 th graders on track to graduate ready for college and career	11%		↑ 2%		Coming December 2009
Sophomores passing exit exam – 10 th graders passing both parts of CAHSEE	38%		↑ 8%		Coming December 2009
Freshmen and graduation – 9 th graders on track to graduate	49%		↑ 20%		Coming December 2009
INSTRUCTION, SCHOOL LEADERSHIP AND SCHOOL CULTURE					
Beyond test scores – teaching, leadership and culture	←		→		Coming December 2009
Professional attendance – faculty and staff attendance rate	94%				Coming December 2009
Campus safety – students feeling safe faculty and staff feeling safe	←		→		Coming December 2009
STUDENT AND PARENT CONNECTION					
Student attendance – average student attendance rate students with 95% attendance rate (miss fewer than 10 days)	87%	45%	↑ 1%	↑ 3%	Coming December 2009
Student satisfaction – students feeling satisfied and connected with their school and learning	←		→		Coming December 2009
Parent satisfaction – parents feeling satisfied and connected with their child's school and learning	←		→		Coming December 2009

SCHOOL OVERVIEW

# of students: 4,205	
2% African American	49% English Learners (EL)
0% American Indian	10% Special Education
3% Asian	4% Gifted and Talented
4% Filipino	82% Economically disadvantaged
90% Latino	83% Students moving in and out of this school during the year
0% Pacific Islander	
0% White (not Latino)	

STATE TEST SCORES BY GROUP (CST)



Fair Assessment Analysis: Excel Spreadsheet of Data From 20 High Schools



Microsoft Excel - High School data_03

File Edit View Insert Format Tools Data Window Contribute Help Adobe PDF

Open In Contribute Publish To Website Post To Blog

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1																	
2	Code Name	% African-American	% Latino	% White	% English Learners	% Econ Disadvantaged	% Moving In/Out	% Graduating 2007	% Graduating 2008	% Proficient on State Tests 2007	% Proficient on State Tests 2008	2008 % Proficient & Advanced on CST tests	Projected	Actual-Projected (negative => underperformer)	Ratio Diff/Projected	School Grade	School Code Name
3	BACH	54	34	0	13	78	58	27	24	20	2	20	21.0	-1.0	-5%	C	BACH
4	BEETHOVEN	57	42	0	19	66	50	36	34	16	1	16	14.8	1.2	8%	B	BEETHOVEN
5	BERLIOZ	9	90	0	39	78	45	29	28	13	1	12	14.6	-2.6	-18%	F	BERLIOZ
6	BERNSTEIN	0	99	0	34	89	30	45	47	21	5	21	20.5	0.5	2%	C	BERNSTEIN
7	BIZET	4	61	29	24	78	33	50	50	31	7	31	36.6	-5.6	-15%	F	BIZET
8	BRAHMS	7	75	10	33	76	44	44	38	28	7	28	24.4	3.6	15%	A	BRAHMS
9	CHOPIN	10	89	0	45	84	39	27	24	12	1	12	12.0	0.0	0%	C	CHOPIN
10	COPLAND	21	78	0	37	77	46	33	30	14	5	13	13.1	-0.1	-1%	C-	COPLAND
11	DEBUSSY	5	73	11	15	55	25	55	57	36	20	35	38.9	-3.9	-10%	D	DEBUSSY
12	DONIZETTI	1	81	1	32	76	29	45	44	26	10	26	28.7	-2.7	-10%	D	DONIZETTI
13	DVORAK	10	77	1	40	67	48	35	39	17	10	17	18.7	-1.7	-9%	D	DVORAK
14	ELLINGTON	4	82	5	34	71	47	42	40	23	13	23	22.9	0.1	0%	C	ELLINGTON
15	ENESCO	4	71	17	20	74	26	58	58	41	16	41	36.5	4.5	12%	B+	ENESCO
16	FALLA	2	90	3	29	73	30	50	56	29	12	28	25.9	2.1	8%	B	FALLA
17	GERSHWIN	7	73	11	23	65	35	49	47	39	19	39	31.8	7.2	22%	A+	GERSHWIN
18	GOUNOD	16	32	41	10	43	22	60	55	50	29	50	50.4	-0.4	-1%	C-	GOUNOD
19	GRIEG	19	59	10	21	67	27	45	44	38	14	37	32.4	4.6	14%	B	GRIEG
20	GROFE	5	50	13	20	58	32	46	46	44	28	44	44.3	-0.3	-1%	C-	GROFE
21	HANDEL	72	17	8	5	39	23	49	46	27	8	26	27.4	-1.4	-5%	D	HANDEL
22	HAYDN	1	93	1	23	78	30	44	44	24	5	24	28.2	-4.2	-15%	F	HAYDN
23												A>15%	B>5%	C<5%,>5%	D>-15%	F<-15%	
24																	
25	85.21857312	-0.59664	-0.39964	-0.0587	-0.49343	-0.04512	-0.1458										

Ready

start Google

Copyright (c) 2009, by Cliff Lazar

Fair Educational Assessment Acknowledgements:



Jerry B. Wong, US Census Bureau

Cynthia Lim, Executive Director Office of Data and Accountability,
LAUSD

Grace Pang Bovy, Director, School Information Branch Office of
Data and Accountability, LAUSD

Peter Rosenstein, Teacher, Reseda Elementary School, LAUSD

Roger Rasmussen

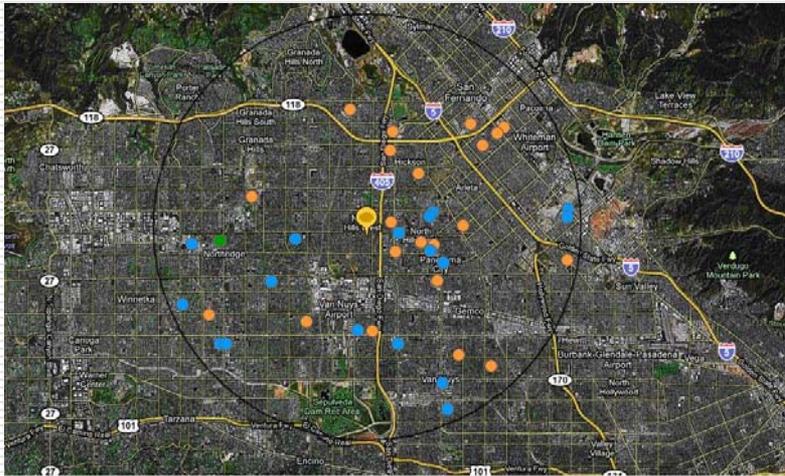
Fred Stern

<http://getreportcard.lausd.net/reportcards/reports.jsp>

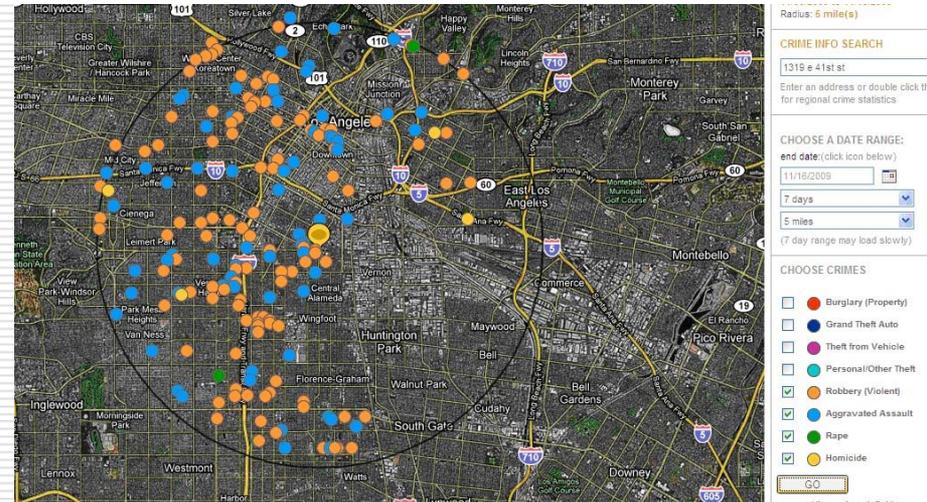
http://factfinder.census.gov/servlet/QTGeoSearchByListServlet?ds_name=DEC_2000_SF1_U&lang=en&_ts=277395136750

Is Crime an Issue?

Crimes in Seven Days within 5 Miles of High School



Monroe High



Jefferson High

You betcha!

Clifford W. Lazar

Brief Resume



Experienced college and university instructor, lecturer in economics, computers and security.

Experienced real estate owner/manager.

Developed numerous successful competitive proposals for federally funded projects.

Candidate for LA Board of Education

Manager, Information Services and Technology, ARCO Transportation

Manager, Management Science Department, ARCO Transportation

Manager, Government Affairs Information System, AtlanticRichfield

Consultant, Management Science, ARCO Petroleum Products Company

Section Chief, Economic Analysis, Advanced Projects, Litton Ship Systems

Senior Systems Analyst, Computer Department, TRW Systems

Financial Analyst, Controller Department, TRW Systems

Publications: Numerous articles and training talks including:

“Political Economics of Computer Security”

“Ten Pages to Windows, Word, WordPerfect, Excel and Access”

“Export Coupons: Solution to Balanced Trade”

US Patent and Patent pending

Education: B.A., M.A., all course work towards Ph.D. in Economics, UCLA

**Talking Points for Jody Papini,
Douglas County Federation, CO
On Behalf of the American Federation of Teachers,
To the U.S. Department of Education
Dec. 1, 2009**



My name is Jody Papini. I have been teaching for 15 years working with this group of students and helping them succeed is my passion. I am a member of the Douglas County Federation and the American Federation of Teachers. I have provided professional development for teachers through the AFT's ER&D Thinking Mathematics courses, some of which I have helped develop. I am currently a math instructional coach in my district.

Today, I speak on behalf of teachers and I ask that as you write the guidelines that will shape the development of the next generation of assessments. Please consider AFT's *Smart Testing* criteria which starts with strong, grade-specific content standards, and includes a number of interrelated pieces:

- Well-developed grade-by-grade curricula;
- Assessments aligned to content standards;
- An efficient, valid, and reliable testing system that does not duplicate testing across education system level;
- Appropriate inclusion of English language learners (ELLs) and students with disabilities in testing programs;
- Timely provision of user-friendly testing results for teachers and students;
- Supportive professional development, including coverage of what the content standards are and how they relate to state curricula and assessments, how to teach to the content standards, and how to use testing data to inform instruction;
- Accountability for results; and,
- Transparency of the system.

Some important pieces of the smart testing criteria have been clearly violated or neglected under the current system.

- Standards are often so broad and ambitious even in places that have grade-by-grade curricula the expectations are unrealistic and overwhelming. There is so much material to cover in a school year; some teachers have expressed concern over not having the time to take advantage of teachable moments; teachers are at times faced with difficult choices such as taking an extra day or two or three to re-teach material that they know students have not mastered, knowing that at the end of the year they will be rushed or simply not able to cover all of the required material.
- Under the current assessment system, states are mandated to administer summative assessments once a school year. However, some states and many districts have developed additional interim and/or benchmark assessments resulting in multiple layers of testing at the classroom level. During focus groups conducted by the AFT, teachers have calculated that up to 20-25% of the school year can be consumed by the summative, interim and benchmark assessments alone. These additional assessments often aim to emulate the summative assessment such that students are tested and retested on similar material. In other cases, these assessments do not align to the summative assessment, so that teachers spend the school year administering assessments and receiving data that does not align or inform progress toward higher achievement on the summative assessment currently used to evaluate schools. This practice does not make the best use of the already scarce instructional time.

In developing the next generation of assessments, require that those overseen the development and implementation of these assessments develop a system that incorporates aligned standards, curricula, assessments, and professional development; tests that do not duplicate across education system levels; user-friendly test results; accountability of results, transparency; and appropriate inclusion of English language learners and students with disabilities. And, require that they take into account the impact of such assessments on the day-to-day classroom experience of our children.

Parents United for Responsible Education (PURE) position paper
Assessment and Accountability under NCLB
December 1, 2009

Submitted by Julie Woestehoff
Executive Director

Parents United for Responsible Education

100 S. Morgan Street Chicago, IL 60607

Tel. 312/491-9101 Fax: 312/491-9404

pure@pureparents.org www.pureparents.org

PURE believes that a high-quality assessment and accountability system is essential to a high-quality public education for all children. We support assessment and accountability systems which are built on high-quality learning standards, incorporate multiple measures of student progress over time, value local assessment, are transparent to the public, and demonstrably support improved teaching and learning.

Our specific recommendations for ways the No Child Left Behind Act (NCLB) can support such a system are as follows:

● **Specify that state test scores may not to be used alone to make important educational decisions about children.**

● **Require that other measures of student progress beyond standardized tests be included in student and school assessment.**

● **Require publication of significant portions of any annual state standardized tests.**

● **Require that states allow parents to opt their children out of any state or local standardized test.**

● **Specify regular public review and revision of state learning standards and related assessment.**

● **Locate the key elements of school evaluation at the local school community level.**

Our detailed rationale for these recommendations follows.

● Specify that state test scores may not to be used alone to make important educational decisions about children.

Rationale: Since 1996, set cut-off scores on first the Iowa Tests of Basic Skills or the Illinois Standards Achievement Tests have been used as promotion barriers for Chicago Public Schools children. This practice violates the test makers' guidelines, sound educational practice, and the standards of the testing profession. It has led to a higher drop out rate of younger children and a narrowing of the curriculum to focus on standardized test skill drill. It has waged emotional warfare on CPS children without improving the overall quality of educational outcomes.

When standardized test scores are the only or the predominant measure of school improvement, as has happened in most states under the No Child Left Behind Act (NCLB), the pressure grows on schools and districts to show increasingly large test score gains. This provides is a powerful motivation for states, districts, and schools to attach high stakes to individual students' standardized test scores. We recommend that this not be an option.

Please see addendum detailing the problems with standardized testing in CPS.

● Require that other measures of student progress beyond standardized tests be included in student and school assessment.

Rationale: Most state learning standards are a fairly comprehensive list of what students should know and be able to do. The majority of these standards cannot be assessed using multiple choice tests, even if they are supplemented by open-ended and essay questions. The pressure to raise test scores has caused states, including my state of Illinois, to emphasize the skills and knowledge which can be assessed by paper-and-pencil tests over other, less “testable” knowledge and skills.

In fact, the Illinois State Board of Education prominently posts “Learning Frameworks” on nits web site:

<http://www.isbe.state.il.us/assessment/IAFIndex.htm>

which ISBE describes as “clearly defining those elements of the Illinois Learning Standards that are suitable for state testing.”

ISBE goes on to state, “They are not designed to replace local curricula and should not be considered state curricula. ” While the caveat is laudable, it has hardly discouraged teaching to the test in Illinois.

NCLB must require that state assessment and accountability systems are comprehensive of the widest range of educational content. This can only be done if the system is required to include forms of assessment such as portfolios and demonstration, which successfully evaluate critical areas of learning that cannot be assessed using standardized tests. NCLB must require states to adhere to assessment best practices including use of multiple measures for assessment and

accountability; NCLB must define what that means and enforce this provision of the law.

● Require publication of significant portions of any annual state standardized tests.

Rationale: Many states include a requirement that their testing system be open and transparent. This is essential if there is to be public trust in standardized tests. There are countless examples of incorrect, racist, or otherwise bad questions that have appeared on state standardized tests and have been made public because of transparency laws. This is not the case in Illinois, where bad questions and inappropriate illustrations, for example, have become public knowledge only through leaks to the press. The public has a right to know what the tests look like in context, not just in the outrageous example.

● Require that states allow parents to opt their children out of any state or local standardized test.

No one test should carry enough weight for its absence to make a meaningful difference in the overall evaluation of one child, a school, a district, or a state. Parents must have the ability to determine their child's best interests as it relates to any assessment or other educational program. NCLB has always given parental involvement an appropriately key role in many NCLB areas; student assessment should be included as an area where parental involvement and parents' rights are important,

● Specify regular public review and revision of state learning standards and related assessment.

Rationale: The ISAT and state learning standards are in need of improvement (as per Achieve, for example). Most parents we work with are quite unaware of the state learning standards and may or may not agree that they capture the most important things that children should know and be able to do. These statements should be reviewed and revised by all the stakeholders. Special emphasis must be given to involving parents (not just one or two token parent representatives, or other stakeholders who claim to represent parents, too, because they also have children). Parents have the most at stake in what their children are being taught. We need to know and understand what is expected of students if we are to support their learning at home. There must also be greater opportunity for parents and other members of the public to consider how those standards should be assessed.

● Locate the key elements of school evaluation at the local school community level.

Rationale: Because so many key areas of state learning standards cannot be effectively assessed through multiple choice tests, and to increase the involvement of the public in evaluating their schools, we recommend a return to emphasis on annual on-site school reviews as a key component of school accountability.

In Chicago, the school improvement plan (SIP) is the central accountability document for each

local school, and the elected local school council (LSC) is the body that oversees the school review process. Through the SIP process, the LSC brings the school community together to review the school's current status, develop focused plans and strategies, monitor and evaluate the effectiveness of the school's educational services, and, based on that review, plan for the coming year's programs. This individualized, qualitative system is fundamental to local school improvement. We support the return of the local school improvement plan to its position as the central accountability document for the state, the district, and the local school community. This gives back the primary role in student assessment to those who know the students best — their teachers, other school professionals, and families.

Parents United for Responsible Education (PURE) position paper
Addendum to PURE's proposals
Assessment and Accountability under NCLB
December 1, 2009

Submitted by Julie Woestehoff

Executive Director

Parents United for Responsible Education

100 S. Morgan Street Chicago, IL 60607

Tel. 312/491-9101 Fax: 312/491-9404

pure@pureparents.org www.pureparents.org

PURE has recommended that Chicago Public Schools implement true multiple measures of student (and school) performance including high-quality formative and summative assessments in the various subjects, as well as other indicators to provide evidence of improved student learning and school quality. These assessments should be based on state standards and the local curriculum, assess higher order thinking and other 21st century skills, and provide multiple approaches for students to demonstrate their learning. The primary use of these assessments should to improve instruction and enable teachers to better address each student's strengths and needs.

We recommend a balanced combination of measures over time to determine a students' placement including portfolio reviews, classroom-based assessments, and occasional district-wide project-based demonstrations such as the ones proposed in 2003 by the CPS Commission on Curriculum-based Assessments.ⁱ

The problem: The way CPS uses standardized tests to retain students violates accepted standards for test use

CPS uses student scores on the 3rd, 6th, and 8th grade reading and mathematics SAT-10 test, which is embedded in the ISAT, to determine whether or not a student will be promoted. According to the test makers themselves as well as state and federal education agencies, this practice is improper, violates professional testing standards. The policy ignores better, sounder, less discriminatory means of identifying students who need the most help.

The SAT-10 was not designed to determine student promotion status. Using a test for a purpose for which it was not designed is considered an improper use by the test makers, the nationally-accepted standards for the testing profession, the state of Illinois, and the U. S, Department of Education.

The test makers, Harcourt Assessment, state in their *Guide for Organizational Planning*,

Another misuse of standardized achievement test scores is making promotion and retention decisions for individual students solely on the basis of these scores. This is an undesirable practice for a number of reasons. Perhaps the most important reason is that national standardized achievement tests are not built to serve this purpose...they cannot provide complete coverage of any local curriculum.ⁱⁱ

In a letter written to PURE on May 11, 2009, Marcilene Dutton, Deputy General Counsel, Illinois State Board of Education, stated:

Using ISAT scores as the basis for student promotion and retention is not an ISBE policy or practice.ⁱⁱⁱ

A January 27, 2009 e-mail from Judith Steinhauser, representing ISBE, to parent Wade Tillett, stated:

the purpose of ISAT, its reliability and validity authenticated by a staff of psychometricians, is to calculate school accountability which is reported to the federal government as Adequate Yearly Progress. It is not the intention of the state to use the test for anything else.

The USDE manual, "Taking Responsibility for Ending Social Promotion," states:

When a statewide or districtwide test is being used to determine student promotion, the state or district must be able to provide professionally acceptable evidence that the test is valid and reliable for the purpose for which it is being used. If a state or district chooses to use a test as a principal criterion for decisions about student promotion, the test must be designed for this use and there must be evidence that it is appropriate to use the test as a sole or principal criterion.^{iv}

CPS improperly uses the SAT-10 as a sole criterion for making promotion decisions, a practice opposed by the test maker, state officials, and national experts.

The makers of the SAT-10 state:

Achievement test scores may certainly enter into a promotion or retention decision. However, they should be just one of the many factors considered and probably should receive less weight than factors such as teacher observation, day-to-day classroom performance, maturity level, and attitude.^v

The ISAT "professional practices" manual lists under "Prohibitions: Actions that must be avoided when reporting test results":

- No person or organization shall make a decision about a student or educator on the basis of a single test.^{vi}

The National Research Council, in their major study on student assessment, states this principle clearly:

(A)n educational decision that will have a major impact on a test taker should not be made solely or automatically on the basis of a single test score. Other relevant information about the student's knowledge and skills should also be taken into account.^{vii}

Standard 13.7 of the *Standards for Psychological and Educational Testing* reads as follows:

In educational settings, a decision or characterization that will have a major impact on a student should not be made on the basis of a single test score.^{viii}

The Code of Fair Testing Practices in Education prepared by the Joint Committee on Testing Practices calls on test users to

Avoid using a single test score as the sole determinant of decisions about test takers. Interpret test scores in conjunction with other information about individuals.^{ix}

CPS has established multiple barriers to promotion, while falsely contending that they are multiple measures. After PURE filed a discrimination complaint against the policy in 1999, CPS began to include classroom grades and attendance in the promotion decision. But instead of using these other criteria as true multiple measures, which testing experts recommend, the policy uses them as multiple barriers.

It is critical to understand the difference between multiple barriers and multiple measures. Under multiple barriers, the student must meet all of several listed criteria. Under multiple measures, also called multiples sources of evidence, the various measures are combined, not used separately. True multiple measures may, for example, use a weighting system to reflect the proportionate usefulness of different assessments. Alternatively, results may be added together using a point system to come up with a total number, or one or more positive results may compensate for, or “outweigh,” a less positive outcome.

As noted above, the test makers themselves say that the test

should be just **one of the many factors** considered and **probably should receive less weight** than factors such as teacher observation, day-to-day classroom performance, maturity level, and attitude^x (*emphasis added*)

In fact, in the CPS promotion policy, **each measure operates as a single deciding factor, each of which on its own can be used to retain the student.** In other words, CPS students must meet district-wide assessment (DWA) cut scores **and** grade standards **and** attendance standards in order to be promoted without attending summer school.

Test scores alone are explicitly used in several of the policy's high-stakes decisions. For example, eighth grade students are banned from graduation with their classmates if they do not meet **all** of these measures. Students whose DWA scores were below the cut off point must pass **one end-of-summer-school test** in order to be promoted to the next grade.

Other useful information as student attendance, academic performance throughout the school year, and faculty recommendations are readily available. These factors are indeed considered when a student successfully exceeds the cut-off score, but then only in a negative sense; low attendance or a failing grade will **also** bar that student from graduation or send him or her to summer school.

Stated simply, students can be hurt by their attendance and academic performance, but these measures cannot help them. They are multiple barriers, not multiple measures, which means that **each one of the measures is a single high-stakes measure.**

SAT -10 results can differ from overall ISAT results. The SAT-10 consists of only 30-40 questions embedded in the ISAT. PURE has learned that, after attending summer school for low SAT-10 scores in 2008, some students receive their ISAT scores – scores from the same test – stating that they meet state expectations.

In a response to a PURE request under the Freedom of Information Act about the correlation of SAT-10 results with ISAT results, PURE found that CPS sent 26,992 students in the “benchmark grades” to summer school in 2008. However, 1,412 of those same students who scored below the CPS cutoff point in math were also found by the state to meet the standard in math. And 13,071 students who scored below the CPS cutoff point in math were also found by the state to fall in the state's 'below standards' category rather the lowest category, “academic warning.’ The state found only 3,430 students to be at the academic warning level in math, and even fewer in reading. The difference in results was similar in 2006 and 2007.

The discrepancy occurs because CPS bases its promotion policy on only two small subsets of the overall test (30 or 40 questions each) that are graded quickly to determine who must attend summer school. These scores don't necessarily match with final overall ISAT scores.

When asked about the correlation between CPS cutoff score and the state standard levels, CPS responded that the correlation is “an ISBE matter.”^{xi}

CPS's use of ISAT scores as a pass-fail barrier is not justified by any compelling educational reason, and less discriminatory alternatives are available. In its 1999 agreement with OCR, CPS agreed to monitor the policy for any discriminatory impact, and to annually report on their findings. Unfortunately, these reports have not been prepared annually. It took CPS four months and one letter from the Illinois Attorney General to produce a response to our request under FOIA for the reports. We were disappointed with the one-page document that we received (attachment E). We were also deeply disturbed that our cursory analysis of the data clearly showed a continued disparate impact of the policy.

Some assert that standardized tests scores are the only “objective” measures of student progress, and so are educationally necessary. Education experts disagree. In 2004, the Joint Organizational Statement on NCLB was developed which is currently supported by 151 education, civil rights,

and civic organizations across the nation. The Joint Statement calls for the use of multiple measures which could include classroom, school, district and state tests; extended writing samples; tasks, projects, performances, and exhibitions; and selected samples of student classroom work, such as portfolios. Gathering this rich information would enable states, communities, schools, parents, teachers and students to know more about student learning and better improve schools. In addition, using such high-quality information could allow states to test less frequently, as many states did before NLCB.^{xii}

Parents United for Responsible Education (PURE) is a parent-organized, parent-run public school advocacy group established in 1987 and based in Chicago. PURE's overall goal is to assure a high-quality education for all children. Our main strategy is to support active, informed, meaningful parent participation in the public schools. PURE has a special role in focusing on issues from the parents' point of view. PURE's membership and constituency are multiracial, multi-cultural and economically diverse.

- i Commission on Improving Classroom-based Assessment. 2003. *Enhancing Teaching and Improving Learning: A Proposed System of Curriculum-Based Assessment for the Chicago Public Schools*.
- ii *Stanford Achievement Test Series, Ninth Edition: Guide for Organizational Planning* Harcourt Brace Educational Measurement. 1997. Pp. 43-44.
- iii Letter of May 11, 2009, to PURE executive director Julie Woestehoff from Marcilene Dutton, Deputy General Counsel, Illinois State Board of Education.
- iv U. S. Department of Education, *"Taking Responsibility for Ending Social Promotion."* 1999. p. 19
- v *Stanford*.

[Professional Testing Practices for Educators ISAT – 2009](http://www.isbe.state.il.us/assessment/pdfs/2009/Prof_Testing_Prac.pdf), Posted 11/18/08. Available at http://www.isbe.state.il.us/assessment/pdfs/2009/Prof_Testing_Prac.pdf

- vii National Research Council, *High Stakes: Testing for Tracking, Promotion, and Graduation*, 1999. Washington, D.C.: National Academy Press. p. 3.
- viii *Standards for Psychological and Educational Testing*. 1999. American Psychological Association, Association for Educational Research and Assessment, National Council on Measurement in Education
- ix *Code of Fair Testing Practices in Education*. (2004). Washington, DC: Joint Committee on Testing Practices. (Mailing Address: Joint Committee on Testing Practices, Science Directorate, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242; p.9.
- x *Stanford*.
- xi Letter from Elizabeth Calhoun, Freedom of Information Officer, Chicago Public Schools, to Julie Woestehoff, dated June 16, 2009.
- xii *Joint Organizational Statement on No Child Left Behind (NCLB) Act*. October 21, 2004.

Denver – General Assessment Meeting

Tuesday December 1, 2009

Public Speaker Testimony

Jim Ysseldyke, PhD

University of Minnesota



Thank you for the opportunity to testify. My name is Jim Ysseldyke and I am Birkmaier Professor of Educational Psychology and Director of the School Psychology Program at the University of Minnesota. For 35 years I have directed major federally funded research centers focused on ways to use assessment information to enhance the competence of individual students and to build the capacity of systems to meet individual student needs. My research is targeted specifically at improving results for struggling students, especially students with disabilities, and struggling schools.

I wish to make four points:

- 1. There is a fundamental disconnect between the stated intent/purpose of the proposed assessment development activity and the criteria specified for the assessments that are to be developed.**
- 2. We need a balanced assessment system that includes ongoing progress monitoring and formative assessment in addition to proposed summative assessments.**
- 3. We should take advantage of advances in technology-enhanced assessments.**
- 4. Formative and summative assessments should consist primarily of multiple choice items.**

In my remarks I focus on matters that can be dealt with **now**, over the next two years rather than assessment development activities that will take much longer. Brian Gong, Randy Bennett and others commented about this at the Boston and Atlanta meetings and I think it deserves emphasis. We need to focus on what we can implement **now** to improve student learning.

THE DISCONNECT

The stated goal of the proposed assessment activity is that the information gathered should be useful in influencing teaching, learning, and program improvement. The desired end of the assessment development process is improved educational outcomes for all students, including students with disabilities and English Language Learners.

There is a well-confirmed knowledge base on the many important components of effective instruction that must be in place to attain enhanced student performance. And, there is consensus that the three most important are:

1. Instructional match (matching instruction to the skill development of the learner),
2. Academic engaged time with extensive relevant practice (first guided, then independent) along with direct immediate feedback, and
3. Ongoing progress monitoring and use of the data to adapt instruction. Teachers need lots of information to provide differentiated instruction and make adaptations on the fly.

Assessment plays a critical role in each of the above. Yet, the federal register announcement calls for development and implementation of summative assessments. There is widespread consensus in the professional literature and in the assessment community that summative assessments do not and cannot inform instruction. They are not intended to do so. They serve an accountability purpose. Summative information is too little/too late for making a difference in instruction.

THE NEED FOR A BALANCED ASSESSMENT SYSTEM THAT INCLUDES PROGRESS MONITORING AND FORMATIVE ASSESSMENTS

What teachers and administrators need most is information during instruction, information that will enable them to make adjustments.

The kinds of assessments that contribute most to instructional improvement are those we used to call mastery measures or curriculum-based measures and now refer to under the umbrella term “formative assessment”. The approaches all entail data-driven decision making and are now an important part of the very important framework for school improvement called “Response to Intervention” (RTI).

In our research we have learned, for example, that there is a 9-year range in academic performance in math of the 6th grade students enrolled in the Minneapolis Schools. Teachers have an enormous task in matching instruction to student skill level and providing differentiated instruction for such a diverse group. Yet, this can be done using computer adaptive tests to pinpoint skill level, using existing technology to match level of instruction to individual skill level, and using existing progress monitoring and instructional adaptation procedures.

TAKE ADVANTAGE OF ADVANCES IN TECHNOLOGY

Advances in technology now enable us to do a more efficient and effective job in assessment. To improve instruction, we need more interim, benchmarking, and formative assessments. At the same time we need to reduce the assessment burden on teachers. The only way these two goals can be met is computer testing, and in particular computer adaptive testing or CAT. Existing low-cost assessment technology already in use in thousands of schools is helping teachers gather more information in less time and make instructional decisions that our research consistently shows improves instructional outcomes for all students.

FORMATIVE AND SUMMATIVE ASSESSMENTS SHOULD CONSIST PRIMARILY OF MULTIPLE CHOICE ITEMS

Multiple choice items have been shown to be reliable, valid, inexpensive, short, saving teacher time and able to test critical thinking skills as well or better than other types of items. Yet, there is a history of attempts to do away with multiple choice items and replace them with constructed response items, and the history of such efforts is not great. Constructed response and performance tasks are in vogue from time to time, but never have staying power because they are not sustainable or scalable. They also are not amenable to the accommodations that are required for accurate assessment of students with disabilities and English Language Learners. Attention also should be paid to efficiency. In doing so, consider the findings of more than 80 years of research on item types showing that 3 response multiple choice items are technically comparable to those providing 4 response alternatives (Rodriguez, 2006).

The call is for development of innovative assessments which is fine. We need innovation, but more importantly we need to implement interim, progress monitoring, screening, and formative assessments now. Students can't wait, teachers can't wait, and most certainly elected officials

can't wait five to ten years for supposedly better assessment technologies to be invented. Multiple choice items are proven. If we want anything that will improve student achievement in the next few years, there really is only one way to do that and that is to use multiple choice item types., in technology-enhanced progress monitoring, interim and formative assessments.

Contact Information

Jim Ysseldyke, PhD
Birkmaier Professor of Educational Leadership
University of Minnesota
Department of Educational Psychology
342 Education Sciences Building
56 East River Road
Minneapolis, MN 55455

Email: jim@umn.edu
Office Phone: 612-624-4014
Home Phone: 763-792-9849

Resources for Ysseldyke Testimony

- Burns, M.K., Klingbeil, D. & Ysseldyke, J.E. (In Press). The effects of technology enhanced formative evaluation on student performance on state accountability math tests. *Psychology in the Schools*.
- Rodriguez, M.C. (2006). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*. Pp. 3-13
- Spicuzza, R., Ysseldyke, J. E., Lemkuil, A., McGill, S., Boys, C., & Teelucksingh, E. (2001). Effects of curriculum-based monitoring on classroom instruction and math achievement. *Journal of School Psychology, 39*(6), 521-542.
- Ysseldyke, J. E., Betts, J., Thill, T., & Hannigan, E. (2004). Use of an instructional management system to improve mathematics skills for students in Title I programs. *Preventing School Failure, 48*(4), 10-14.
- Ysseldyke, J. & Bolt, D. (2007). Effect of technology-enhanced continuous progress monitoring on math achievement. *School Psychology Review, 36* (3), 453-467.
- Ysseldyke, J.E., Burns, M.K., Scholin, S. & Parker, D. (In Press). Instructionally valid assessment within RTI. *Teaching Exceptional Children*.
- Ysseldyke, J., & Tardrew, S. (2007). Use of a progress-monitoring system to enable teachers to differentiate math instruction. *Journal of Applied School Psychology, 24*(1), 1-28.
- Ysseldyke, J. E., Tardrew, S., Betts, J., Thill, T., & Hannigan, E. (2004). Use of an instructional management system to enhance math instruction of gifted and talented students. *Journal for the Education of the Gifted, 27*(4), 293-310.
- Ysseldyke, J.E., Spicuzza, R., Kosciolik, S., & Boys, C. (2003). Effects of a learning information system on mathematics achievement and classroom structure. *Journal of Educational Research, 96*(3), 163-173.
- Ysseldyke, J. E., Spicuzza, R., Kosciolik, S., Teelucksingh, E., Boys, C., & Lemkuil, A. (2003). Using a curriculum-based instructional management system to enhance math achievement in urban schools. *Journal of Education for Students Placed at Risk, 8*(2), 247-265.

Testimony for US Dept of Education
The Role of Public Service Media in a National Assessment System



I am presenting testimony today as both a former Superintendent of Instruction for 10 years in the State of Ohio and the Senior Vice President for Education for the Corporation for Public Broadcasting (CPB), where I serve as chief education policy advisor and consultant to the public service media system. It is my pleasure to provide comments to the U.S. Department of Education regarding the proposed Race to the Top (RTTT) assessment initiative. The Corporation for Public Broadcasting is a private, non-profit corporation that was created by Congress in 1967. It promotes universal access to public telecommunications services (television, radio, and on-line) by supporting over 1100 radio and television stations across America. CPB has a long and well documented record of funding for diverse and innovative educational programming that is second to none. However, beyond programming, public service media helps teachers, caregivers, parents, and communities educate children. CPB is a strong ally in raising the academic bar and closing achievement gaps for all students, particularly the underrepresented and underserved.

Because I have been a policy leader in both public education and now public service media, I understand how our publicly funded television and radio stations can enhance a national system of student assessment. A national assessment system can provide for the better integration of curriculum, instruction, assessment, and educator development, and public service media can provide the digital content and technological know-how to assist with this innovative and digitally-based system.

As state superintendent, I wanted a coherent, comprehensive assessment system that assured that all students had the opportunity to learn. In Ohio, we saw a future assessment system that was built upon clear and succinct academic content standards that incorporate 21st century skills such as problem

solving, innovation, and collaborative learning. We envisioned performance tasks embedded in mini-curricular units that would be crafted to allow teachers to individualize learning opportunities for students through individualized student. We began to craft a focused, professional development system for superintendents, principals, teachers and parents to better understand and participate in the development of this new assessment process. We set up a system that provided differentiated reports to superintendents, principals, teachers, students and parents to report districts', schools' and students' strengths and weaknesses and improve professional practice for educators and give reliable information to parents to support their children's learning.

As State Superintendent of Public Instruction, I saw how funding constraints limited Ohio's opportunity to develop formative and summative assessment systems that could use multiple measures such as portfolios and performance based assessments. Through a grant from the Gates and Hewlett Foundations, Ohio is now working with Stanford University and 27 school sites to develop performance assessment tasks, with strong statistical validity and reliability systems modeled after the moderations panels found in Queensland, Australia, Finland, and other higher performing countries.

To show the value of public service media on a national assessment system, I would like to address:

- Technology and Innovation in Assessment
- Project Management and National Consortia

Public service media has rich and trusted digital content such as video and audio programming, online games, simulations, podcasts and other digital learning objects that allow for multiple representations of the same concepts. These resources, much of which is in the public domain, motivate and engage the audience. Some of this content has been subject to rigorous evaluations that demonstrate its efficacy for enhancing the learning outcomes of poor and underserved children. Our public media system is in the process of aligning this content with academic standards through the PBS Digital Learning Library

and CPB's American Archive program. In addition, we can customize digital learning objects for assessment projects. These resources can be used for performance tasks, performance portfolios, constructed responses and essays. Our content can blend both academic and technical studies and test subject matter competency, and can create tasks that stress habits of mind for collaboration, design, invention, and entrepreneurship – skills essential for success in the 21st century.

The public broadcasting system has a long history of educator development. Systems like PBS TeacherLine, ThinkPort (Maryland Public Television), E-Learning for Educators (a collaboration of southern and mid-western stations as well as Delaware and New Hampshire) and Teachers' Domain (WGBH, Boston) are but a few examples. Our system is and can be even more helpful in facilitating teacher and administrator training in assessment literacy. Specifically, our system can provide online training on developing items for formative and summative assessments, scoring of performance tasks, the interpretation and use of results for all types of assessments, and the creation of curricular materials that can be shared electronically within and across districts, schools, and states. Our system can also provide important information to parents and community leaders. Our stations are community based and are experienced in convening stakeholders around a host of educational issues and facilitating both professional and social networking.

Public broadcasting is now experimenting with new digital media, such as I-pods, cell phones, mobile TV, and other handheld devices that can be of service to test developers as they continuously adapt to the ever-changing technology. Our local stations have experience using adaptive technologies with individuals with special needs. A leader in this area is WGBH in Boston. Public service media can be a valuable partner in multiple consortia that will provide differentiated assessment models for special populations and be an active participant in the development, design, research and evaluation of a national assessment system. We have a close working relationship with the Council of Chief State

School Officers, the Council of Great City Schools, and the Partnership for 21st Century Skills. We are a knowledgeable and cost-effective public partner in a system that holds the promise to improve instruction for all students and holds everyone accountable for results.



Listening. Learning. Leading.®

Growth in Student Achievement: Issues of Measurement, Longitudinal Analyses & Accountability

Damian W. Betebenner
National Center for the Improvement
of Educational Assessment

Robert L. Linn
University of Colorado, Boulder

Dr. Betebenner gratefully acknowledges The Center for Assessment, the Colorado Department of Education, and the Massachusetts Department of Elementary and Secondary Education for their support of this work.

Dr. Linn gratefully acknowledges the National Center for Research on Evaluation, Standards and Student Testing and the University of Colorado at Boulder for their support of this work.

**K – 12 Assessment and
Performance Management Center**

An independent catalyst and resource for the improvement of
K – 12 assessment and performance management systems

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

OVERVIEW

Overview

Data, Data, Everywhere . . . and not a drop of information

Overview

- ❑ Enhanced data acquisition and management has enabled:
 - ❑ Historical records of student achievement
 - ❑ Historical records of student demographics, teachers, schools, educational programs ...
 - ❑ Stakeholder interest in an examination of this longitudinal data
- ❑ Interest in examining student achievement over time (student growth) derives from data availability

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

What is growth and why measure it?

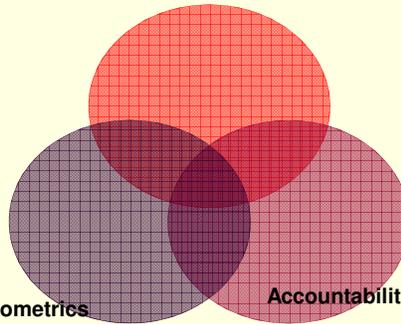
Overview

- ❑ Student learning is a central goal of education
- ❑ Assessments of student achievement provide evidence of the current status of student knowledge and understanding
- ❑ Learning is demonstrated by growth in student achievement from one point in time to another point in time – not by status at either point time alone

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Discussions of student growth lie at the intersection of three topics

Longitudinal Data Analysis/Applied Statistics



Measurement/Psychometrics

Accountability/Education Policy/Data Use

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Overview

Measurement/Psychometrics

Examining student growth requires multiple
measurement of the same individual

- Growth in what?
- How much growth? (How is scaling involved in answering this question?)
- Is it enough growth?

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Overview

Longitudinal Data Analysis/Applied Statistics

Overview

Many methods for analysis of longitudinal data

- ❑ What are the relevant questions?
- ❑ Are the analytic techniques capable of answering those questions?
- ❑ Does the data possess properties sufficient for the analytic techniques employed? (e.g., vertical scale)
- ❑ Does the analysis sustain the inferences made from the data?

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Accountability/Education Policy/Data Use

Overview

Education Policy & Accountability have many goals and purposes

- ❑ Why growth in accountability?
- ❑ What are the goals and purposes of accountability?
- ❑ What is the theory of action behind accountability?
- ❑ How can we judge the validity of the accountability system?
- ❑ What about the current policy context?

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

TECHNICAL CONSIDERATIONS



Three Intersecting Issues

Student Growth brings together related issues from three areas:

- measurement/psychometrics
- longitudinal data analysis/applied statistics
- accountability/education policy/data use

Technical Considerations

Measurement/Psychometric Issues

- ❑ Growth in what?
- ❑ How much growth?
- ❑ Scales for measuring growth
 - ❑ Ordinal (within-year, across year)
 - ❑ Interval (within-year, across year)
 - ❑ Vertical
- ❑ Growth magnitude versus growth norm
- ❑ Is it enough growth? Norm- versus criterion-referencing (intersection of Accountability and Measurement)

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations

Growth in what?

- ❑ Beneath any notion of change (i.e., growth) is a construct that is changing over time
- ❑ Height and weight are common points of reference
- ❑ Constructs in education are “slippery”
- ❑ Need, at a minimum, an underlying semantical referent (e.g. reading or math)

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations

How much growth?

- ❑ Are growth magnitudes possible in education?
- ❑ If calculable, are they interpretable absent some norm?
- ❑ Approaches to growth magnitudes:
 - ❑ Performance standards
 - ❑ Vertical scale with interval properties
 - ❑ Learning progressions (qualitative growth)

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations

How much growth?

Performance Standards

Strengths	Limitations
<ul style="list-style-type: none"> ❑ Anchors reference points for discussions about performance ❑ Growth is embedded in accountability metric 	<ul style="list-style-type: none"> ❑ Few levels, mask substantial range within levels thus masking student growth within level ❑ Vary greatly in stringency from state to state so that “proficient” performance lacks meaning

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations	<h2>How much growth?</h2> <p>Scale Scores</p>	
	<h3>Strengths</h3> <ul style="list-style-type: none"> ❑ Semi-continuous scores (many score points) ❑ Can be used to create vertical scales across grade levels ❑ Give the appearance of interval scales needed by some analytical models 	<h3>Limitations</h3> <ul style="list-style-type: none"> ❑ Difficult to interpret or explain to users ❑ Vertical scales are hard to defend ❑ Claims of interval measurement properties don't hold to close scrutiny
Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.		

Technical Considerations	<h2>How much growth?</h2> <p>Vertical Scale</p>	
	<p>Vertical & Interval scales required for some analytic techniques:</p> <ul style="list-style-type: none"> ❑ Gain score calculation (magnitude of growth) ❑ Growth curve analysis (rate of growth) (e.g., Willett & Singer, 2003) <p>Vertical & Interval scales required for some questions:</p> <ul style="list-style-type: none"> ❑ Matthew effects: Do higher achievers grow faster than lower achievers? ❑ Growth rates relative to student age: Do students grow more in later grades than earlier grades? 	
Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.		

Technical Considerations

How much growth?

Vertical Scale

Vertical and/or Interval scales NOT required for some analytic techniques:

- Value-Added analyses: Most require interval, but not vertical, scale. See Ballou (2008), Briggs & Betebenner (2009).
- Auto-regressive analyses, growth norms

Vertical and/or Interval scales NOT required for some questions:

- Is a student's progress (ab)normal?
- Is a student's growth sufficient to put them on track to reach/maintain proficiency?
- See Yen (2007) for an excellent list of questions

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations

How much growth?

Magnitudes versus Norms

Physical growth

- 9 year old boy grew 5 inches in past year
- Average increase in height for boys between years 8 and 9 is 4 inches

Achievement growth

- 4th grader grew 25 scale score points since 3rd grade
- Average 4th grade scale score is 21 points higher than average 3rd grade score

Two Growth Quantities

- Magnitude of growth
- Relative amount of growth

How much growth?

- People expect an answer of magnitude
- People need magnitude embedded within a norm

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations

How much growth?

Growth norms

Although normative comparisons are spurned by criterion-referenced and standards-based measurement advocates, norms can provide a useful interpretive framework, especially in the interpretation of student growth

“Scratch a criterion and you find a norm”
W. H. Angoff (1974)

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations

How much growth?

Norm- and Criterion-Referenced

- ❑ Defining enough growth is a standard setting procedure
- ❑ Growth standard setting should be informed, at least in part, by norms
- ❑ Criteria superimposed over norms provide a transparent and fair mechanism to communicate simultaneously:
 - ❑ What is?
 - ❑ What should be?
 - ❑ What is reasonable?

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations

Longitudinal Data Analysis Issues

Technical Considerations

Many Questions

- ❑ How much annual growth did this (these) student(s) make in reading?
- ❑ Is (Are) this (these) student(s) making sufficient growth to reach/maintain desired achievement targets? (Growth-to-standard & Growth Model Pilot Program)
- ❑ Are students in particular subgroups (e.g., minority students) making as much progress as other students?
- ❑ How much did this teacher/school contribute to students' growth over the last year? (Value-Added)
- ❑ Again, see Yen (2007) for an excellent list of questions

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Many Techniques

Numerous data analysis techniques for use with longitudinal data:

- ❑ Gain scores (suitable scale required)
- ❑ Cross-tabulation based upon prior and current categorical achievement level attainment (e.g., value-tables, transition matrices)
- ❑ Regression based approaches: growth-curve analysis (HLM), fixed/mixed-effects models, growth norms

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Questions 1st, Analyses 2nd

- ❑ Different growth analysis techniques often address different questions
- ❑ Different questions lead to different conversations which lead to different uses and outcomes

“It is better to have an approximate answer to the right question than a precise answer to the wrong question.”

J. W. Tukey

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations

Model Purpose

Three general uses associated with statistical models (Berk, 2004):

Description: An account of the data. Model is true to the extent that it is useful. Model quality judged by craftsmanship (de Leeuw, 2004)

Inference: Sample to Population. Model is true to the extent that the assumed chance process reflects reality (super-population fallacy)

Causality: *A* causes *B* to happen. Model is true to the extent that plausible causal theory exists and design criteria are met

- Models are rarely descriptive despite minimal requirements
- Inference and causality require information external to the data. Can't be validated solely from data
- Models are often causal in nature but rarely meet rigorous criteria necessary for such inferences

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations

Value-Added Models

Causality

- Value-Added Models (e.g., EVAAS) are a frequently discussed type of growth model
- Value-Added Models attempt to quantify the portion of student progress attributable, usually to a teacher or school
- Value-Added is about the inferences made and not the actual model
- Causal attributions make value-added models well suited for accountability discussions
- In the absence of random assignment causal attributions are always suspect and subject to challenges (see, for example, Raudenbush, 2004; Rubin, Stuart & Zanutto, 2004)

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations

Value-Added Models

Causality

- ❑ Value-added models return norm-referenced effectiveness quantities
- ❑ With regard to schools, quantities indicate whether a school is significantly more or less effective than the mean school effectiveness in the district or state
- ❑ In a standards based assessment environment, how much effectiveness is enough?
- ❑ Especially important in light of universal proficiency policy mandates
- ❑ Growth-to-standard models created to provide criterion-referenced growth models

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations

Growth Model Pilot Program

Growth-to-standard

- ❑ In response to requests for growth model use as part of AYP, USED allowed states to apply to use growth models
- ❑ Fifteen states had models accepted
- ❑ Models required to adhere to the “bright line principle” of universal proficiency (growth-to-standard)
- ❑ Yen (2009) provides an excellent overview of the models
- ❑ Growth-to-standard models returned, in general, results that closely aligned with AYP status results.

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations

Growth versus Value-Added Models

Description & Causality

- ❑ Growth measures are descriptive
- ❑ Accountability has skewed discussions of growth from description toward responsibility (i.e., causality)
- ❑ All measures (even VAM) are potentially descriptive. However, some measures are specially crafted for causal inference/attribution
- ❑ Good descriptive measures are interpretable, informative and capable of multiple uses

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations

Growth versus Value-Added Models

Description: Colorado Growth Model

- ❑ The Colorado Growth Model uses student growth percentiles to quantify student growth
- ❑ Percentiles are familiar to stakeholders
- ❑ Separating description from responsibility has led to broad public acceptance including teacher's unions
- ❑ Asking what schools or teachers are associated with students demonstrating the highest growth percentiles moves from description toward value-added

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations

Growth versus Value-Added Models

Description: Colorado Growth Model

- ❑ Analysis employs quantile regression to calculate conditional quantile relationships between current and prior achievement
- ❑ Student growth percentiles can also be criterion referenced to accommodate growth-to-standard
- ❑ This approach formed the basis of Colorado's successful application as part of the Growth Model Pilot Program
- ❑ Student growth percentiles provide a bridge connecting value-added and criterion-referenced interests

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations

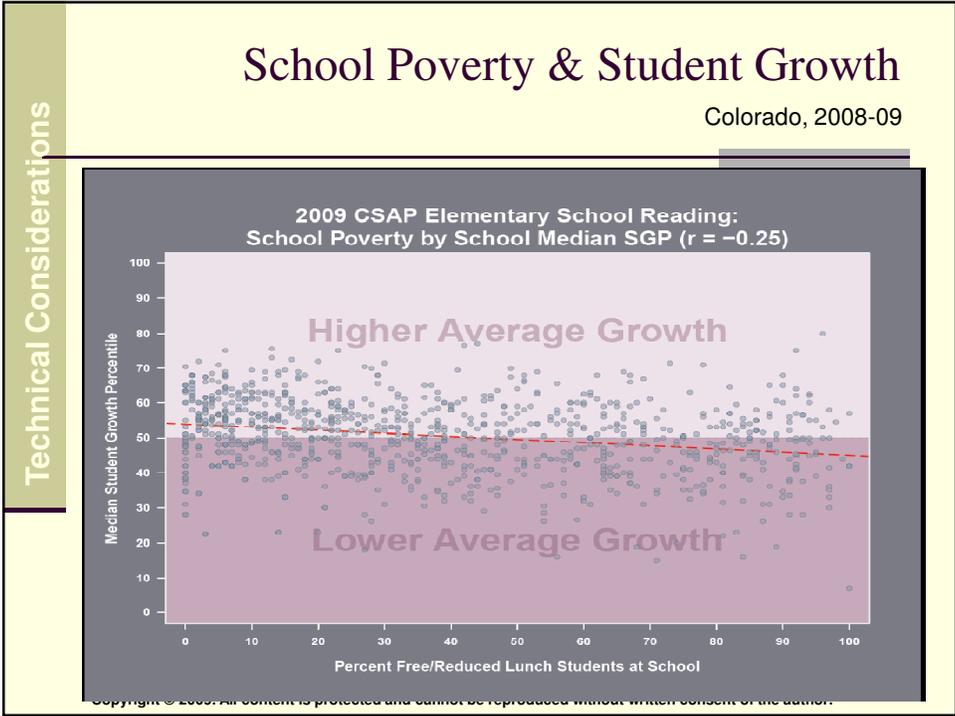
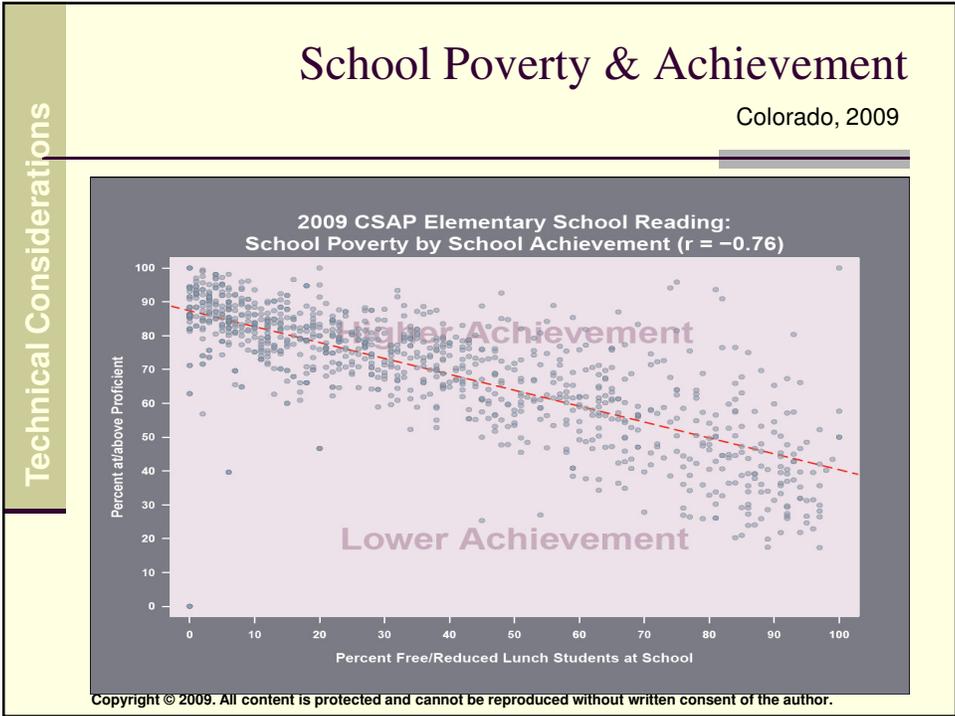
Validating Models

- ❑ There is no "gold standard" against which to judge value-added or growth model results
- ❑ Statistical model specification goes only part way toward validation
- ❑ Results should have face validity
- ❑ Because of their importance in accountability, utility is a primary component of model validity

"All models are wrong but some are useful."

G. E. P. Box

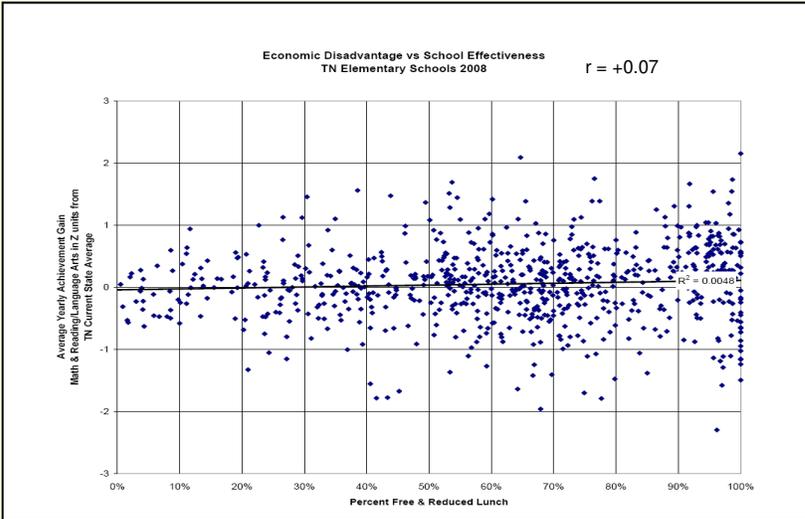
Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.



Technical Considerations

School Poverty & Value-Added

Tennessee, 2007-08



Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations

Accountability/Policy/Data Use

Technical Considerations

Accountability and Growth

Why growth instead of status?

- ❑ Enthusiasm for growth in accountability stems from the belief that growth and teacher/school quality are more closely related
- ❑ Enthusiasm for growth also stems from its potential diagnostic uses
- ❑ How do we judge the use of growth related measures within an accountability system?
- ❑ What are the features of a valid accountability system that uses student growth?

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations

Accountability and Growth

What type of growth?

- ❑ Value-added provides a norm-referenced lens judging growth/effectiveness against district/state averages.
- ❑ Growth-to-standard provides a criterion-referenced lens judging growth toward community endorsed achievement goals
- ❑ Inferences about education quality based upon value-added make judgments relative to students reaching statistical expectation
- ❑ Inferences about education quality based upon growth-to-standard make judgments relative to students reaching criterion-referenced destinations

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations

Accountability and Growth

What type of growth?

- ❑ States currently employ a variety of growth models in service of accountability
- ❑ Current policy mandates like NCLB are criterion-referenced---establishing achievement targets/destinations for all students
- ❑ Need BOTH norm- and criterion-referenced growth to reconcile individual focused policies like NCLB with imperatives to judge education quality at the group level (e.g., teacher or school)

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations

Accountability Systems

Purpose and requirements

- ❑ Intended to improve education
 - ❑ Increase student achievement
 - ❑ Reduce achievement gaps
 - ❑ Increase efficiencies
- ❑ Externally mandated and designed to hold educators responsible for student learning
- ❑ Impose sanctions and rewards based upon results from large scale assessment outcomes

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations

Accountability Systems

Theory of action

- ❑ Theory of action connects interpretations, uses, and consequences (Gong, 2008)
- ❑ Details connecting punishments/rewards and outcomes are usually vague/incomplete
- ❑ It is exactly these details that are critical to validating the theory of action associated with the accountability system's use of growth any growth/value-added metric

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations

Accountability System Validity

Systemic Validity

“Assessment practices and systems of accountability are systemically valid if they generate useful information and constructive responses that support one or more policy goals (Access, Quality, Equity, Efficiency) within an education system, without causing undue deterioration with respect to other goals.”

H. Braun (2008)

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations

Accountability System Validity

Systemic Validity

- ❑ Through careful consideration, assessment and accountability systems can be engineered to maximize systemic validity.
- ❑ Requires meticulous pre-specification of the desired “useful information and constructive responses”
- ❑ This anticipates and incorporates many of the lessons learned about unintended consequences

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations

Descriptive Accountability

Building in systemic validity

“Accountability system results can have value without making causal inferences about school quality, solely from the results of student achievement measures and demographic characteristics. Treating the results as descriptive information and for identification of schools that require more intensive investigation of organizational and instructional process characteristics are potentially of considerable value. Rather than using the results of the accountability system as the sole determiner of sanctions for schools, they could be used to flag schools that need more intensive investigation to reach sound conclusions about needed improvements or judgments about quality.”

R. L. Linn (2008)

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations

Descriptive Accountability

- ❑ Part of a broader research program
- ❑ Helpful in spotting provocative associations
- ❑ A part of advocacy/informative discussions (e.g, growth-gaps by ethnicity)
- ❑ Informs policy goals and initiatives

The descriptive growth norms of the Colorado Growth Model are an example of this type of accountability metric

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Technical Considerations

Policy Context

Race To The Top

Differentiating teacher and principal effectiveness based on performance.... The extent to which the State, in collaboration with its participating LEAs, has a high quality plan and ambitious yet achievable annual targets to (a) Determine an approach to measuring student growth (as defined in this notice); (b) employ rigorous, transparent, and equitable processes for differentiating the effectiveness of teachers and principals using multiple rating categories that take into account data on student growth (as defined in this notice) as a significant factor; (c) provide to each teacher and principal his or her own data and rating; and (d) use this information when making decisions. (Section III.C.(C)(2), p. 37809)

- Race-to-the-top has embraced the use of large scale assessment to make high stakes judgments (or has it?)
- Notice the terms teacher and principal effectiveness

Accountability 2.0

Recommendations

Recommendations

"They say that genius is an infinite capacity for taking pains. It's a very bad definition, but it does apply to detective work."

Sherlock Holmes,
in "A Study in Scarlet"

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

We've heard it all before

Recommendations

❑ Despite admirable goals, high stakes accountability based on large scale assessment outcomes has unintended negative consequences (Linn, 2000, Mintrop & Sunderman, 2009):

- ❑ Narrowed curriculum
- ❑ Gaming the system (e.g., bubble students)
- ❑ Emphasis on test preparation
- ❑ Frustration and de-moralization of those most involved with school improvement efforts

Campbell's Law: The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor.

D. T. Campbell (1976)

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Stepping Back

Some Axioms

Recommendations

- ❑ Distance = Rate x Time

When establishing a common destination for all students, those students starting further from the destination must travel “faster” or “longer” to reach the destination.

- ❑ The Ultimate Goal

We want to witness increases in rates of growth/effectiveness. It’s questionable whether current measurement instruments have the precision to detect such system changes

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Stepping Back

Growth begets more data

Recommendations

- ❑ Growth analyses just produce more data for organizations already drowning in data
- ❑ The data do not speak for themselves and can tell a thousand “stories”!
- ❑ Stories currently told are usually overly simplistic and/or just plain wrong
- ❑ What stories form the basis for continuous improvement?
- ❑ Start at the end with the important stories and develop a growth model from there.

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Stepping Back

Some Basic Questions

Recommendations

- ❑ Who uses the data to improve education?
 - ❑ Administrators, policy makers, researchers (the elites)
 - ❑ All stakeholders (teachers, principals, parents, administrators, policy makers, & researchers)
- ❑ How should the data be used to improve student achievement?
 - ❑ Different stakeholders have different interests
 - ❑ Same top-down theory of action (now fortified with growth) or a paradigm shift?

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Paradigm Shift

Marshalling a consensus for change

Recommendations

“This is the difference between a retrospective question of identifying fault as opposed to a prospective strategy to engineer some corrective measure, almost independent of considering whether there was blame-worthiness. And to move away from the blame-worthiness paradigm toward something that is more regulatory in nature where one might seize upon disparities or circumstances that are for some reason deemed unacceptable and engineer the interventions needed to bring about the necessary change. . . . It’s the no-fault gap closing strategy in which the effort is to build a consensus about a vision of an improved society rather than figure out where’s the person we want to pillory.”

C. Edley (2006)

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Paradigm Shift

Marshalling Collaboration

Recommendations

- ❑ Examining data (broadly) is critical to making sound judgments.
- ❑ What data and stories marshal the consensus for change
- ❑ Overreliance on sanctions can be reduced when policies aim to develop a partnership between government, teachers, and parents, and motivate changes by adhering to the professional values and standards of educators (Mintrop & Sunderman, p. 9).
- ❑ Data visualization is a critical component of developing partnerships and marshalling consensus.

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Paradigm Shift

Collaborative Data Visualization

Recommendations

With a collaborative spirit, with a **collaborative platform** where people can **upload data, explore data, compare solutions, discuss the results, build consensus**, we can engage passionate people, local communities, media and this will raise—incredibly—the amount of people who can understand what is going on.

And this would have fantastic outcomes: the **engagement of people**, especially new generations; it would **increase knowledge, unlock statistics, improve transparency** and accountability of public policies, **change culture, increase numeracy**, and in the end, **improve democracy and welfare**.

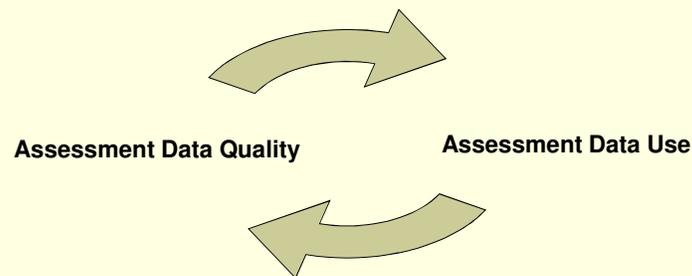
E. Giovannini, Chief Statistician, OECD. June 2007

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Data Quality & Data Use

Recommendations

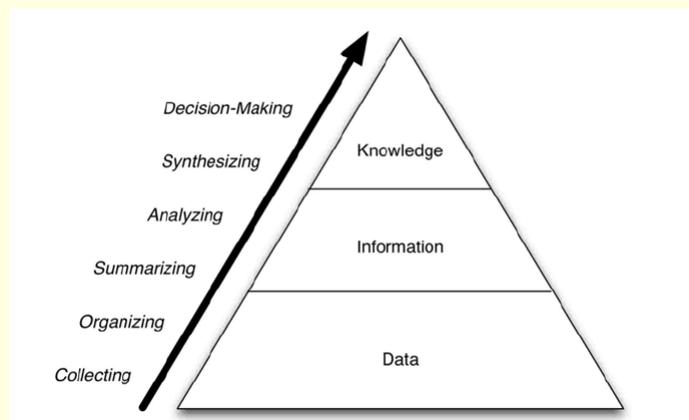
- ❑ Better quality data doesn't necessarily lead to improved use.
- ❑ However, informed use often leads to demands for higher quality data.



Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Data to Knowledge Continuum

Recommendations



Breiter & Light (2006)

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Reporting and Data Visualization

- ❑ Multiple sources of data are required to make sound inferences.
- ❑ The goal is to marshal multiple sources of data in service of:
 - ❑ Individual learning diagnosis and prognosis
 - ❑ Advocacy & consensus building
 - ❑ Program oversight
- ❑ Data visualization forms the basis for collaborative story telling.
- ❑ Experts understanding the “right” stories must participate in the construction of collaborative visualization platforms allowing stakeholders to explore these stories
- ❑ Collaborative story telling allows users to relate and share data of their experiences (e.g., a superintendent and their schools or a teacher and their students)
- ❑ Complicated stories require complicated and multifaceted visualizations.
- ❑ Like with any product development, METICULOUS design is essential to maximizing utility
- ❑ Without dedicated initiatives to promote good data use---synthesizing data from various sources---simplistic (and likely incorrect) data uses will persist
- ❑ We’re all detectives looking at evidence and must show “an infinite capacity for taking pains”

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Bibliography

- ❑ Angoff, W. H. (1974). Criterion-referencing, norm-referencing and the SAT. *College Board Review*, 92:2–5.
- ❑ Ballou, D. (2008). Test Scaling and Value-Added Measurement. Presented at *National Conference on Value-Added Modeling*, Madison, Wisconsin.
- ❑ Berk, R. A. (2004). *Regression analysis: A constructive critique*. Sage Publications, Thousand Oaks, CA.
- ❑ Braun, H. (2008). Vicissitudes of the Validators. Presented at the Reidy Interactive Lecture Series, Portsmouth, New Hampshire.
- ❑ Breiter, A & Light, D (2006). Data for School Improvement: Factors for designing effective information systems to support decision-making in schools. *Educational Technology & Society*, 9 (3), 206-217.
- ❑ Briggs, D. C. M. & Betebenner, D. W. (2009) Is growth in student achievement scale dependent?. Paper presented at the NCME annual conference, April 2009. San Diego, CA.
- ❑ Campbell, D. T. (1976), *Assessing the Impact of Planned Social Change*. The Public Affairs Center, Dartmouth College, Hanover New Hampshire, USA. December, 1976.
- ❑ De Leeuw, J. (2004) Preface to Berk’s “Regression analysis: A constructive critique”. Sage Publications, Thousand Oaks, CA.

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Bibliography

Bibliography

- ❑ Downloaded January 20th, 2009 from <http://www.education-consumers.com/VAAA/Poverty%20vs%20School%20Effectiveness%20006%20chart.pdf>
- ❑ Downloaded January 20th, 2009 from http://www.education-consumers.org/tnproject/poverty_vs_effectiveness_2008.pdf
- ❑ Edley, C. (2006). Educational "Opportunity" is the highest civil rights priority. So what should researchers and lawyers do about it? Retrieved June 22, 2006 from the World Wide Web: <http://www.softconference.com/MEDIA/WMP/260407/#43.010>
- ❑ Gong, B. (2008). Validating accountability systems. Presented at the Reidy Interactive Lecture Series, Portsmouth New Hampshire
- ❑ Linn, R. L. (2000). Assessments and accountability. *Educational Researcher* 29(2): 4-16.
- ❑ Linn, R. L. (2008). Educational accountability systems. In K. E. Ryan & L. A. Shepard (Eds.), *The Future of Test-Based Educational Accountability*, pages 3–24. Taylor & Francis, New York.
- ❑ Mintrop, H. & Sunderman, G. L. (2009). *Why high stakes accountability sounds good but doesn't work—and why we keep on doing it anyway*. Los Angeles, CA: The Civil Rights Project/Proyecto Derechos Civiles at UCLA.

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Bibliography

Bibliography

- ❑ Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press, USA.
- ❑ Yen, W. M. (2007). Vertical scaling and No Child Left Behind. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and Aligning Scores and Scales*, pages 273–283. Springer, New York.
- ❑ Yen, W. M. (2009, March) Growth Models Approved for the NCLB Growth Model Pilot. ETS.

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Issues in Measuring Student Growth and Conducting Productivity Analyses

Henry Braun

Lynch School of Education
Boston College

Exploratory Seminar
ETS
December 7 2009

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

The State Landscape

- Policy
 - Financial exigencies
 - Responding to federal mandates/opportunities
 - Improving instruction and school leadership
- Assessments
 - Variable quality and weakly aligned to content standards
 - Design specs do not focus on growth – although much of current rhetoric references measures of growth
- Performance Standards
 - Variable quality and rigor
 - Generally uninterpretable
 - Poorly articulated across grades

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Federal Policy Environment

- R2T Initiative
 - Funding “radical” systemic change
 - Funding assessment innovation
 - Support for high standards and school choice
 - Call for improved accountability
- Reauthorization of ESEA
 - General unhappiness with NCLB
 - Likely interest in both status and growth indicators
 - No clear direction yet

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author. 3

Framing Assumptions

- Common Core Standards
 - Comprehensive
 - Focused
 - Multi-grade
- R2T Assessment Program
 - Funding for “next generation” of assessments
 - Multiple purposes: Informing both instruction and accountability
 - Competition among state consortia
 - Need to plan for a multi-stage implementation

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author. 4

Purpose

To consider the challenges in enhancing the utility of assessment systems for productivity analyses

Outline

- Assessment design
- Productivity analyses
- Value-added modeling
- Reflections and suggestions

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author. 5

Innovation in Assessment: Prerequisites

- Comprehensive model of each domain
- Models of student learning in the domain (pathways to expertise)
- High quality content standards that are vertically articulated
- Performance standards that are rigorous and vertically articulated

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author. 6

Innovation in Assessment: Assessment Design

- Integration of cognitive and developmental perspectives, in concert with “traditional” psychometric and logistic requirements
- Explicit targeting of a “growth construct” with appropriate cross-grade linkage
- Assessment system components
 - Interim probes (diagnostic)
 - Curriculum-embedded extended exercises (may be on-demand)
 - Summative on-demand with multiple formats
- Existing technology platforms could enable
 - New formats and test structures (e.g. adaptive testing)
 - Improved accuracy and more uniform precision
 - Faster turn-around for both formative and summative assessments

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author. 7

Assessment Design

- Greater challenges in balancing goals and constraints
 - What are trade-offs in obtaining improved cross-grade articulation and better measures of growth?
 - What are implications of different designs?
- Better measurement of growth presupposes theoretical and empirical understanding of pathways to mastery
 - In some domains, research fairly well along
 - Complexity in accommodating multiple pathways

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author. 8

The Measurement of Growth for Accountability: Some (Naïve) Questions

- How do we operationalize growth as a summary indicator of student learning?
 - Difference in scale scores
 - Change in score profiles
 - Progress along a developmental trajectory marked by discrete milestones
 - Conditional (relative) achievement
- How should measures of growth be validated?
 - Psychometric properties
 - Relationship to other measures of learning
 - Predictive power
 - Consequences (direct and indirect)

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author. 9

Productivity Analyses: Goals

- Generate evidence to inform system improvement with respect to both effectiveness and efficiency
- Hold individuals and units accountable for performance

Three questions for education systems:

1. *Where are students in relation to targets?*
2. *How much learning took place?*
3. *What was the (relative) contribution of X (teacher or school) to that learning?*

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author. 10

Productivity Analyses: Considerations

- Methodology should match both the questions and the type/quality of data available
 - e.g. If outcomes are represented on a discrete, ordinal scale, then a transition matrix can be the basis of the analysis
- Results can be referenced to specific targets or normatively defined
 - e.g. Indicator is compared to an absolute threshold
 - e.g. Indicator is compared to a reference distribution
- Simple indicators will always be attractive – despite technical flaws
 - e.g. Percent of students exceeding the proficiency standard

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author. 11

Productivity Analyses: Case Study (1)

Using Standards of Proficiency as Markers for Productivity

- NCLB uses changes in “percent proficient” for a fixed grade (cohort-to-cohort comparison)
 - Technically flawed
 - Can lead to misleading conclusions
- What about changes in “percent proficient” for a specific cohort (grade-to-grade comparison)?
 - Incoherent standards lead to misinterpretations
- If grade-level standards are explicitly (and appropriately) linked to learning trajectories, then longitudinal tracking of grade-level results could generate useful data for productivity analyses

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author. 12

Productivity Analyses: Case Study (2)

- Mastery of material at grade “ n ” is not an end in itself, but a milestone in a student’s trajectory through school.
- Some common-sense meanings of achieving *proficiency* in grade n are:
 - i. Student has met requirements for grade n
 - ii. All things being equal, the student has a high probability of achieving proficiency in grade $n+1$,
- Argues for cross-grade coherence in standard-setting

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author. 13

Productivity Analyses: Case Study (3)

A 3P Paradigm for Standard Setting

- **Prospective:** The domain model and agreement on competencies shape test development through the early specification of performance standards
- **Progressive:** Coordination in content frameworks and performance standards across grades
- **Predictive:** Descriptions of performance standards are explicitly based on theoretical and empirical evidence about trajectories of student learning and development

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author. 14

Productivity Analyses: Case Study (4)

- Starting point for 3P approach is the set of common core “college and career-readiness” standards.
- Develop K-12 standards through “backward” mapping from common core standards for college and career readiness -- so that “meeting” standards in earlier grades signals student is on-track for post high school readiness
- Requires both expert judgment and empirical analysis
- Challenge for assessment design when there are multiple standards for each grade

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author. 15

School Improvement vs. Accountability

- School Improvement
 - Low- to moderate-stakes
 - Growth indicators can be flexibly combined with other measures for evaluation
 - Local context provides richer information for interpretation and action
- Accountability (external)
 - High-stakes
 - Formal and rigid
 - Few indicators
 - Limited information for improvement

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author. 16

Value-added Modeling (VAM)

- Intent is to extract from measures of student learning trajectories the (average) component due to the teacher or unit (school, program, district)
- A step beyond tracking growth because it involves adjustments for selection bias
- Aggregation of individual level data places greater burdens on test design
- Results usually defined normatively
- For accountability, statistical descriptions (parameter estimates) are interpreted as causal effects – problematic when data are derived from an observational study rather than a randomized experiment

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author. 17

Some Approaches to VAM

1. Statistical models for scalar growth
2. Education production functions
3. Modeling of longitudinal achievement trajectories in a multi-level framework
4. Multivariate, longitudinal, mixed effects

**Only (1) and (3) refer to direct measures of growth.
Current research does not favor these approaches!**

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author. 18

Construction of a cross-grade vertical scale

- Requires choices among various methods
- Vertical scale characteristics depend on choices
- Interval scale property difficult to support
- Vertical scale likely lacks instructional sensitivity

How can we take advantage of better test-based measures of growth in order to enhance the utility of summative assessments?

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author. 19

Reflections and Suggestions: Growth

- Intuitively attractive measures of growth are problematic
- Planning and implementing assessments to measure growth will require both conceptual and technical improvements
- The next generation of assessment systems should be designed to support growth-related productivity analyses at different levels of sophistication
- Research needed on how growth-focused assessments impact operating characteristics of different VAMs

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author. 20

Reflections and Suggestions: Accountability

- Accountability systems should be based on an explicit theory of action
- They must be designed with compatible components properly linked
- Don't give up on multiple indicators
- Consequential validity is key – so ongoing systemic evaluation is essential
- Assessment quality is a necessary but not sufficient condition for success
- Transparent reporting and ongoing training/support for users is required to achieve desired cost/benefit ratios

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

21



1201 16th St., N.W. | Washington, DC 20036 | Phone: (202) 833-4000

Dennis Van Roekel
President

Lily Eskelsen
Vice President

Rebecca S. Pringle
Secretary-Treasurer

John I. Wilson
Executive Director

December 2, 2009

Sent via racetothetop.assessmentinput@ed.gov

The Honorable Arne Duncan

United States Department of Education

c/o Office of Elementary and Secondary Education

Attention: Race to the Top Assessment Program

400 Maryland Avenue, SW, Room 3E108

Washington, DC 20202

Dear Secretary Duncan:

On behalf of the 3.2 million members of the National Education Association (NEA), we are pleased to submit the following comments and assessment system plan in response to the request for input on a possible Race to the Top Assessment Program for the development of high-quality assessments based on common standards.

Areas of Support

The NEA supports the idea of a program that is organized around consortia of states. State consortia in the CCSSO SCASS program and the New England Common Assessments Program have been successful in developing and improving assessments. NEA believes that funding multiple consortia will allow for a wider range of possibilities in assessment design. Since the purpose of the Race to the Top funding is the development of high-quality assessments, we agree that supporting multiple courses of development is the best way to spur innovation and improvement.

The NEA supports the focus on assessment systems here rather than the usual focus on *accountability systems* because it brings attention and resources to the entire range of purposes of assessment from diagnosis and instructional planning to reporting final achievement. The development of assessment systems that utilize a variety of assessment tools to gather and report multiple kinds of information and

data for addressing the multiple aspects of education is the only path to achieve systemic education reform.

The NEA believes that assessment systems should be developed with all students in mind. The needs of students with disabilities, English language learners, and other special populations should be considered from the outset as assessment systems are being created rather than after the fact through modifications to the base assessment design. Retro-fitting assessment systems for special populations can be expensive and discriminatory.

The NEA supports assessments that are “internationally benchmarked.” We believe our students should have access to curricula comparable to that offered in other countries and know that assessment systems can be designed to leverage that accessibility. However, we believe that the international assessment benchmarking system should include broader indices of the functioning of education delivery systems in which assessments are administered in order to contextualize assessment results. For example, the current United States education system is dramatically different from systems in other countries in which most or all students have access to high quality curricula. The international benchmarks should therefore include data on such areas as resources and efforts toward curriculum development as well as teacher preparation and development, at a minimum. The Department also should encourage benchmarks of child well-being such as child mortality, child poverty, educational possessions, and teenage birth rates (see www.oecd.org/els/social/childwellbeing).

Areas of Concern

The NEA has two primary concerns about the potential direction of this program related to the overall purpose of the Race to the Top Assessment program and the intended uses of assessment data.

Broaden Focus Explicitly

First, we believe the Department’s focus on college and career readiness, while appropriate, is too narrow in scope as well as approach. With respect to scope: other nations with successful education systems prioritize a far broader range of human and academic competencies, including relating to others, lifelong learning, critical and innovative thinking, and communication and collaboration, than the United States’ current focus on academic skills (primarily language arts and mathematics) (New Zealand Curriculum, <http://nzcurriculum.tki.org.nz/Curriculum-documents>; Singapore, http://seab.gov.sg/SEAB/nLevel/syllabus/2010_GCE_N_Syllabuses/7053_2010.pdf). These competencies, including 21st century skills, are regarded by those nations as essential to student learning and transcend any definition of “college and career readiness.”

With respect to approach: to the extent that other nations with successful education systems *do* emphasize “college and career readiness,” they do so using multiple means of assessment, including performance tasks, portfolios, and ongoing formative assessments beyond large-scale, state-directed summative assessments (Darling Hammond & Wood, 2008). Therefore, at least one consortium should

be funded to explore development of an assessment system that addresses a far broader range of essential content, skills, and competencies than mathematics and English language arts as well as a broader range of assessment tools beyond common summative assessments.

Use Assessments for Valid Purposes

Second, the NEA urges the Department to refrain from encouraging the use of large-scale, summative assessments for invalid purposes. While we enthusiastically support the development of effective ways to measure teacher and school leader performance as well as overall school performance, existing state accountability assessments of student achievement are not designed to evaluate any of these important barometers of success. The use of existing state tests of student achievement as they are currently designed and constructed is not valid for evaluating teachers or schools. The American Psychological Association has made it clear in its guidelines on *Appropriate Use of High Stakes Testing in Our Nation's Schools* that the use of a single test for high-stakes decisions is inappropriate since those tests are only snapshots of student performance and they are subject to measurement errors and false conclusions.

The misuse of tests has the potential to negatively impact the entire education system (American Psychological Association, www.apa.org/pubinfo/testing.html). More research is needed on the impact of large-scale, high-stakes assessments before expanding their use to the evaluation of teachers. In addition to improperly evaluating individuals or schools, attempts to use current assessments (or future assessments that are not properly designed) for these purposes may produce unintended consequences, such as driving the most talented professionals away from challenging assignments or schools as well as further narrowing the curriculum and instruction in recognition of the ever larger stakes surrounding the results of student performance on large-scale, summative assessments.

Therefore, in the development of an assessment system, the Department should insist on articulation of specific, detailed, and integrated requirements to foster not only general validity (i.e., do the assessments measure the domains they intend to measure?) but also consequential validity (i.e., are assessments valid and accurate for the purpose for which they are being used?). If the Department wishes to support designing a system for assessing teacher, principal, and school performance, it should insist on the independent reliability and consequential validity of that system. Such systems should certainly emphasize impact on student learning, but they should do so in a way that is reliable and valid.

The NEA believes that serious consideration of the above issues is essential to promote innovation, access for all students, and excellence in our schools, and we are committed to providing whatever support necessary to ensure the success of this funding effort.

Responses to Questions:

General Assessment Questions

QUESTION 1

An essential aspect of any assessment system is that it is guided by a clear articulation of the purpose of education, which then drives the design of the system. While the goal of ensuring that students are “college and career ready” may be a reasonable starting point, it is too narrow to encompass the broad purpose of K-12 education overall. Countries that achieve significant results on international student assessments invariably include broad purpose statements in their national documents on standards or curriculum. Without this broad, directive guidance, the components of assessment systems become overly diffused or focused on small enabling skills and never gauge the achievement of larger, essential education goals.

Purposes

The entire process of assessment system development must begin with a clear, unambiguous statement of the purposes to be served by each assessment component, with specific methods to respond to each of these concretely stated purposes. The time has long passed when we can expect a single assessment, regardless of how broad or technically strong, to serve multiple purposes as broad and disparate as “accountability” and “instructional planning.” If we have learned nothing else from the proliferation of state-mandated assessments over past 20 years – or from eight years of NCLB-driven assessments – we should at least have proven that such instruments are instructionally inert. This is not a reflection on the low quality of the instrumentation, but of the fact that the assessments have been designed specifically to serve accountability purposes. Instructionally useful instruments can be built; however, they cannot evolve from assessments built for accountability purposes.

A complete assessment system would include five major components that must all be addressed simultaneously: summative assessments, formative assessments, teacher capacity development, effective data systems, and evaluation systems that include the analysis of context variables and other measures of effective practice in addition to student assessment results. It would be unwise to focus on summative assessment while leaving formative assessments, teacher capacity development, or effective data or evaluation systems for later attention. Indeed, the “state of the art” with respect to the development of summative assessments is far more advanced than it is for formative measures. It is formative assessments that require a jolt of federal assistance to bring them into the mainstream of assessment systems.

A complete system should incorporate the concept of assessment *of, for, and as* learning. This concept is explained and supported in assessment literature (Bennett & Gitomer, 2009). It is also embraced by several high-achievement countries such as Singapore, New Zealand, and Canada and is integrated into the assessment system description below.

Components

Summative assessment or assessment of learning should include state accountability tests, interim assessments, and end-of-course tests. These assessments should consist mainly of rich, open-ended tasks that require the application of skills and knowledge to solve problems, create projects, and think critically. These assessment “tasks” should be administered throughout the school year and reported as a cumulative score across the tasks. Doing so would eliminate the practice of single, summative assessments that provide only a snapshot of student achievement. It would also allow the use of more complex tasks since the assessment would not be administered in one sitting and therefore would not be subject to the time constraints of a one-time, end-of-year assessment or even a two-time administration solely to mark “growth” in a perfunctory way. Funding for innovation in assessment should also include support for the development and use of performance tasks and/or projects as summative assessment tools.

There should be encouragement of efforts to develop assessment tasks that can be used to measure learning across disciplines such as math-science and social studies-writing. This can increase cost effectiveness and time effectiveness of assessments as well as support the incorporation of real world tasks into assessments.

Additionally, assessments should be scored at least in part by teachers. The most efficient way to incorporate these assessments into the assessment system is to make them available on line.

The assessment system should also have a formative component that is intended for learning. The essential aspects of formative assessment are that it provides direct feedback to students as well as teachers and occurs at a point where additional instruction or learning activities can be identified and used to address learning weaknesses or next steps. Both teachers and students can use formative assessment data to plan learning. These assessments should take place at the classroom level and should be generated either by individual teachers or chosen from a common pool of assessment resources and adapted by teachers to use with specific students. In countries such as Singapore, New Zealand, Great Britain, and the Netherlands, a bank of formative assessment tasks is available to teachers via the Internet. Formative tasks can consist of paper and pencil exercises, performance tasks, demonstrations, and projects. They should be as rich as possible, allowing students to demonstrate and analyze their own learning related to standards and to the content underpinning the large-scale assessments described above. These assessments must take into account the changing needs of individual students and support teachers’ responses to those needs in timely manner. (Black & Williams, 1998) Ultimately, these formative assessment tasks should help the teacher to predict and improve their students’ performance on summative assessments.

The presence of job-embedded, continuous professional development to enable teachers to use assessments and assessment data (assessment as learning) is an essential component of a complete, effective assessment system. There are several forms of building teacher capacity as part of an

assessment system. One is the use of teachers to score summative assessments. This provides them with opportunities to develop deeper understandings of the assessments, learn how students respond to assessment tasks, and determine what might be done in the classroom to improve student learning. This type of professional development currently is rare in the United States but characteristic of assessment systems in other countries. Described as “moderation” in countries such as Singapore, this activity requires that time be added to the school year solely for the purpose of teachers participating in scoring activities. There is evidence that developing and scoring assessments is an effective investment in professional development (Darling-Hammond & Rustique-Forrester, 2005).

The second type of professional development occurs when teachers meet to share and discuss students’ performance on formative assessments. This is assessment *as* learning. It increases teachers’ capacity to make connections among assessments, standards, and curricula. The ultimate implementation of such a system would lead to professional development for teachers with regard to the single most important “assessment” done by teachers—classroom grading.

A third component of professional development addresses the capacity to understand and analyze data from all types of assessments and then to use these multiple forms of data to inform instructional planning. This can be accomplished through effectively presenting information, making resources available to teachers, and facilitating teacher discussion of assessment data.

An effective data system is an essential component of a complete assessment system. An effective data system not only helps stakeholders keep track of student scores on summative assessments but also enables teachers to use technology to choose among optional formative assessment tasks and resources, keep track of data on formative assessments, and share resources and insights with other teachers.

Singapore spent over \$100 million to develop its educational data system, and it appears to have helped that country promote high levels of learning and strong teacher capacity.

For an assessment system to be effective, it should also address contextual variables such as teacher capacity, school climate, community support, and school health and safety. Data on these variables are part of the accountability and assessment systems in Alberta, Canada; Queensland, Australia, and high achieving countries around the world. Incorporating these factors into the assessment system acknowledges the reality that what takes place in classroom cannot be disconnected from the conditions and experiences of students outside of school.

Measures of teacher effectiveness should be based on factors such as those required to become a National Board Certified Teacher. This would allow the use of sound, validated empirical data on teacher expertise rather than trying to rely solely or primarily on linking student performance to specific teachers over a limited time without attention to other variables that are known to affect student achievement beyond the influence of teachers. School climate is another contextual variable that should be included in the data system.

An inspectorate such as those in Britain and New Zealand and suggested by the Broader, Bolder Approach to Education, http://www.boldapproach.org/report_20090625.html, would allow for guided observation of teachers in the classroom and characteristics of school climate, health, and safety. Parental support and involvement also can be assessed through the use of parent surveys.

QUESTION 2

The summative assessments recommended above should be criterion-referenced tests that are aligned with standards. Criterion-referenced assessments make sense when the purpose of the assessment is to determine whether students are meeting or progressing toward meeting standards. Norm-referenced assessments, designed to spread students along a normal curve, are not appropriate for this purpose. The criterion-referenced assessment should, however, allow for students to demonstrate achievement of criteria in multiple ways in order to honor the guidelines of universal design for learning (UDL). These assessments should consist of complex tasks administered three or four times a year and reported as cumulative scores (Bennet, 2009). The tasks can be administered through technological platforms when that is valid and should include small projects that require students to use multiple media to demonstrate achievement. This type of task allows students to demonstrate their learning in several possible ways and thus conforms to the principles of UDL (see Students with Disabilities section below).

The formative assessments should consist of complex tasks as well as quick checks that allow both teachers and students to confirm achievement, note progress, and set next steps in learning. Teachers must be allowed total flexibility to determine which tasks to use, when to use them, and how to adapt them. They should be aligned with, but also go well beyond, standards and summative assessments.

QUESTION 3

LEA teams of educators should be supported in developing formative assessments and provided time to score and discuss the assessments and the implication of the results for instructional planning. Teachers can share assessment tasks and related instructional strategies via the Internet as well as face to face.

QUESTION 4

There are two important requirements that are critical in order to involve teachers in the scoring of summative assessment tasks such as constructed responses and performance tasks. The first requirement is that assessments should be at least partially delivered and taken online. This allows for speedy distribution of assessment tasks and quick delivery of scores once they are completed. Several other countries such as Singapore, New Zealand, and Britain already are doing this.

The second requirement for teacher involvement in scoring is that teachers have expertise in using rubrics and scoring student work. The optimal approach to developing this capacity is to use three types of professional development. First, there should be information and practice sessions for teachers to learn about scoring. Second, there should be opportunities for teachers to use scoring guides aligned with those on the summative assessments for formative assessment tasks. This aspect must be

accompanied by time for teachers to discuss student work in the light of the formative tasks and their scoring of those tasks. The third type of professional development is woven into the actual scoring of students' summative assessment tasks. Teacher scorers need to be given time to discuss key or confusing items and provide input to modification of assessment tasks in subsequent administrations.

QUESTION 5

NEA supports the notion of competency-based student testing rather than grade-level testing. Competency-based testing allows for less frequent assessments, keeps the focus on learning goals rather than grade-level expectations, and acknowledges the fact that students' cognitive development does not proceed in a constant, uniform trajectory. New Zealand's system of bands of grades associated with levels or stages of competency is an excellent model. For example, students taking the first level of competency can pass the assessment anytime during school years 1 and 3. The bands of grades are not exactly the same, which reflects the ranges in levels of achievement across years of schooling.

QUESTION 6

Assessments should be designed to reflect the actual application of students' knowledge and skills. They should be administered at least three times during the school year, and the scores should be combined. To see the true benefit of using growth models, assessments should not consist of only two administrations with a calculation of growth. Using only one assessment as a baseline and only one as a growth indicator is not a reliable way to gather data on growth. The assessments should not be administered every year in every grade but should be administered when students appear ready, based on formative assessment data, to take them within the bands of grades mentioned above. This should occur across bands of years. The New Zealand National Curriculum provides an excellent model of this approach.

Responses to Questions:

High School Assessments

QUESTION 1

Readiness not only for college and career but also for life-long learning should be determined through quality formative assessments linked to comprehensive assessments in all major curricular areas, including civics, world studies, history, foreign language, fine arts, as well as assessments of career technical education fields as appropriate. This would allow a focus on a range of knowledge and skills needed for a productive, satisfying life.

Responses to Questions:
Assessment of English Language Learners

QUESTION 1

It is essential to ensure that the unique factors that impact the performance of English language learners are specifically addressed in the assessments that are used to measure and report the academic achievement of these students. Assessments must be sensitive to the various forms of diversity, including cultural, both within (e.g., Hispanic students of different ethnic and familial backgrounds) and across subgroups (e.g., ELL students with learning disabilities). It cannot be assumed that assessment accommodations adopted for one subgroup will be effective or valid for other subgroups.

QUESTION 2

NEA supports the use of native language assessments when appropriate, but it is important to note that it is difficult to determine the effectiveness of assessments in native languages since there are many different languages represented in this country. Significantly, the validity of using first language for assessments of ELL students depends on the language of instruction and the level of students' fluency in English. Here are several factors to consider:

1. Ensure that the unique factors that impact the performance of English language learners (ELLs), and ELLs with learning disabilities are specifically addressed in the assessments that are used to measure the academic achievement of these students and report the results.
 - a. When developing assessments, consider the specific characteristics of ELLs, in conjunction with standards. Assessments must be sensitive to various forms of diversity, including cultural, both within and across subgroups such as ELLs and ELLs with learning disabilities. It cannot be assumed that assessment or accommodations developed or adapted for one subgroup will be effective and valid for other subgroups. For example, the issues to be addressed in assessments and accommodations for ELLs and ELLs with learning disabilities are not the same.
 - b. Align and integrate standards and assessments that are specifically crafted for ELLs (such as ELLs or ELLs with learning disabilities) into the overall assessment system.
 - c. Incorporate available research, evidence, and principles of fairness and equity for ELLs into assessment systems. (For example, use results from empirical research to indicate when ELLs may be tested in English on content-based assessments based on their level of English language proficiency.)
 - d. Provide the opportunities and resources necessary to ensure that ELLs have meaningful access to the content that is based on state standards.
 - e. Require multiple forms of evidence in the assessment of ELLs, including results of classroom-based assessments and performance of ELLs in the native language and/or in English,

- consistent with the language(s) in which they receive instruction or are best able to indicate their learning.
- f. Understand the diversity within the ELL student population (such as linguistic and cultural differences; and the continuity of educational experiences inside and outside the United States) and act accordingly.
2. Require states to provide research-based recommendations for selecting and using appropriate accommodations for ELLs to ensure that these students have access to valid assessments of their content knowledge.
 - a. While the principles of universal design for learning should be applied to the assessment system for ELLs with learning disabilities, assessment tasks or accommodations should be based on the specific needs of the students being tested.
 - b. Provide specific guidance for selection of assessments and/or accommodations for students with dual classifications (e.g., twice exceptional: ELLs with reading disabilities).
 3. Require states to validate assessment systems for ELLs.
 - a. Include large enough numbers (95%) of ELL students in the validation process.
 - b. Control factors that negatively impact assessment outcomes for ELLs so that variables that are not the primary interest in assessments of achievement do not affect assessment results. For example, a test in English is a test of English for ELLs; therefore, English language proficiency may affect students' ability to demonstrate their academic achievement in English.
 - c. Require that states develop accountability systems that incorporate both growth and status measures. For example, emphasize growth when students are acquiring English language proficiency since language is a developmental process, and then shift the emphasis to a mix of status and growth when students have achieved the necessary proficiency (as determined through validation studies) to learn academic content taught entirely in English.

Responses to Questions:

Assessment of Students with Disabilities

QUESTION 1

Summative and formative assessments must address the needs of students with disabilities, a very heterogeneous group. Even students who are identified under the same disability category can be vastly different from one another in capabilities. Therefore, NEA recommends the following critical considerations for Race to the Top Assessment-funded projects:

- Accessible assessments should be developed for all students using the principles of universal design for learning (UDL) which provide for proactive design that minimizes the need for

accommodations. [For information on UDL, see the UDL guidelines published by the National Center on UDL at <http://www.udlcenter.org/aboutudl/udlguidelines>.]

- When students do require accommodations, they should be provided with the widest range of assessment accommodations feasible. To do this, tests must be designed with a clear specification of the constructs and skills that are being assessed and validated using a variety of accommodation options. This allows teachers to identify which target skills are necessary to successfully participate in the assessment and which accommodations might be needed for each student. [Recently published by the National Accessible Reading Assessment Projects (NARAP), a set of principles for creating accessible reading assessments can help guide the development of future reading assessments. This document is available at <http://www.narap.info/publications/reports/NARAPprinciples.pdf>.]
- Test design should consider how construct elements affect accessibility. For example, word choice, alternate answer choices, graphics, and cognitive demand can dramatically interfere in test item difficulty and yet have very little to do with the skill being assessed.
- Test items should require authentic demonstration of skills and knowledge for all students, including students with disabilities.
- Alternate assessments can be effective vehicles for measuring student skills, knowledge, and growth. However, the administration time for alternate assessments should not decrease the instructional time that these students receive.
- Engage special education professionals in all aspects of standards and assessment development, including scoring.
- Since Individualized Education Program (IEP) teams determine which assessment students will take and what accommodations will be provided, educators need to be provided with professional development on the use of accommodations in instruction and assessment. In particular, classroom teachers and members of IEP teams need to understand the impact of the accommodation recommendations and alternative assessments defined in students' IEPs.
- Please note that international comparisons are problematic for students with disabilities because few countries include students with disabilities to the extent that the United States does. And, disability definitions for the subgroup of students with disabilities can be significantly different from how the U.S. views disabilities.

Responses to Questions:

Technology and Innovation in Assessment

QUESTION 1

Using technology as a base for assessments allows for greater complexity and range in both assessment tasks and student responses. Tasks can include scientific observations and experiments, integrating information from multiple sources and media, and opportunities to make corrections when necessary, based on feedback. Students can respond with recordings of oral answers or speeches, written responses, graphic responses, and explanations of selected responses.

QUESTION 2

The technology platform for assessment systems should include the means for students to take both formative and summative assessments online, allowing teachers and students to access results quickly, and the means for teachers to aggregate and relate data from different types of assessments and assessment tasks.

Computer adaptive testing is only one possibility for the use of technology. While it works well for formative assessment generally, it is not very compatible with the notion of formative, criterion-referenced assessment.

CONCLUSION

The NEA respectfully submits these comments on the Race to the Top Assessment program. If you have any questions, please do not hesitate to contact me at kbrilliant@nea.org or 202-822-7946.

Sincerely,

A handwritten signature in cursive script that reads "B Kay Brilliant".

B. Kay Brilliant, Director

NEA Education Policy and Practice

References

Bennett, R. (2009). The next generation of K-12 reading assessment. Presentation at the ETS Research Forum, Washington, DC, May 3, 2009.

Bennett, R. & Gitomer, D. (2009). Transforming K-12 assessment: Integrating accountability testing, formative assessment, and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century*. New York: Springer.

Black, P. & William, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5, 7-74.

Darling-Hammond, L. & Wood, G. H. (2008) *Refocusing accountability: Using performance assessments to enhance teaching and learning for higher order skills*. The Forum for Education and Democracy.

From: Christopher Camacho, PhD [ccamacho@childrensprogress.com]
Sent: Wednesday, December 02, 2009 12:32 PM
To: Race To The Top Assessment Input
Subject: Race to the Top Assessment Program
Attachments: Race to the Top - Dynamic Assessment.pdf; ATT00001..htm

Name of Submitter:
Christopher Camacho, PhD

Title of the Document Submitted:
Computer Dynamic Assessment for Early Childhood

Topic Addressed:
General Assessment Input

--

Dear Committee Members:

As you consider innovative methods for assessment in our schools, it is vital that early childhood be given special attention and the included document provides information about an innovative method of assessment for early childhood.

The outlined approach is a computer-dynamic approach that is meant to provide objective, valid, and reliable information to help teachers individualize instruction. In particular, a computer-dynamic assessment for early childhood, the Children's Progress Academic Assessment, is mentioned.

Attached is a PDF that contains my written responses. I thank you for your allowing me to provide input and I look forward to the the forthcoming solution that will be provided by the Department.

Sincerely,

Christopher Camacho, PhD

Christopher Camacho, PhD
Director of Research
Children's Progress
Tel: 646.443.9312
Fax: 646.895.7583
www.childrensprogress.com

Inc. 500 Award Winner (2009)
NYER Technology Best Practice Winner (2008)



Children's Progress

: **Christopher J. Camacho, PhD**
: *Director of Research*
: 646.443.9312
: ccamacho@childrensprogress.com

: 108 West 39 Street, Suite 1300
: New York, NY 10018
: 866.427.4787
: www.childrensprogress.com

December 2, 2009

re: Race to the Top - Assessment Program

Dear Committee Members:

As you consider innovative methods for assessment in our schools, it is vital that early childhood be given special attention and the included document provides information about an innovative method of assessment for early childhood.

The outlined approach is a computer-dynamic approach that is meant to provide objective, valid, and reliable information to help teachers individualize instruction. In particular, a computer-dynamic assessment for early childhood, the Children's Progress Academic Assessment, is mentioned.

I thank you for your allowing me to provide input and I look forward to the the forthcoming solution that will be provided by the Department.

Sincerely,

Christopher Camacho, PhD

--

Name of Submitter: Christopher Camacho, PhD

Title of Document Submitted: Computer-Dynamic Assessment for Early Childhood

Topic Addressed: General Assessment Input

Computer-Dynamic Assessment for Early Childhood

The Individuals with Disabilities Education Improvement Act (IDEA, 2004) emphasizes the importance of improving the quality of education by ensuring that schools are held accountable for monitoring the progress of all students. Through IDEA 2004 schools are allowed to use as much as 15% of their special education budgets to fund early intervention activities. Thus, more support is allocated to general education and students identified as “at-risk” are receiving support in general education prior to referral to special education to a greater extent than required in the past. The federal provision also enables schools to move away from an IQ-achievement discrepancy to identify children with learning disabilities and instead use a response to intervention (RTI) model for making these decisions. The reauthorization of IDEA 2004 has led to increased adoption of RTI in schools for the prevention, identification, and early intervention of children with learning disabilities in reading.

Although research supporting the efficacy of RTI for the prevention and remediation of reading difficulties is promising, challenges such as inadequate time and personnel have limited its widespread use. An urgent need exists to examine how technology can support and enhance the implementation of RTI models, particularly at tier 1 when the least amount of support is provided to the classroom teacher. Research on the efficacy of RTI is well-established for improving student performance, reducing referral rates for special education (Telzrow, McNamara, & Hollinger, 2000), and decreasing the disproportionate representation of minorities in special education (Marston, Muyskens, Lau, & Canter, 2003). Further, research has demonstrated that RTI models are effective in detecting academic problems early and preventing problems from worsening.

RTI models were initially formulated to address concerns that some students are misidentified as learning disabled and instead fail to learn because they have not been taught the fundamental skills. Thus, the intention of RTI approaches is to provide students with the least restrictive learning environment by making general education instruction more accessible to diverse learners. In addition, RTI models aim to ensure that students are provided with high-quality instruction and intervention matched to student need. Dynamic assessment is perfectly suited to achieve the aims of RTI because it is able to target each student’s current skill level and provide direct links to instructional decision-making.

RTI and Dynamic Assessment. While the benefits of an RTI approach have been researched in the field, it is unfortunate to note that most teachers do not know how to engage in such assessment practices. Current research shows that high-quality RTI is relatively rare in classrooms and teachers often do not know how to engage in such assessment (Popham, 2008). Researchers have also reported numerous challenges faced by schools that have begun to implement RTI models including a lack of adequate infrastructure (such as adequate space, time, and personnel) to support service delivery (Mathes, Denton, and Vaughn, 2008), differences in intervention implementation (Kethley, Mathes, Nimon, Denton, & Ware, 2008), and too much burden placed on classroom teachers to administer assessment probes and associated interventions during Tiers 1 and 2.

However, the RTI movement has been accompanied by the tremendous innovation and wide-spread use of computers in the classroom to support instructional and assessment activities. The

introduction of classroom technology has allowed the field of educational assessment to grow and develop beyond paper-and-pencil tests and observational records. In particular, computer-delivered dynamic assessments have been shown to have significant benefits for early childhood assessment and instruction (Grigorenko, 2009). Dynamic assessment refers to an assessment procedure that provides corrective feedback in response to student failure to measure both the product and process of learning (Caffrey, Fuchs, & Fuchs, 2008). Dynamic assessment is hypothesized to be the ideal type of assessment to support RTI approaches given that measuring adequate instruction and determining a student's learning potential with assistance is at the core of the RTI approach. Because dynamic assessment is a test-teach-test process, it has the potential to (1) provide direct links to educational intervention and planning decisions; (2) account for the influence of prior learning or educational background; and (3) predict how well students will respond to educational interventions. These are just a few of the reasons that dynamic assessment is believed to be a viable alternative to RTI (Caffrey, Fuchs, & Fuchs, 2008).

Dynamic Assessment Versus Static Assessments. As the field moves from an IQ-Achievement discrepancy model to an RTI model, the type of assessments that are used for helping to determine intervention must also be reconsidered. Researchers have demonstrated that dynamic assessment can be used to identify students who will respond to instruction (Bain & Olswang, 1995), predict future educational placement (Samuels et al., 1992), and can contribute to unique variance in the prediction of future achievement above and beyond traditional tests (Byrne et al., 2000; Meijer, 1993; Resing, 1993; Swanson & Howard, 2005). Further, dynamic assessment may be more advantageous than current RTI approaches, such as fluency assessment because it may be possible to determine responsiveness within one single test administration.

Dynamic assessments are designed to generate information that can be readily used by teachers to tailor instruction to target student's unique learning needs. A critical feature of a dynamic assessment approach is to link assessment with instruction effectively by providing actionable information to support instructional adjustments (cf. Black and William, 1998). In addition, dynamic assessment has been named as a promising alternative to summative assessments for the adoption of an RTI approach because this type of assessment may be able to determine the adequacy of students' response to the intervention (Fuchs and Fuchs, 2008; Fuchs, Fuchs, and McMaster, 2003). Specifically, research has shown that dynamic assessments may predict an individual's potential to learn better than static measures (Grigorenko and Sternberg, 1998; Sternberg and Grigorenko, 2002). In particular, dynamic assessment is able to account for variations in performance that may be due to factors such as prior instruction or a misunderstanding of the task directions, which cannot be accounted for by traditional cognitive or academic testing (Samuels et al., 1992). Therefore, it might be possible to use dynamic assessment to better predict whether the student is likely to respond to instructional interventions or whether a more intensive intervention should be prescribed earlier on in the intervention process.

Principles of Dynamic Assessment. Although there are many different approaches that fall under the umbrella of dynamic assessment, there are a few key aspects that are consistent. Generally, dynamic assessments are identified by the objective to quantify a child's learning potential, whereas summative (or static) assessments typically gauge a child's state of pre-existing knowledge. The primary difference between the two approaches lies in the fact that dynamic assessments adopt a Vygotskian approach to assessment. That is, dynamic assessments typically provide various types of scaffolding after incorrect responses to dissociate what a child can do independently versus what a child can do when provided with scaffolding. Through the scaffolding procedures, dynamic assessments can shed light on particular misunderstandings that may be responsible for a child's

incorrect response. On the other hand, because static assessment does not provide scaffolding after incorrect responses, they are only able to reveal two states of understanding, unaided success and unaided failure (Fuchs, Compton, Fuchs, Hollenbeck, Craddock, and Hamlett, 2008).

Another key aspect of dynamic assessment is that it provides feedback throughout the assessment. In the traditional field of measurement, it is paramount to ensure standardized testing conditions. To this end, teachers and proctors are not allowed to provide feedback to children, as doing so would invalidate the assessment. However, dynamic assessment takes feedback as one of its principles. That is, it is essential to dynamic assessment to mimic the learning environment that children live in - and in such learning environments children receive constant feedback when they are engaged in activities. The question for dynamic assessment is how does the child respond to the feedback and scaffolding procedures - and thus it is this the reaction to feedback and scaffolding that dynamic assessments attempt to quantify. By providing feedback and scaffolding, dynamic assessment makes the evaluation a bi-directional process which mimics the child's daily learning environment.

Computer-Dynamic Versus Computer-Adaptive. It is important to note the differences between computer-dynamic assessment and computer-adaptive assessment. Although both are administered on a computer and often share similar characteristics (e.g., interactive design, immediate reports), they have distinctly different approaches that result in distinctly different objective functions. Adaptive assessments present items in a sequence that is dependent upon the correctness of the examinee's response to the preceding item (typically guided by item response theory). This adaptive process can lead to an accurate measure of a child's level of achievement and are effectively used as summative assessments. However, it is often difficult to use this information to individualize instruction. Dynamic assessments, on the other hand, examine more information about a student's responses (including the accuracy of their responses, the type of errors they make, and how they perform with assistance) and utilize this information to individualize the presentation of items in a way that generates instructionally relevant information about a student's strengths and weaknesses.

Proposed Dynamic Assessment Method. Children's Progress has developed and patented (Patent No. 6,511,326) a dynamic approach to assessment that differs from traditional, norm-referenced testing. Whereas traditional testing refers to the administration and scoring of individual tests based upon the comparison of an individual test with a normative group, our dynamic assessment examines several factors and integrates these results to gain a deeper understanding of the child's learning (cf. Naglieri, Drasgow, Schmit, Handler, Prifitera, Margolis, and Velasquez, 2004). The Children's Progress Academic Assessment (CPAA) utilizes an assessment approach whereby incorrect responses are followed-up with scaffolded questions to gain deeper insight into the child's content understanding. On the other hand, if a child demonstrates mastery of a particular concept, he/she will progress to more advanced content until he/she commits an error - at which time the patented hinting process is initiated. The patented assessment technologies that underlie our assessment examine concepts in language arts that are essential to a child's learning. This process is designed to identify a child's zone of proximal development (where instruction will be most effective) across a range of concepts.

There is as much (if not more) information to be gained by the examining a child's pattern of incorrect responses as there is from looking at his/her correct responses (see Piaget, 2002; Piaget and Inhelder, 2000; Vygotsky, 1962, 1978). However, the vast majority of current testing measures cannot adequately describe a child's pattern of responses and do not provide insight about the child's misunderstandings. The dynamic approach of the CPAA analyzes student responses and

identifies patterns of misunderstanding and areas of difficulty. The patented assessment procedures provide scaffolding after incorrect responses, and examine if the scaffolding was beneficial for children.

Development of the CPAA. Children's Progress is an educational software company located in New York City that aims to improve the meaning of assessment in early childhood education through the use of classroom technologies. Children's Progress was founded in 1999 at Columbia University by renowned Professor Eugene Galanter. In 2003 the company was awarded a patent for the dynamic approach of the academic assessment (U.S. Patent No. 6,511,326 B1). Also that year, Children's Progress released the Children's Progress Academic Assessment, the culmination of years of dedicated research and development to examine practical ways that technology can improve and support early childhood education. Children's Progress maintains this dedication to providing scientifically based assessment in its development of the CPAA.

Every item developed for the CPAA undergoes an extensive development process that can be broken down into four components: content development, production, field testing and analysis, and content alignment. The first process, content development, begins with the creative process of storyboarding and content documentation by education professionals who have graduate degrees in education and/or extensive experience in early childhood education. Careful attention is paid to the alignment of the content with NCTE, research in early literacy development, and other developmental guidelines (see Developmental Indicators of Early Literacy Difficulty/ Content Covered by the CPAA below for more information). The second step involves the production team taking the storyboards to develop the necessary artwork, animations, and voiceovers to create each item. Voiceover quality standards of intonation, pace and fidelity are maintained through custom audio recording by experienced professionals. [For a more detailed description of the nature of the artwork and voiceovers used for the assessment, see Universal Design, below.]

Once the individual items have been produced, they are field tested in two different phases. The first phase involves think aloud testing in which a field researcher presents items individually to a one child at a time. From the initial field-testing, the item may be further modified by the production and/or content teams. The second phase of field-testing involves presenting a single item to several hundred children across the country. From this field test, the item parameters of the item are gathered (e.g., difficulty of the item, guessing parameter of the item, DIF analysis, etc). Based on the data collected from this second round of field testing, additional modifications can be made. The fourth and final stage of development is the content alignment. Children's Progress work with experts in the field to perform a content validity study. Using these results of these studies along with the empirical evidence that is gathered, the dynamic assessment maps for each administration period for each grade is created. (This information is synthesized in the Scope and Sequence document provided in Appendix B.)

Data on Validity and Reliability. Children's Progress has recently completed a three-year validation study on the CPAA (through a grant from the National Institutes of Health, NICHD [SBIR Program]). A final report is currently being prepared; however, a brief summary of the results is presented here. The CPAA demonstrated a reliability (Cronbach's alpha) between 0.89 and 0.92 for children in pre-kindergarten through third grade. In addition, the construct validity of the CPAA was measured against the New York State 3rd Grade Language Arts Test. In this analysis, over 1,400 children in third grade were assessed with the CPAA and with the NY Language Arts Test. The analysis revealed a significant correlation of about 0.7 between the two measures. (Additional information about the validity and reliability of the CPAA can be found on the Children's Progress

website at www.childrensprogress.com.) This data, along with other data collected by the CPAA over the past several years demonstrates that the CPAA is a valid and reliable assessment for early literacy.

Universal Design Principles. Computers have become a ubiquitous feature of our lives. Computers are accessible in all public schools and computer use for instruction and assessment has increased rapidly in recent years (US Department of Commerce, 2002 US Department of Education, 1999). In fact, computer use is more widespread among school-age children than among adults and children are becoming increasingly more comfortable with educational technology (DeBell and Chapman, 2003). The manner in which material is delivered via the computer is designed to be accessible and easily navigated by the youngest children with no prior computer training. A thorough understanding of interactive design and significant research has led to development of an assessment tool that is well suited to meet children’s technical and cognitive abilities. There are a number of elements that are monitored in achieving this end. These elements include, but are not necessarily limited to, the following:

Visibility: What does the child see? In all questions, the background is presented first and then the response choices and/or targets are presented on top of the background. This interactive design ensures that the child’s attention is drawn to the key features of the screen. Additionally, when each response choice or target is presented on the screen, there is an accompanying voiceover prompt that specifically identifies what the image is – this procedure ensures that there is no ambiguity in the child’s mind regarding what the image represents. In all items careful attention has been paid to distinguish foregrounds (e.g., response choices, characters, etc) from background (e.g., settings). This design principle helps to familiarize the child with the content and hones the child’s attention to the significant assets presented on the screen. (See Appendix A for a sample of content that the child sees in the assessment.)

Language: What does the child hear? All instructions are presented by a professional voice actor with several years experience in early childhood interactive design. Encouraging audio feedback is provided after every response and additional prompts are provided to children throughout the assessment to keep them engaged and on task. No reading is required during the assessment apart from items intended to assess reading skills (e.g., the phonemic awareness items contain no aspects related to reading). Verbal instructions are designed to require minimal language processing and can be repeated at the child’s option. In addition, the formatting of questions has been developed to be easily understood.

Accessibility: How does the child respond to the information? All responses entered by the child are “point and click” and only a mouse is needed to answer any question. The vast majority of items are Basic Multiple Choice questions, with the number of options ranging from three to nine. In addition, there are some ordering questions (e.g., alphabetic order, numeric order) that require a sequence of responses. Both of these types of items have been demonstrated to be accessible and understandable by children. Care has been taken to design a screen layout that accommodates inexperienced computer users.

Computer Readiness Screening. Prior to administering the assessment, teachers are given guidelines regarding how to introduce the computer activities and ensure that all the children know the basics regarding how to use the mouse interface. Then, when children begin the assessment, they are presented with a computer readiness assessment. This computer readiness screening begins with a “teacher” that gives general information to the child about the interactive task and some directed instruction on how to use the mouse or touch screen interface. The computer readiness screener

then asks children to follow simple verbal instructions and use the interface to click on different objects (e.g., “click on the balloon to make it pop!”). If the child successfully passes the computer readiness screener, he/she will begin the assessment. However, if the child has some difficulty with the screener, there is additional directed instruction from the teacher avatar. If, after being given additional directed instruction, the child still demonstrates difficulty with the computer interface, the teacher is notified and given specific activities that he/she can work with the child on to build up the child’s computer skills and try the assessment at a later date.

Typical Use Case. The dynamic format of the CPAA makes this assessment efficient at covering a wide range of concepts in a relatively short amount of time. Administration for the CPAA takes less than half a class period to complete language arts components - about 10-12 minutes for children in kindergarten and growing to about 15-18 minutes for children in third grade. The CPAA assessment timeframe is short enough to ensure that children can stay on task and engaged while still covering the key concepts in each grade. The CPAA can be administered in a large group, small group or one-on-one setting for all grade levels. Classroom administration in a computer lab is recommended in order to save time and collect the most comparable assessment data for all students in each classroom. Follow-up administrations and/or progress monitoring is typically conducted in small groups or individually. Regardless, the technology is flexible enough to allow the administration of the assessment to fit a school’s schedule and/or technology organization (e.g., computer lab versus classroom computer administration).

Features and Components of the Automatic Reporting Mechanism. CPAA results are reported in a user-friendly format that was developed to meet specific educator needs. The Children’s Progress web-based reporting mechanism was built around a user-centric design paradigm. The development of the reporting interface was guided by teachers and administrators via extensive guided interviews relating to the uses of assessment information in real-world classroom and school settings. The resulting interface has been extremely well received by existing users in its ability to address specific questions clearly and concisely and allow comprehensive exploration of the meaning of assessment results. Perhaps most importantly, the interface provides information to the teacher as to how the results line up with expectations and suggest the next steps that can bring a student up to individualize instruction.

The reports were designed to complement classroom instruction by providing teachers with immediate, actionable information about each student and the whole classroom without requiring any time for grading. All information is easy to read and interpret without any additional tools. Aside from displaying a quick snapshot of performance, CPAA reports also make student and classroom progress over time easy to follow. All reports are available online immediately for teachers to view upon a student’s completion of the assessment. Results update in real time, so if necessary, teachers can even view partially completed data while a student is being assessed. Furthermore, to ensure maximum flexibility, reports are accessible from any computer with internet access.

Teachers can access “at-a-glance” information about individual student’s performance regarding performance in major areas (e.g., phonemic awareness, phonics, etc.) where the child might be having general difficulty. To examine result more closely, the teacher has access to a highly detailed narrative report that provides highly detailed information about the child’s performance on the assessment. The narrative report provides an automatically generated summary of the child’s performance on each of the sub-concepts evaluated. These summary statements provide specific information about the types of questions the child saw and insight about how the child answered the question. In addition, the narrative report provides specific information about every question the

child saw on the assessment - and whether the child answered the question correct independently, correct with scaffolding, or incorrect with scaffolding. In this way, the teacher can have detailed information about every response entered by the child. Moreover, the narrative report references specific state learning standards. That is, the report identifies what the child was able to do and what the child should be able to do according to state standards. Finally, teachers have access to a progress report. The progress report can track the child's performance in specific content areas across a school year. In addition, the child's performance can be tracked across years - i.e., from kindergarten through third grade. The highly specific progress monitoring give teachers highly detailed information about the child's response to intervention. [Additional information about the individual student reports can be found in Appendix B.]

Tying Assessment Information to Instruction. The zone of proximal development is defined as the difference between what a child can do with scaffolding versus what a child can do independently. Discovering a child's zone of proximal development is important because instruction and intervention have been discovered to be most effective when it is targeted to the child's zone of proximal development (Chaiklin, 2003; Kaprov, 2003; Kozulin, 2003). By individualizing instructing and intervening within the zone of proximal development, a teacher can ensure that the content being presented to the child is at the appropriate level of difficulty and the appropriate level of competency. Thus, there is reduced risk that the material being presented would be too difficult (potentially leading to frustration) or too easy (potentially leading to boredom). The collection of highly detailed information about the child's incorrect responses allows us to identify the child's specific and unique needs that can be used to individualize intervention protocols for the child.

The dynamic approach of the CPAA attempts to identify the child's zone of proximal development across a range of concepts in early literacy. Further, it is important to note that this dynamic approach is applied to every item contained within the CPAA. Using this information, the CPAA automatically provides targeted recommended activities for children based on their performance on the assessment. Each activity is selected based on the child's performance on the assessment and the type of activity that is selected is determined by the level of understanding the child has on that particular subconcept. For example, a child can complete the rhyming component and either answers the questions correctly without scaffolding (i.e., independent understanding), correctly with scaffolding (i.e., scaffolded understanding), or incorrectly with scaffolding (i.e., no understanding demonstrated). To this end, the selection of the recommended activity will reference the child's pattern of responses. The individual student report contains 1-2 pages of activities and instructional strategies for each child (see Appendix B for a sample of recommended activities that are automatically generated by the CPAA).

References

- Anthony, J., & Lonigan, C. (2004). The nature of phonological awareness: converging evidence from four studies of preschool and early grade school children. *Journal of Educational Psychology*, 96, 43-55.
- Berninger, V. W., Abbott, R. D., Brooksher, R., Lemos, Z., Ogier, S., Zook, D., & Mostafapour, E. (2000). A connectionist approach to making the predictability of English orthography explicit to at-risk beginning readers: Evidence for alternative, effective strategies. *Developmental Neuropsychology*, 17(2), 241–271.
- Black, P. & William, D. (October, 1998). Inside the Black Box: Raising Standards Through Classroom Assessment, (Phi Delta Kappan, October 1998) Retrieved from: <http://www.pdkintl.org/kappan/kbla9810.htm>
- Burns, M. K & Ysseldyke, J. E. (2006). Comparison of Existing Response-to-Intervention Models to Identify and Answer Implementation Questions. *NASP Communiqué*, 34, 5.
- Byrne, B. and Fielding-Barnsley, R. (2000). Phonemic awareness and letter knowledge in the child's acquisition of the alphabetic principle. *Journal of Educational Psychology*, 81, 313-321.
- Caffrey, E., Fuchs, D., & Fuchs, L. (2008, February). The predictive validity of dynamic assessment: A review. *The Journal of Special Education*, 41(4), 254-270.
- Chaiklin, S. (2003). The zone of proximal development in Vygotsky's analysis of learning and instruction. In A. Kozulin, B. Gindis, V. S. Ageyev, and S. M. Miller (Eds.), *Vygotsky's Educational Theory in Cultural Context*, pp. 39-64. New York: Cambridge University Press.
- Démonet, J., Taylor, M., & Chaix, Y. (2004). Developmental dyslexia. *Lancet*, 363(9419).
- Eldredge, J. L. (2005). Foundations of fluency: An exploration. *Reading Psychology*, 26, 161–181.
- Fuchs, L. S., & Fuchs, D. (2007). The role of assessment in the three tier approach to reading instruction. In D. Haager, J. Klingner, & S. Vaughn (Eds.), *Evidence-based reading practices for response to intervention* (pp. 29–44). Baltimore: Brookes.
- Fuchs, D., & Fuchs, L. (2006). Introduction to Response to Intervention: What, why, and how valid is it?. *Reading Research Quarterly*, 41(1), 93-99.
- Galanter, E. and Galanter, M. (2003). Adaptive evaluation method and adaptive evaluation apparatus. United States Patent, no. 6,511,326 B1.
- Good, R. H., Simmons, D. C., & Smith, S. B. (1998). Effective academic interventions in the United States: Evaluating and enhancing the acquisition of early reading skills. *Educational and Child Psychology*, 15, 56–70.
- Hammill, D. (2004). What we know about correlates of reading. *Exceptional Children*, 70(4), 453-468.

Jenkins, J. R. & Johnson, E. (n.d.) Universal Screening for Reading Problems: Why and How Should We Do This? Retrieved from: <http://www.rtinetwork.org/Essential/Assessment/Universal/ar/ReadingProblems>

Individuals With Disabilities Education Improvement Act (IDEA), Pub. L. No. 94–142.

Kaprov, Y.V (2003). Vygotsky's doctrine of scientific concepts: Its role for contemporary education. In A. Kozulin, B. Gindis, V. S. Ageyev, and S. M. Miller (Eds.), *Vygotsky's Educational Theory in Cultural Context*, pp. 39-64. New York: Cambridge University Press.

Kozulin, A. (2003). Psychological tools and mediated learning. In A. Kozulin, B. Gindis, V. S. Ageyev, and S. M. Miller (Eds.), *Vygotsky's Educational Theory in Cultural Context*, pp. 15-38. New York: Cambridge University Press.

McCandliss, B., & Noble, K. (2003). The development of reading impairment: A cognitive neuroscience model. *Mental Retardation and Developmental Disabilities Research Reviews*, 9, 196-204.

Naglieri, J. A, Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., and Velasquez, R. (2004). Psychological testing on the Internet. *American Psychologist*, 59, 150-162.

National Reading Panel. (2000). Teaching children to read: An evidence based assessment of the scientific research literature on reading and its implications for reading instruction. Washington, DC: National Institute of Child Health and Human Development.

Paris, S. G. (2005). Reinterpreting the development of reading skills. *Reading Research Quarterly*, 40, 184 – 202.

Piaget, J. (2002). The language and thought of the child: Jean Piaget: Selected works. London: Routledge.

Shinn, M. R. (2007). Identifying students at risk, monitoring performance, and determining eligibility within response to intervention: Research on educational need and benefit from academic intervention. *School Psychology Review*, 36(4) 601-617.

Simons, P.G., Fletcher, J. M., Bergman, E., (2002). Dyslexia-specific brain activation profile becomes normal following successful remedial training. *Neurology*, 58, 1203-1213.

Snow, C. F., Burns, M. S., & Griffin, P. (Eds.). (1998). Preventing reading difficulties in young children. Washington, DC: National Academies Press.

Swanson, H., & Howard, C. (2005, December). Children with Reading Disabilities: Does Dynamic Assessment Help in the Classification?. *Learning Disability Quarterly*, 28(1), 17-34. Retrieved May 23, 2009, doi:10.2307/4126971

Torgesen, J. K. (2002). The prevention of reading difficulties. *Journal of School Psychology*, 40, 7–26.

Torgesen, J., & Mathes, P. (1999). What every teacher should know about phonological awareness. *Reading research: Anthology: The why of reading instruction* (pp. 54-61). Novato, CA, US: Arena Press.

Vygotsky, L. S. (1962). *Thought and language*. Cambridge, MA: MIT Press.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

VanDerHeyden, A. M., Snyder, P. A., Broussard, C., Ramsdell, K. (2007). Measuring response to early literacy intervention with preschoolers at risk. *Topics in Early Childhood Special Education*, 27, 232-249.

US Department of Commerce (2002). *A nation online: How Americans are expanding their use of the internet*. Washington, DC: US Department of Commerce.

US Department of Education (1999). *Office of Educational Technology, Progress Report on Educational Technology*. Washington, DC: US Department of Education.

Appendix A: Sample Assessment Item

The Children’s Progress has developed an assessment approach whereby incorrect responses are followed-up with scaffolded questions. The type of scaffolding presented to the child depends upon the child’s incorrect response. The example presented in Figure 1 is a screenshot from a sample rhyming question.

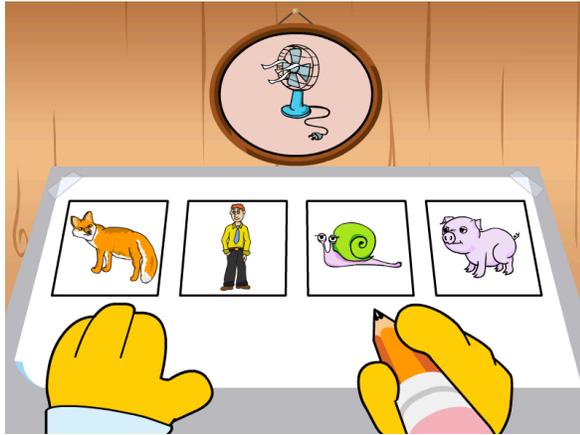


Figure 1. A rhyming question from the CPAA. In these questions, the child is asked to identify a word that rhymes with a target word. If the child answers the Independent Question incorrectly, then the child is presented with the Scaffolded Question.

Independent Question Audio Script: “Click on the picture that rhymes with the word ‘fan.’”

Scaffolded Question Audio Script: [presented when the child incorrectly clicks on “fox”]. “Fox. Fan. They sound the same at the beginning, but not at the end. Fan rhymes with can and pan. Click on the picture that rhymes with the word ‘fan.’”

Questions like this one were presented to children in kindergarten in the fall. All these questions began with the Independent presentation of the question and followed up by the Scaffolded presentation of the question only if the child answered the question incorrectly. The data from these rhyming questions is presented below.

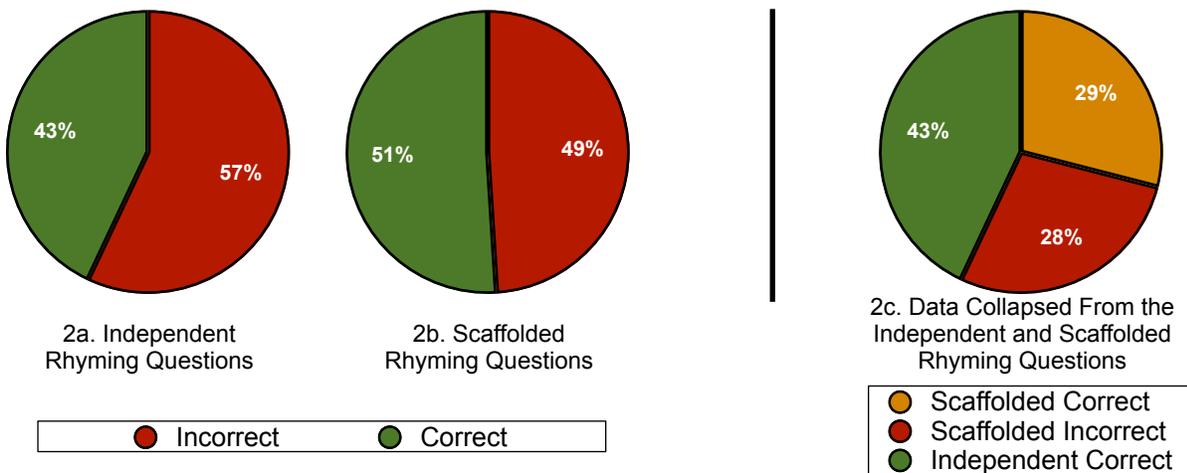


Figure 2a-c. Data collected from Independent and Scaffolded questions on rhyming from children in kindergarten. Figure 2a (left) presents the data collected from all the Independent rhyming questions. Figure 2b (center) presents the data collected from all the Scaffolded rhyming questions (children only see the rhyming questions after an incorrect response to the Independent question). Figure 2c presents the data collapsed from the Independent and Scaffolded questions. By presenting data with three different outcomes (as in Figure 2c), we can gain a deeper insight into the children’s understanding of the content. Certainly, children who answer correctly independently are different from children who answer scaffolded question correctly from children who answer scaffolded question incorrectly.

Appendix B: Sample Children's Progress Online Teacher Reports

The CPAA generates user-friendly reports for teachers, administrators and instructional specialists. All reports are available instantly (as soon as students complete the assessment). Below are a few examples of the reports available to teachers.

Fig 1. Class Summary Report

An overview of a classroom's latest assessment, with colorful charts representing performance levels by concept

- Teacher Tools
- Edit My Profile
- Download Software
- View Help
- Manage Roster [Add]
- This Report
- Print All Full Reports
- My Class
- Grades / Classes
- Alexander Johnson
- Students
- A - L
- [Abati, Trinity](#)
- [Axon, Yoshiko](#)
- [Bennick, Rosario](#)
- [Bernacchi, Oliver](#)
- [Brown, Samantha](#)
- [Copeland, Velma](#)
- [Dahlberg, Buffy](#)
- [Debraga, Lizeth](#)
- [Enix, Jed](#)
- [Greenleaf, Fred](#)
- [Locsin, Ulysses](#)
- M - Z
- [Niwa, Genia](#)
- [Schellhase, Leda](#)
- [Schrantz, Damian](#)
- [Storto, Frederic](#)
- [Streicek, Shalanda](#)
- [Trumbull, Gavin](#)
- [Wesner, Sherell](#)
- [Zike, Hilma](#)

Fall Gr2 | Winter Gr1 '07-'08 | Show All
Recent Assessments

Alexander Johnson's Class Report

Winter Gr2

Print version

Proctor:	Alexander Johnson	Legend: Above expectation (3.5 - 4.0) At expectation (2.5 - 3.5) Approaching expectation (1.5 - 2.5) Below expectation (1 - 1.5)
Date:	01/18/09	
Assessment:	CPAA Grade 2 Winter	

Report Areas

Report Card
Class Roll
Activities
Progress

Click on concept to see details.

Language Arts - Class's Concept Scores Score scales 1 to 4

Concept	Graph	Level	Class Avg.	School Avg.
Phonemic Awareness	<div style="width: 100%; height: 15px; background-color: #76b82a;"></div>	At Expectation	2.6	2.2
Reading	<div style="width: 100%; height: 15px; background-color: #76b82a;"></div>	At Expectation	2.5	2.7
Writing	<div style="width: 100%; height: 15px; background-color: #76b82a;"></div>	At Expectation	2.9	2.5

Mathematics - Class's Concept Scores Score scales 1 to 4

Concept	Graph	Level	Class Avg.	School Avg.
Measurement	<div style="width: 100%; height: 15px; background-color: #f1c40f;"></div>	Approaching Expectation	2.3	2.4
Numeracy	<div style="width: 100%; height: 15px; background-color: #f1c40f;"></div>	Approaching Expectation	2.4	2.4
Operations	<div style="width: 100%; height: 15px; background-color: #76b82a;"></div>	At Expectation	2.7	2.6
Patterns and Functions	<div style="width: 100%; height: 15px; background-color: #f1c40f;"></div>	Approaching Expectation	2.1	2.2

Fig 2. Class Roster

An interactive roster, sortable and printable by performance in any concept.

Teacher Tools

- [Edit My Profile](#)
- [Download Software](#)
- [View Help](#)
- [Manage Roster \[Add\]](#)

This Report

- [Print All Full Reports](#)

My Class

- [Grades / Classes](#)
- Alexander Johnson

Students

A - L

- [Abati, Trinity](#)
- [Axon, Yoshiko](#)
- [Bennick, Rosario](#)
- [Bernacchi, Oliver](#)
- [Brown, Samantha](#)
- [Copeland, Velma](#)
- [Dahlberg, Buffy](#)
- [Debraga, Lizeth](#)
- [Enix, Jed](#)
- [Greenleaf, Fred](#)
- [Locsin, Ulysses](#)

M - Z

- [Niwa, Genia](#)
- [Schellhase, Leda](#)
- [Schrantz, Damian](#)
- [Storto, Frederic](#)
- [Strejcek, Shalanda](#)
- [Trumbull, Gavin](#)
- [Wesner, Sherell](#)
- [Zike, Hilma](#)

Fall Gr2 | Winter Gr1 '07-'08 | Show All

Recent Assessments

Alexander Johnson's Class Report

Winter Gr2

 Print version

Proctor: Alexander Johnson

Date: 01/18/09

Assessment: CPAA Grade 2 Winter

Legend:

- Above expectation (3.5 - 4.0)
- At expectation (2.5 - 3.5)
- Approaching expectation (1.5 - 2.5)
- Below expectation (1 - 1.5)

Report Areas

[Report Card](#) | [Class Roll](#) | [Activities](#) | [Progress](#)

View: Language Arts Mathematics

Click on the concept headers to sort by that concept.
Click on the student name to see that student's individual report.

Language Arts - Concept Scores Per Student Score scales 1 to 4

Students	Phonemic Awareness	Reading	Writing
Strejcek, Shalanda	1	2	3
Zike, Hilma	1	2	3
Abati, Trinity	2	3	3
Copeland, Velma	2	1	3
Enix, Jed	2	3	3
Niwa, Genia	2	4	3
Schrantz, Damian	2	1	4
Trumbull, Gavin	2	4	2
Wesner, Sherell	2	4	4
Axon, Yoshiko	3	2	2
Bennick, Rosario	3	4	3
Bernacchi, Oliver	3	2	2
Brown, Samantha	3	1	2
Greenleaf, Fred	3	3	4
Schellhase, Leda	3	2	2
Storto, Frederic	3	4	3
Dahlberg, Buffy	4	2	4
Debraga, Lizeth	4	2	3
Locsin, Ulysses	4	2	2

Fig 3. Class Activity List

A list of recommended activities for a classroom (generated based on assessment performance). Each activity can be opened and printed, complete with a list of suggested participants. Activity lists can also be viewed for individual students.

Teacher Tools

- [Edit My Profile](#)
- [Download Software](#)
- [View Help](#)
- [Manage Roster \[Add\]](#)

This Report

- [Print All Full Reports](#)

My Class

Grades / Classes

- Alexander Johnson

Students

A - L

- [Abati, Trinity](#)
- [Axon, Yoshiko](#)
- [Bennick, Rosario](#)
- [Bernacchi, Oliver](#)
- [Brown, Samantha](#)
- [Copeland, Velma](#)
- [Dahlberg, Buffy](#)
- [Debraga, Lizeth](#)
- [Enix, Jed](#)
- [Greenleaf, Fred](#)
- [Locsin, Ulysses](#)

M - Z

- [Niwa, Genia](#)
- [Schellhase, Leda](#)
- [Schrantz, Damian](#)
- [Storto, Frederic](#)
- [Streicek, Shalanda](#)
- [Trumbull, Gavin](#)
- [Wesner, Sherell](#)
- [Zike, Hilma](#)

Fall Gr2 Winter Gr1 '07-'08 Show All
Recent Assessments

Alexander Johnson's Class Report

Winter Gr2

Print version

Proctor:	Alexander Johnson	Legend: Above expectation (3.5 - 4.0) At expectation (2.5 - 3.5) Approaching expectation (1.5 - 2.5) Below expectation (1 - 1.5)
Date:	01/18/09	
Assessment:	CPAA Grade 2 Winter	

Report Areas

[Report Card](#) [Class Roll](#) [Activities](#) [Progress](#)

Click on each of the activities to see recommended participants.

Language Arts - Recommended Activities

Phonemic Awareness	Reading	Writing
Building Words	Active Reading	Dear Pen Pal
Dissecting and Creating Words	Beginning, Middle and End	Editing Scavenger Hunt
Singing Words	Boring Word Pit	Fill-in-the-Blank
Syllable Steps	Can You Arque With That?	Capitalization
Vowel Changes	Complete the Sentence	Fix the Mistakes
Word Jumble	Dice Roll	Grammar Reinforcer
Word Shuffle	Does It Belong?	Guess My Word
	How Are We Alike?	Invent-a-Poem
	I Wonder?	Missing Vowels
	Let's Make a Poem/Song	Mistakes Galore
	Listen To This	Period Hunt
	One Sentence	Sentence Stumpers
	Summaries	Spelling Bee
	Personal Dictionaries	Who's At Bat?

Fig 4. Student Detailed Report

A detailed, state standards-referencing narrative, outlining an individual student’s assessment experience and highlighting specific strengths and weaknesses. Recommended activities are included based on concept-specific performance

- Teacher Tools
- [Edit My Profile](#)
- [Download Software](#)
- [View Help](#)
- [Manage Roster \[Add\]](#)
- This Report
- [Print All Full Reports](#)
- My Class
- Grades / Classes
- * [Alexander Johnson](#)
- Students
- A - L
- [Abati, Trinity](#)
- [Axon, Yoshiko](#)
- [Bennick, Rosario](#)
- [Bernacchi, Oliver](#)
- [Brown, Samantha](#)
- [Copeland, Velma](#)
- [Dahlberg, Buffy](#)
- [Debraqa, Lizeth](#)
- [Enix, Jed](#)
- * [Greenleaf, Fred](#)
- [Locsin, Ulisses](#)
- M - Z
- [Niwa, Genia](#)
- [Schellhase, Leda](#)
- [Schrantz, Damian](#)
- [Storto, Frederic](#)
- [Streicek, Shalanda](#)
- [Trumbull, Gavin](#)
- [Wesner, Sherell](#)
- [Zike, Hilma](#)

Winter Gr2
Fall Gr2
Winter Gr1 '07-'08
Show All

Recent Assessments

Fred Greenleaf's Report

Winter Gr2

[Print version](#)
[Go to class report](#)

Proctor:	Alexander Johnson	Legend: Above expectation (3.5 - 4.0) At expectation (2.5 - 3.5) Approaching expectation (1.5 - 2.5) Below expectation (1 - 1.5)	
Assessment:	CPAA Grade 2 Winter		
Date/Time:	01/18/09 7:11pm		

Report Areas

Report Card
Full Report
Activities
Progress

View: Language Arts Mathematics

Language Arts

Phonemic Awareness
Reading
Writing

Phonemic Awareness At Expectation

Open all sub concepts (details view) |
 Close all sub concepts

[1]Fred added a phoneme to an existing word to create a new word containing a blend. [3]In the following section, Fred decoded a nonsense word containing a complex rime and a digraph without assistance.

- ✔ Correct answer
- ✔ Correct answer with hint
- ✘ Incorrect answer

Phonemic Addition

Fred Greenleaf was able to:	Fred Greenleaf should be able to:	Recommended Activities:
Fred added a phoneme to an existing word to create a new word containing a blend. (PA.10.1.a PA.10.1.b PA.10.3.a)	Fred should blend sounds using knowledge of letter-sound correspondences in order to decode unfamiliar, but decodable, multisyllabic grade-level words (NY Learning Standard Reading 1-4).	Building Words
✔ If you add the sound /b/ to the beginning of "ring", what new word do you get?		
✔ If you add the sound /s/ to the beginning of "pot", what new word do you get?		

Fig 5. Student Progress

An individual student's progress in literacy and mathematics, sortable by concept and time period

Teacher Tools

- [Edit My Profile](#)
- [Download Software](#)
- [View Help](#)
- [Manage Roster \[Add\]](#)

This Report

- [Print All Full Reports](#)

My Class

Grades / Classes

- * Alexander Johnson

Students

A - L

- [Abati, Trinity](#)
- [Axon, Yoshiko](#)
- [Bennick, Rosario](#)
- [Bernacchi, Oliver](#)
- [Brown, Samantha](#)
- [Copeland, Velma](#)
- [Dahlberg, Buffy](#)
- [Debraqa, Lizeth](#)
- [Enix, Jed](#)
- * [Greenleaf, Fred](#)
- [Locsin, Ulysses](#)

M - Z

- [Niwa, Genia](#)
- [Schellhase, Leda](#)
- [Schrantz, Damian](#)
- [Storto, Frederic](#)
- [Streicek, Shalanda](#)
- [Trumbull, Gavin](#)
- [Wesner, Sherell](#)
- [Zike, Hilma](#)

Winter Gr2 | Fall Gr2 | Winter Gr1 '07-'08 | Show All

Recent Assessments

Fred Greenleaf's Report Winter Gr2

[Print version](#)
[Go to class report](#)

Proctor: Alexander Johnson

Assessment: CCAA Grade 2 Winter

Date/Time: 01/18/09 7:11pm

Legend:

- Above expectation (3.5 - 4.0)
- At expectation (2.5 - 3.5)
- Approaching expectation (1.5 - 2.5)
- Below expectation (1 - 1.5)



Report Areas

[Report Card](#) | [Full Report](#) | [Activities](#) | **[Progress](#)**

View: Selected Year Over The Years

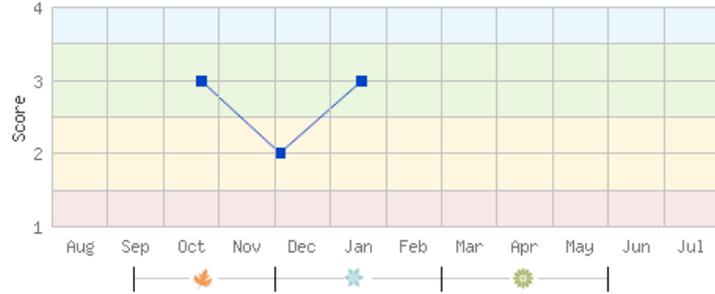
Subject scores for the selected year



Month	Language Arts	Mathematics
Oct	2.5	2.8
Nov	2.5	1.8
Jan	3.5	2.5

Concept scores for the selected year

Subject: Language Arts Mathematics



Select Concept:

- Phonemic Awareness
- Reading
- Writing

Appendix C: Sample Children’s Progress Print Reports

Any Children’s Progress report can be printed. Below are some examples of commonly printed reports

Fig 1. Student Detailed (Narrative) Report

A list of the concepts the student was tested in, the corresponding state learning standard, and how the student responded (correctly, correctly after seeing a hint, or incorrectly)

← Return to normal view | Print this page

Amelia Bedelia's Report

CPAA-K-Fall

Proctor: Teacher Preview
 Assessment: CPAA Kindergarten Fall
 Date and Time: 10/21/08 5:28am

View: Summary Full Details

**Student Narrative Report:
Full Details View**

This report provides teachers with an additional level of detail. They can identify exactly which questions each student saw, the corresponding **CALIFORNIA CONTENT STANDARD**, and how the student responded (correctly, correctly after seeing a hint, or incorrectly).

Language Arts

Writing

Amelia was asked to find some letters. She identified 2 of 3 letters on the first try. She moved on to the letter-sound section. On the first try, Amelia was not able to match either of two presented letters to its respective sound.

Approaching Expectation

- ✔ Correct answer
- ✔ Correct answer with hint
- ✘ Incorrect answer

▼ Letter ID		
Amelia Bedelia was able to:	Amelia Bedelia should be able to:	Recommended Activities
Amelia identified 2 of 3 letters on the first try.	Amelia should recognize and name all uppercase and lowercase letters of the alphabet (CA ELA Content Standard for K – Reading 1.6)	Letter Hunt Matching Memory
<ul style="list-style-type: none"> ✘ Click on the letter "g". ✔ Click on the letter "S". ✔ Click on the letter "I". 		
▼ Letter-Sound: Single Letter		
Amelia Bedelia was able to:	Amelia Bedelia should be able to:	Recommended Activities
On the first try, Amelia was not able to match either of two presented letters to its respective sound.	Amelia should match all consonant and short-vowel sounds to appropriate letters (CA ELA Content Standard for K – Reading 1.14)	Alphabet Taboo Toss a Letter
<ul style="list-style-type: none"> ✔ What letter makes the sound /d/ as in dog? ✔ What letter makes the sound /t/ as in top? 		

Phonemic Awareness

Amelia matched two words with the same initial letter sound without assistance. She had difficulty with the rhyming section and was not able to rhyme a one-syllable word, even with guidance.

Approaching Expectation

- ✔ Correct answer
- ✔ Correct answer with hint
- ✘ Incorrect answer

Fig 2. Student Recommended Activities

A sampling of the activities recommended for a particular student based on his or her assessment performance, organized by subject and concept.

[Return to normal view](#) | [Print this page](#)

 **Amelia Bedella's Report**
CPAA-K-Fall

Proctor: Teacher Preview
Assessment: CPAA Kindergarten Fall
Date and Time: 10/21/08 8:28am

Mathematics > Measurement

Long and Longer
Length Comparison Instructional Activity: Have the class sit in a circle that includes you. Name an object that is not very long, e.g. a paper clip. Then go around the circle and have each person name an object (or distance) once you have exhausted long objects) that is longer than the previous one. In the second round, players have to name an object that is shorter than the last. Make sure you will make it around the circle by drawing attention to and preventing excessively large jumps in size. For example, a school bus is longer than a paper clip, but that may end the game. Challenge students to think of something just a little bit longer or shorter than the previous object.

Making Shapes
Shape ID Instructional Activity: Cut shapes out of felt. Divide the class up into groups. Give each group one of the felt shapes and challenge each group to lie on the floor and use their bodies to make that shape. When the group of students is done with their shape, take a picture of them. Then make a book of shapes with all of the pictures.

More Letters
Quantity Comparison Instructional Activity: Divide the students into pairs. Ask them to write their names on a piece of paper. If they struggle with writing, encourage them to copy their name from an already printed place. Next, have the students count the letters in their name and compare the quantities. Whoever has more proceeds to the child with the longest name in a nearby pair. Continue with the comparisons until the child(ren) with the longest name in the class has been determined.

Shoobox and A Ball
Positions - Reference Instructional Activity: Give one of your students a shoebox and a ball. Tell them to arrange the objects based on what positional term you use (i.e. put the ball inside the shoebox). Make this activity more challenging by asking students to remember the order of positions used. Similarly, you can have two students alternate in thinking of arrangements.

Mathematics > Numeracy

Matching Cards
Number ID Instructional Activity: Create a set of cards. Show the digit on one card and the matching number of dots on the second card. Make a pair of cards for the digits 1 - 10. Children then use the deck to match the digit with the corresponding pictures.

Reorganizing Objects
Subitizing Instructional Activity: Give children a set number of small circular objects such as marbles or checkers pieces. Ask the students to organize the objects into different shapes such as into a triangle, a house, or a rectangle. Ask the students which shape makes it easiest to identify the quantity. Then show them flash cards of different amount of circles and see who can guess the quantity of each one.

Sequential Surprise
Correct Order Instructional Activity: Write the numbers 1-25 on individual pieces of paper. Have the students write four blank lines on a sheet _____. Then pick a number and call it out. The students have to decide where to place the number. For example, if the number 25 was called, the student should place it last; the number four first. Once the number is put down, it cannot be moved. When you are finished calling out all four numbers, see who has the numbers in the correct order from smallest to largest. Take this opportunity to introduce probability concepts if you feel it is appropriate, for instance, by asking if most of the numbers in the 1-25 are above or below 5. How about 10, 20?

Switch Seats
Ordinality Supportive Activity: Ask the group to sit in one long line, and allow one student volunteer to stand outside of the line. Explain that you will be playing a game where the children will have to switch their seats. You will be calling out different ordinal numbers for the children to switch with, one at a time. Tell your volunteer to tap the third person in line. The third child then stands up and gives his seat to the original volunteer. Then you call out another ordinal number, "Tap the tenth person in line" - and the tapping and switching continues. Switch your wording around to reinforce different ways of ordering numbers. Instead of saying twenty seventh, say last. Instead of saying twenty sixth, say second to last. For older children, you could ask them to tap the child exactly in the middle.

Student Narrative Report: Recommended Activities

This report includes a sampling of the activities recommended for a particular student based on his or her assessment performance. Activities are organized by subject and concept.

From: Christopher Camacho, PhD [ccamacho@childrensprogress.com]
Sent: Wednesday, December 02, 2009 2:10 PM
To: Race To The Top Assessment Input
Subject: Re: Race to the Top Assessment Program
Attachments: Race to the Top - Dynamic Assessment Addendum.pdf; ATT00001..htm

December 2, 2009

re: Additional Information for "Computer-Dynamic Assessment for Early Childhood" Statement

Dear Committee Members:

This information is provided as an addendum to our statement submitted on December 2, 2009 entitled "Computer-Dynamic Assessment for Early Childhood."

To add additional clarification, the statement was written to address Questions 1 and 2 related to General Assessment. In particular, my hope in writing this statement was to bring the Committee's attention to an innovative approach to assessment. Computer-dynamic assessment holds great promise for the state of educational assessment. By implementing assessments that have a pedagogical emphasis while also delivering results that states can use for accountability and to inform school improvement initiatives, teachers and states can spend their less time interpreting results and more time helping students.

Further, I would like to note that the Children's Progress Academic Assessment (CPAA) is currently being used nationwide. The CPAA has tripled in usage each year over the past four years. Currently, the CPAA is used in over 40 states across the country by over 500,000 children in over 1,200 schools. Of notable adoption has been the State of Mississippi which is currently using the CPAA to assess every child in kindergarten through third grade. The CPAA is one of the very few computer-dynamic assessments in early childhood literacy assessment that has widespread national use.

Finally, I would like to mention one last final point about the CPAA. Children's Progress has recently completed a three-year validation study on the CPAA (through a grant from the National Institutes of Health, NICHD [SBIR Program]). A final report is currently being prepared; however, a brief summary of the results is presented here. The CPAA demonstrated a reliability (Cronbach's alpha) between 0.89 and 0.92 for children in pre-kindergarten through third grade. In addition, the construct validity of the CPAA was measured against the New York State 3rd Grade Language Arts Test. In this analysis, over 1,400 children in third grade were assessed with the CPAA and with the NY Language Arts Test. The analysis revealed a significant correlation of about 0.7 between the two measures. (Additional information about the validity and reliability of the CPAA can be found on the Children's Progress website at www.childrensprogress.com.) This data, along with other data collected by the CPAA over the past several years demonstrates that the CPAA is a valid and reliable assessment.

Thank you for your time.

Sincerely,

Christopher Camacho, PhD



Children's Progress

: **Christopher J. Camacho, PhD**
: *Director of Research*
: 646.443.9312
: ccamacho@childrensprogress.com

: 108 West 39 Street, Suite 1300
: New York, NY 10018
: 866.427.4787
: www.childrensprogress.com

December 2, 2009

re: Additional Information for "Computer-Dynamic Assessment for Early Childhood" Statement

Dear Committee Members:

This information is provided as an addendum to our statement submitted on December 2, 2009 entitled "Computer-Dynamic Assessment for Early Childhood."

To add additional clarification, the statement was written to address Questions 1 and 2 related to General Assessment. In particular, my hope in writing this statement was to bring the Committee's attention to an innovative approach to assessment. Computer-dynamic assessment holds great promise for the state of educational assessment. By implementing assessments that have a pedagogical emphasis while also delivering results that states can use for accountability and to inform school improvement initiatives, teachers and states can spend their less time interpreting results and more time helping students.

Further, I would like to note that the Children's Progress Academic Assessment (CPAA) is currently being used nationwide. The CPAA has tripled in usage each year over the past four years. Currently, the CPAA is used in over 40 states across the country by over 500,000 children in over 1,200 schools. Of notable adoption has been the State of Mississippi which is currently using the CPAA to assess every child in kindergarten through third grade. The CPAA is one of the very few computer-dynamic assessments in early childhood literacy assessment that has widespread national use.

Finally, I would like to mention one last final point about the CPAA. Children's Progress has recently completed a three-year validation study on the CPAA (through a grant from the National Institutes of Health, NICHD [SBIR Program]). A final report is currently being prepared; however, a brief summary of the results is presented here. The CPAA demonstrated a reliability (Cronbach's alpha) between 0.89 and 0.92 for children in pre-kindergarten through third grade. In addition, the construct validity of the CPAA was measured against the New York State 3rd Grade Language Arts Test. In this analysis, over 1,400 children in third grade were assessed with the CPAA and with the NY Language Arts Test. The analysis revealed a significant correlation of about 0.7 between the two measures. (Additional information about the validity and reliability of the CPAA can be found on the Children's Progress website at www.childrensprogress.com.) This data, along with other data collected by the CPAA over the past several years demonstrates that the CPAA is a valid and reliable assessment.

Thank you for your time.

Sincerely,

Christopher Camacho, PhD



COMMENTS ON THE RACE TO THE TOP ASSESSMENT PROGRAM

General Assessment Input

Submitted by the College Board

December 2, 2009

Wayne Camara, Vice President for Research and Development, The College Board

(wcamara@collegeboard.org)

Kevin Sweeney, Executive Director of Psychometrics, The College Board

(ksweeney@collegeboard.org)

COMMENTS ON THE RACE TO THE TOP ASSESSMENT PROGRAM
General Assessment Input
Submitted by the College Board

The College Board is a national non-profit membership association of more than 5,600 schools, colleges and universities with more than a century of experience in the areas of standards and assessment. The College Board’s mission is to connect students to college success and opportunity, and it sponsors the SAT and SAT Subject Tests, PSAT/NMSQT, Advanced Placement (AP), ACCUPLACER, CLEP and other national assessments that reach more than seven million students annually. The College Board has strong partnerships with hundreds of states and school districts that rely on its assessments and other teaching & learning programs to prepare students for enrollment and success in college. The College Board has been a major participant in the Common Core State Standards project.

In the following text we would like to address seven major areas that the United States Department of Education should consider as it gathers information to inform the components of a request for proposals from states for a collaborative summative assessment program. These areas are:

- 1 Use the AERA/NCME/APA standards as authoritative guidance
- 2 Clearly define the purpose of the assessment system
- 3 Devise alternative methods of measuring student growth
- 4 Design a unified and integrated assessment system
- 5 Understand the importance of validity evidence in the design of the high school assessment
- 6 Incorporate innovation in the assessment system
- 7 Include teacher involvement, where appropriate

I. Use AERA/NCME/APA standards

The College Board recommends that the Department of Education formally recognize the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) as providing definitive professional guidance on the development and use of any assessments related to this initiative. In the request for input, the Department has called for “high quality summative assessments” that are based on “best practices in assessment.” In addition, the request appropriately requires that such assessments provide evidence relating to their validity, reliability, and fairness. The *Standards for Educational and Psychological Testing* have served as the definitive source for assessment professionals across a variety of applications (e.g., education, employment, licensure, psychological), and they delineate the appropriate types of evidence that are required to support statements by test publishers and users concerning these and other claims (e.g., comparability, use of cut scores). The Department of Education should ensure that any proposed summative assessments appropriately address these standards, and a technical oversight group should be established to review the proposed use(s) and evidence.

The *Standards* recognize that new assessments initially may not have all the documentation and evidence required to support inferences about validity. However, such evidence can be gathered over time and should be required of any assessment or accountability system. Indeed, because of the likely political and other pressures that will be placed on such a system, high quality validity evidence is essential to maintaining the integrity of the assessment system. In addition, the *Standards* note that “the applicability of the Standards to an evaluation device or method is not altered by the label applied to it, . . . the degree to which stimulus materials are standardized . . . or the type of response format (p. 3).” The *Standards* have been widely recognized as the definitive guidelines for the development, validation and use of assessments in a wide range of settings. They provide appropriate guidance and definitions on important technical issues associated with test development and use. It is important that the Department prominently cite the *Standards* in order to maintain a consistent level of quality, ensure common understanding about the types of evidence and documentation required, and ensure that any assessment practices adhere to current scientific findings and best practices. The alternative would be to allow each organization or consortium to define validity and fairness in its own way and thereby threaten the integrity and quality of assessments.

We recommend establishing a technical advisory committee of national assessment and content experts whose role would include adherence to the standards. The National Technical Advisory Committee (NTAC) or some other similar group could provide the Department with advice in developing RFPs and establishing criteria for their evaluation and use. It is important to note that, because not all aspects of the assessment system are driven by technical and psychometric issues, this committee should be advisory in nature and not a committee to determine final policy, although any ultimate policy committee should have representation from this technical advisory group.

II. Clearly define the purpose of the assessment system

Specifying the intended purposes of the summative assessment is the first step in designing a quality assessment. At least nine purposes have initially been mentioned in the Department’s call for inputs for the summative assessment:

1. To inform teaching and learning
2. To determine school effectiveness
3. To determine teacher and principle effectiveness
4. To determine student readiness for college and careers
5. To determine if a student is on track for college and career readiness
6. To measure student growth or change in achievement
7. To determine high school graduation
8. To determine college course placements
9. To inform college admissions

A single summative assessment or assessment system cannot serve all of these purposes equally well. There are tensions between many of these uses, and there are constraints that impose significant operational requirements for other uses. For example, summative assessments are not

designed to provide instructionally rich and actionable information. Typically, results are not available until the end of a school year, while diagnostic information is needed from the beginning and throughout the year.

Another constraint and conflict exists between the desire for innovative assessments that take advantage of technology and the use of the same assessments for very high-stakes individual decisions. Many state assessments are delivered by computer (although very few, if any, have achieved the desired goal of delivery exclusively on computer), but only when states permit schools to administer the same form (and/or items) over an extended testing window. There are simply not enough computers in schools to administer the same test to all 8th graders, for example, in a state on a single date (or even 3-4 different dates). School calendars also vary greatly within a state and flexibility in administration is required to accommodate local demands. Contrast this requirement with the security demands placed on tests used for college admissions, college credit and college placement. National testing programs have extensive procedures to ensure the security of test content and results for such high-stakes programs. The same items and forms cannot be administered over an extended window without greatly compromising security. In addition, the number of item pools and items required to maintain security of adaptive programs that offer the same level of flexibility for administrative dates would be cost prohibitive. These and other trade-offs need to be considered in determining the final requirements and purposes for an assessment system. The Department should identify a limited number of desired uses for a summative assessment system. In each instance, the consortium of states should then describe the types of evidence that will be used to support the validity of inferences that will be made for each purpose.

Testing at different grade levels may also need to take on different purposes. We believe that a summative assessment is not the best vehicle for providing diagnostic information to teachers and schools, and this issue is addressed later in the paper. However, a summative assessment in earlier and middle grades *can* be used to determine if students are on a path that will lead to college readiness. A summative test would ideally provide comprehensive information about student skills and mastery at a particular point in time, a measure of student growth during the academic year, an indication of whether a student has the knowledge, skills, and abilities required for success at the next grade level, and a metric that can be used as part of an accountability system for schools and teachers. At the high school level, a summative assessment may ideally be administered at the end of 10th grade to serve the above purposes, as well as to determine whether a student is prepared for college and career success. We believe that states should avoid attaching high stakes for students to this type of assessment during any transitional period. Moreover, when tests are used to determine graduation or college admissions, many operational and technical constraints arise that will reduce the flexibility and innovation desired for this assessment program. Graduation and admissions tests include a significant incentive to perform well on the test at all costs. Such proposed uses would require significantly more test items, test forms, and security, and they would also introduce significant operational constraints (limit dates of testing, require longer tests, greater reliability) and significant additional costs (more test items, more test forms).

III. Alternative methods of measuring student growth

Measuring student growth has long been an explicit goal of many state testing programs. However, because of technical, logistical, and cost constraints, this goal has been achieved with only mixed success. There are many lessons to be learned from the attempts to measure student growth, and we encourage the Department to speak with states and technical experts who have done it successfully, as well as with those who have not. Done properly, student growth data can be useful in both accountability programs and in providing information about individual student achievement; done poorly, student growth data will distort (either exaggerating or disguising) the amount of growth obtained.

Any meaningful discussion of student growth, however, requires a careful use of language. The term “student growth” is sometimes used as if its meaning were clearly understood by all parties and has a common definition. A cursory review of the research literature indicates that this clearly is not the case. Minimally, for example, student growth can be defined as relative to an achievement standard (e.g., student X scored five points closer to proficiency than on a previous test), relative to content standards (e.g., student X has displayed mastery on 4 of 5 objectives compared to mastery on 2 of 5 objectives on an earlier test), or relative to other students (e.g., student X is now at the 75th percentile, compared to the 60th percentile on a previous test). To be meaningful, all of these examples require that there be at least two points in time at which a student is assessed and that the results of these assessments be compared. It goes without saying that, for such comparisons, the results of the two tests must be comparable. A full discussion of what makes test results comparable is beyond the scope of these comments, but there are many important technical and logistical issues to be considered, and any assessment system purporting to measure student growth will need to work through these.

Each measurement of student growth provides answers to slightly different questions. There are at least three different questions that one can ask about student growth:

- 1) How much did student X learn this year?
- 2) How much more does student X know this year compared to last year?
- 3) How does student X compare to other students?

It is important to note here that each of these questions focus on different aspects of growth, and one cannot substitute for another. Consequently, we recommend that the RFP be clear as to what is meant by growth and what types of student growth are important.

Measuring student growth does not require the establishment and use of a vertical scale (i.e., placing test results from all grades onto a single scale). A vertical scale, while useful in many circumstances, has some limitations in measuring student growth in a K-12 standards-based assessment. Chief among these limitations is that in comparing, for example, the end of grade 3 to the end of grade 4, there is an important assumption that the grade 3 test is a good measure of grade 4 content (and vice versa). Because the content taught in grade 3 differs from that taught in

grade 4, this will rarely, if ever, be the case. With a vertical scale, any content that a student may have learned in grade 4 that does not overlap with content in grade 3 will not be captured in any measures of growth comparing end of grade 3 to end of grade 4. Grades 3 and 4 are used as examples here, but the logic applies to any pair or sequence of grades and may be of even greater concern in middle school and high school, where separate courses are taught (e.g., Algebra, Geometry). Several researchers (e.g., Lissitz & Huynh, 2003, Schaeffer, 2006) have discussed this issue extensively and make a compelling case for not using vertical scales in K-12 standards based assessments.

Despite their limitations, one of the reasons for the desirability of vertical scales is that they allow for statements of cross-grade growth. Often vertical scales have been adopted for this type of efficiency, and the instructional and curricular differences across grades have been overlooked. However, cross-grade growth can be measured in other ways (e.g., vertically moderated standards, growth percentiles), all of which have pluses and minuses. Whichever cross-grade growth model is employed (should one be employed at all), it is important that it be consistent with the stated purpose of the assessment system and that the strengths and limitations be clearly articulated.

In addition to cross-grade growth models, student growth can be measured—and depending upon the stated purpose of the assessment system, arguably, should be measured—via a within-grade growth model. This is consistent with a notion put forward by Laress Wise during his testimony in Boston. The measurement of within grade growth is a simple idea:

At the beginning of each school year, assess students on the material to be covered that year and use this initial measure as a baseline. At the end of the year, compare the end of year, summative test to the baseline measure to determine how much a student grew that year.

Various metrics can be established to ascertain how much improvement is adequate growth. This approach is direct in the interpretation of results and removes the troublesome problem of placing tests that measure different content standards on the same scale. Done properly, this approach can also provide initial diagnostic, actionable information about a student's areas of strengths and weaknesses at the beginning of the school year, when teachers can use that data to help students.

IV. Design a unified and integrated assessment system

We believe that the goals and intended purposes of this new assessment will be best served through an integrated assessment system that includes summative, interim and formative tests. In addition, we believe that the integrated assessment should be strongly aligned with the curriculum and that professional development will be essential to assist educators in connecting these elements. However, we will restrict our comments to the assessment system. The summative assessment can best provide useful information to students, parents, and schools on college and career readiness. Valid and reliable inferences can be produced for student and school level decisions. This information may also inform other decisions in time, such as course

placement, teaching and learning, and student growth or changes in achievement, if additional information is incorporated into the system beyond that collected during a single summative assessment. For example, a math test administered in 11th grade may not be the most precise way to predict how well a student will perform in a college math class some 18 months into the future. This is especially true when students score close to the cut point or when they fail to continue to take a math course in their senior year. Interim assessments can provide snapshots of how students are doing in mastering skills or providing more in-depth analysis of student weaknesses at a point in time. The formative components of such an integrated system can complement the summative and interim assessments and provide instructionally actionable information to schools and districts. A carefully designed integrated system is needed to ensure all components are complementary and consistent. Formative and interim assessments could utilize a common bank of assessment tasks and scoring rubrics available for teacher use.

The way in which the components of an integrated system are designed and work together will contribute greatly to the success or failure of the entire system. Consequently, although the current guidance is focused upon summative assessments, it would be short-sighted to not specify certain critical aspects of how the summative, interim, and formative components should work together.

Consistent with the above comments, we recommend an integrated assessment system comprised of the following three inter-related components:

- 1) Summative end of year
 - a. Grades 3-8: end of year
- 2) HS: end of domain (administered in grade 10)
- 3) Interim/Benchmark
 - a. Grades 3-8: minimum of 2 tests: baseline (at beginning of year) and midterm
 - b. HS: minimum of 4 tests:
 - i. Grade 9: baseline and end-of-year
 - ii. Grade 10: baseline and midterm
 - c. HS interim tests are not course specific but focus on college readiness
 - d. Test items are calibrated onto the same scale as the summative tests
- 4) Formative.
 - a. Most teacher involvement
 - b. Teacher scored

Ultimately, the summative and interim tests should be computer administered and, if possible, the summative assessment should be computer adaptive. The interim tests should be content focused and may not need to be adaptive. For the summative tests, the item types would be designed so that they are computer scorable. This summative design would facilitate: (a) quick turnaround of results; (b) increased use of innovative item types; (c) lower operational costs with higher fidelity items; and (d) greater ‘diagnostic-type’ information on college and career readiness.

For the interim and formative assessments, we recommend that decisions be made locally as to the item types and degree of teacher involvement in scoring. Allowing such local or state control will promote greater buy-in to the entire system and allow schools and districts to make the determination of valuing quicker turnaround time over teacher involvement in scoring more complex item types. Projects and performances can easily be integrated into interim assessments, and once they have been refined and evaluated, they could be integrated as a component of a summative assessment. However, this type of transition will require additional time to ‘try-out’ and evaluate the model and tasks, which is best done before they are incorporated into a summative assessment.

In the proposed system, items available for the interim assessments would be scaled onto the same theta metric as the summative test to allow for growth comparisons. These items would come from a common item bank, which would accept contributions from teachers and others. Projects, performances and extended tasks (e.g., out-of class assignments, in-class research) could also be included if standardized with well-developed rubrics for scoring. Additionally, it should be possible for off-the-shelf tests that demonstrate content and psychometric congruence to be used as interim assessments. In this instance, these instruments must be scaled (via a special study) to the summative scale.

Within this proposed system, all components should be designed to assess the same content standards. In this model, within-year growth can be measured by comparing the interim baseline assessments to other interim tests and baseline to the summative end of year test.

In this proposed system, the high school summative assessments would focus exclusively on college and career readiness and not be course specific. This is in keeping with one of the stated purposes of the assessment system.

The main advantages of the system outlined above are that it allows for measuring student growth, enables the measurement of performance against standards, and has the capacity to track students for college and career readiness. Additionally, it includes the capability of teacher scoring, but does not require it for interim assessments.

The main disadvantages of the system are that it requires universal access to technology, requires innovative item types to be developed and piloted, and requires a sophisticated database infrastructure to support relationships between interim and summative assessments.

V. Design of the high school assessment and the importance of validity evidence

The Department has stated that demonstration of college and career readiness is a priority of the RTTT assessment system, and that the high school test should focus on college and career readiness (CCR). We fully support this position and believe that the high school assessment should be consistent with this vision. A focus on CCR in high school, coupled with the options for differential course taking patterns in high school, is logical for this component of the assessment system. We must recognize that high school assessments must begin to evolve in a

different manner and design than the K-8 portions if assessments are to be relevant for higher education and career training programs.

Because the Department desires a system that supports the assessment of CCR, we recommend that assessments be made at the beginning and end of each grade to assess each student's status relative to college and career readiness. End-of-course assessments, while valuable, do not assess the same standards at the same level as an assessment focused on college and career readiness. If one wants to know the status of a student relative to college and career readiness, then assess that directly. End-of-course tests will present additional challenges to measuring student growth and obtaining agreement across schools, districts, and states. True "opportunity to learn" requires that students are allowed to take an end of course test at the completion of a course and not have to wait several years to take the test. This means that some students in middle schools may be taking the same Algebra and Geometry end-of-course tests as students in upper high school grades. It also means that schools may be administering different tests to different students in the same grade. All of these issues will complicate the use of such test results for school or teacher accountability. Our research often illustrates that students taking a test in 9th and 10th grade outperform students taking the same test as 11th or 12th graders. This phenomenon is more related to differences between the students than differences in school or teacher effectiveness. Students who are taking advanced math courses in earlier grades are generally at a higher ability level than the population of all high school students.

To assure that the defined purpose(s) of the assessment are being met, a comprehensive program of validity research must be established. Because there will be many pressures for test scores to be used for purposes that the system was not designed to support, it is important that validity evidence exist to support each intended test use and to refute possible improper uses—otherwise, appropriate and inappropriate test score uses become a matter of opinion and not a matter of fact. Such a state of affairs ultimately undermines the credibility of any testing program. The purpose of validity evidence is to establish the parameters for what are legitimate and illegitimate interpretations to be made from test scores, as it provides an empirical basis for the veracity of the claims being made. If the evidence does not support a particular interpretation, then there exists an empirical basis to refute bogus claims. Similarly, if the evidence does support a particular interpretation, then there exists an empirical basis to support such claims. In the best case, this evidence becomes foundational data on which solid policy decisions are made. Without these data, important policy decisions are based on untested beliefs and hearsay.

While validity evidence is among the most important information about a testing program, currently most state testing programs provide a very limited amount of validity evidence to support the claims made from statewide test scores (Sweeney, 2009; Sireci, et al, 2009). There are a variety of reasons for this state of affairs, chief among them being the costs and difficulty in obtaining good data to do strong validity work. Another limitation with current state assessments lies in the criteria. Current state assessments are designed to measure state standards not future outcomes. Therefore, the vast majority of validity evidence to support state assessments comes solely from a content validation strategy. States review assessment frameworks to ensure that they adequately map to state standards. When gaps are found between state assessments and standards, they are often justified as constructs that cannot be measured

with a summative assessment. Contrast this validation evidence with the type used in other settings (e.g., admissions, employment) where the outcome of the test score is compared against empirical outcome data. That is, admissions and employment tests incorporate concurrent and predictive validity evidence in their design because they are used to predict performance in a future setting (e.g., college, organization).

Given that the primary purpose of the high school assessment will be to determine if students are college and career ready, we believe high school assessments should be evaluated in large part by their relationship to student performance in college and career training programs. That is, states must break away from relying solely on subject matter experts to decide if their assessment frameworks are comprehensive and if their proficiency levels are rigorous. Instead, empirical results from future performance must be incorporated in this validation plan. Consequently, we urge the Department to make validity research a fundamental component of any assessment program and to provide funding specific to the collection of validity evidence. Validation efforts at the state and local level should not be used as the primary focus because, for example, students in New Jersey do not just think about going to college in New Jersey, but are often applying to public and private colleges throughout the country. College readiness results must be generalizable across colleges and states, and meta-analysis is a far more robust and superior validation strategy than supporting local studies that will produce slightly different results because of sampling and other methodological issues (Hunter and Schmidt, 2004).

In sum, we believe that external evidence must be collected in order to establish the validity of high school assessments in measuring CCR. Students who are considered “college or career ready” based on these assessments should be able to demonstrate college proficiency on a variety of external indicators. For example, students who are considered college ready in 11th grade should be able to attain a grade of 3 or higher on an AP course the subsequent year (or a corresponding grade in the International Bachelorette degree program). They should also be able to attain the prerequisite score on most college placement tests, and, ultimately, they should be much more likely to attain higher grades in freshmen courses across a wide range of colleges and universities. If students are deemed proficient based primarily on current content-based evidence and judgmentally derived standard settings, but do not achieve these outcomes, then it is evident that the tests and proficiency levels are simply not established as CCR. Similarly, if the proficiency levels are set so high that students who are successful on these external metrics are not deemed CCR, then the tests and proficiency levels are set at a level beyond what is currently required for post secondary academic success. The best way to evaluate the validity of high school assessments is by conducting large scale meta-analytic studies of performance across institutions (2-yr, 4-yr, career training) using external data on CCR. Certainly validation evidence based on content, construct, instruction, and consequences are also important in this effort, but predictive evidence is directly relevant in supporting future predictions. We do not believe that funding hundreds of local or state validation efforts will be effective, because the conflicting data by state and institution will lead to a false perception that CCR differs by state and institution and will lead to greater confusion among students and parents. In today’s global environment we must establish college and career readiness indicators that generalize across state and national lines.

VI. Incorporate innovation in the assessment system

Innovation can be realized most efficiently in a large scale testing program if it is delivered exclusively on computer. Innovative item types, extended performances and different response formats can be more efficiently captured and scored with the use of technology. Innovation in large scale assessment has been hampered by the requirement to produce comparable forms on paper. If the assessment is administered solely on computer (with the exception of paper administration as a special accommodation), it will be easier to introduce new item types such as simulations, scenario-based tasks, or performance tasks. Ideally such tasks in the summative assessment can largely be scored by computer to increase efficiency and reduce turnaround time. Teacher scoring of formative and/or interim assessments can be best utilized in a distributed scoring network or through an audit function.

Many of the emerging skills contained in the draft Common Core State Standards can likely not be measured with paper-based assessments alone. Maintaining parallel paper and computer systems would likely limit innovation and the range of emerging skills that could be measured. This is another example of the trade-offs that must be considered in the final design of assessment systems that will be proposed by state consortia.

Another option is to incorporate results from interim assessments or actual student performances that occur throughout the year into the summative assessment score. Currently, summative assessments are based on what a student does at the end of the year on a single test date, while some high performing nations have incorporated student performance at several different points in time into their summative assessment. Results from interim assessments or tasks completed during the year or student performance on a highly structured in-class or out-of-class assignment (e.g., research paper, literary report, laboratory report, presentation) that is scored by teachers using a detailed scoring rubric could be incorporated into the results of summative assessments. Clearly such models present operational challenges in terms of security and when students transfer into a school midway through the year, yet such models could increase the instructional relevance of assessments and work for the vast majority of students.

“Computer scored” items does not necessarily mean multiple choice items, but may also include simulations and other performance tasks that can be objectively scored.

An important question is how states can transition from their current assessment system to a more innovative and integrated system as proposed by the College Board and other leaders in education. Clearly each state will need to address the specific mechanisms for transitioning data systems, proficiency standards, and reports. State specific approaches may ultimately be required for many of the operational issues, but we believe that such a transition can be accomplished more easily if technology is incorporated in the assessment system and there is additional federal support to prepare states for such a migration. A transitional or interim approach may be required but it is doubtful that it will meet many of the desires expressed by the Department for comparability and a new generation of assessments that are directly mapped to the Common Core State Standards. It may be more effective to spend limited resources to fund development

on the integrated assessment and supporting curriculum for 2014 than to build an interim solution for 2012 that will fall far short of most goals and objectives.

VII. Include teacher involvement where appropriate

Teacher involvement will be a critical component of the success of any assessment system of the type proposed. However, current state testing program experiences have taught us that teacher involvement is most beneficial when it allows teachers to learn from each other and to develop skills needed in the classroom.

Several states have learned that teacher involvement in operational scoring of summative tests is neither cost- nor time-efficient. Including teachers in the scoring process for constructed response items is problematic from both a technical and logistical perspective. Logistically, getting teachers out of the classroom for the necessary training time and scoring time within the bounds of an operational testing program will likely be challenging. Technically, the characteristics that make one a good teacher are different from the characteristics that make one a good scorer; and good scorers are needed to assure that the assessment is scored validly and reliably. This is not to say that teachers cannot be good scorers, but merely points out that one's ability to teach and one's ability to score are independent attributes.

If the intent to include teachers in the scoring process is for the professional development of teachers, then this should be done outside of the operational scoring window. Goals of professional development can be better met by including teachers in the development of the constructed response or performance task items or in an audit function of the scoring of those items. Teachers play an integral part in the development and scoring of Advanced Placement exams. We believe that professional development that provides teachers with greater insight into the assessment frameworks and student performance levels can be accomplished through involvement in the development of assessment tasks and scoring of formative and integrative components in the short term. As described in an earlier section of this paper, the interim or formative tests are much better suited for teacher involvement than are the summative tests. As noted earlier, once such a system has been in place we can then examine ways to more effectively integrate interim or benchmark tasks or projects into the summative component, as well as use teachers for scoring these elements.

In closing, the College Board is pleased to have the opportunity to share these views, and we would welcome the opportunity to respond to any questions you might have about our comments. We deeply appreciate the Department's strong leadership in pursuing common core state standards and establishing common state assessments through this competitive grant process.

As an organization, the College Board has a history of working with states, districts, and schools in a variety of capacities related to assessment practices. For example, we currently work with the state of Maine in providing the SAT for use as its high school NCLB assessment and have statewide agreements for use of the PSAT/NMSQT and Advanced Placement exams.

Wayne Camara is the Vice President of Research and Analysis at the College Board. He serves on several state technical advisory committees and is a member of the technical advisory committee for the Achieve end-of-course algebra assessment. Dr. Camara is the president-elect of the National Council of Measurement in Education (NCME), which works closely with state leaders on issues of assessment.

Kevin Sweeney recently joined the College Board as the Executive Director of Psychometrics. Prior to joining the College Board, he worked for 11 years providing psychometric expertise in K-12 statewide assessments. In that work, Dr. Sweeney was responsible for the psychometric design and implementation of NCLB assessments for several states, including the Massachusetts Comprehensive Assessment System (MCAS) and the New England Common Assessment Program (NECAP).

References:

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). Standards for educational and psychological testing. Washington, D.C.: American Educational Research Association.

Hunter, J.E., and Schmidt, F.L. (2004). Methods of meta-analysis: Correcting error and bias in research findings. Thousand Oaks: SAGE.

Lissitz, R. W. & Huynh, H. (2003). Vertical equating for state assessments: Issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research, and Education*, 8, 1-10.

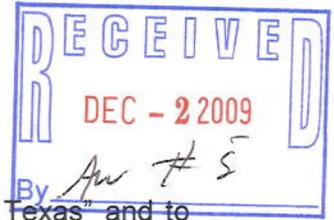
Schaefer, W. D. (2006). Growth scales as an alternative to vertical scales. *Practical Assessment, Research and Evaluation*, 11, 1-6.

Sireci, S.G., Meng Y., Hanwook Yu, H. and Zenisky, A. (2009). Building Validity Arguments for Educational Testing Programs. Paper presented at the annual conference of the Northeastern Educational Research Association, October 22, 2009, Rocky Hill, CT.

Sweeney, K. P. (2009). Focus on Validity: State Practices in Obtaining Validity Evidence. Paper presented at the annual conference of the Northeastern Educational Research Association, October 22, 2009, Rocky Hill, CT.

December 2, 2009

Good morning.



Thank you for the invitation to fly to the “Mile High City” from our “warm Texas” and to share with you our passion for students and for their success. I am here to discuss our Limited English Proficient student population – their challenges, their successes, and their assessments.

I represent *Responsive Education Solutions*, a public charter system that serves over 5000 students from kindergarten through 12th grade on 36 campuses throughout the state of Texas. Our *Vista Academies* serve students in elementary grades; *Quest Academies* serve middle school students; *iQ Academies* provide virtual programs; and our *iSchool High* and *Premier High Schools* serve middle and high school students.

Currently, over 90% of our Premier High School student population is classified as *at risk of dropping out of school*, and 43% of our Premier High School population is Limited in their English Proficiency (LEP).

2009 official data reports 90% of all ResponsiveEd students passed or projected to pass the Texas Assessment of Knowledge and Skills (TAKS). Eighty-nine per cent of the Premier High School student population passed /projected to pass the TAKS, and the same percentage – 89% – of the Premier High School LEP population passed or projected to pass their State Assessment. That's 89% of the entire student population, and 89% of the LEP student population. Thus, in 2009, ***Responsive Education Solutions closed the gap between the all student and LEP student performance groups.***

How did we close this gap? ResponsiveEd utilizes individualized, self-paced, teacher-assisted methodology to provide an environment in which each individual student can be academically successful.

Upon enrollment, each ResponsiveEd student is individually assessed to determine his or her academic strengths and/or academic weaknesses. English knowledge abilities are clarified; learning styles are determined; difficulties are diagnosed. This is done for every student in a personal and individual manner.

The ResponsiveEd student learns to set goals – his daily, weekly, and yearly goals. He begins with the end in mind – his high school diploma in hand, and his higher education and career goals established.

For our LEP student population, we conscientiously shelter instruction through strategy teaching and modeling. By appropriately scaffolding content, and by thoughtfully asking questions that require the LEP students to interpret, apply, and synthesize, we increase the likelihood that they will become critical thinkers.

Now, let's consider the performance for the overall Texas LEP student population in comparison to that of all students. In the latest Texas State Performance Report, though 91% of the entire state student population successfully mastered the Reading/ELA TAKS, only 72% of the Texas State LEP population was successful. In Math, though the general student population showed an 80% success rate, Texas state LEP students demonstrated only a 68% success rate.

In Writing, the differential was 93% down to 84%; Social Studies was 91% down to 63%; and in the all-important content area of Science though 74% of the total student population was successful, only 42% of the Texas total Limited English Proficient population demonstrated success.

ResponsiveEd closed this gap last year by treating each student as an individual – not just a “member of the class”. His needs are individual, his strengths are individual, and more importantly, he or she is an individual – unique and exceptional, the only “one of a kind” – and worthy of individual assessment and individual understanding.

So, what am I here to suggest, to share, or to offer to those who have spent their lives in pursuit of educational excellence? One thing – one very important thing. Today, you will hear much from many. You will hear only one important suggestion from us.

LEP students need more time. They need time to be recognized when they show growth, time to be appreciated when they show progress. Cognitive language acquisition takes time. LEP students will fulfill our high expectations if given time. ResponsiveEd LEP students did just that. They stayed the course, they progressed, and they grew in academic language.

LEP students will not necessarily do it all “at the same time”, “in the same way”, and “on the same designated timeline.”

When older LEP students are recognized only for “pass” or “fail”, and they fail, and this happens repeatedly, they invariably become discouraged. They consider themselves “failures”. They want to give up. They want to stop making the academic effort. They may eventually choose to drop out of school and to enter the workforce as “unprepared high school dropouts”. Many times this is the result of LEP students being required to “march to the beat of one, and only one drum” – lock stepping their way to high school completion ... or not.

When older LEP students are recognized for growth, for progress, for accomplishment, for success, they will logarithmically increase their academic performance to meet high expectations. We do not lower our high expectations. We expect continual growth; we expect continual progress; we expect them to meet these high expectations. It is the predictable result of cumulative growth and progress.

A LEP student may need to remain in high school five years or attend school through the summers – progressing and growing in cognitive academic language – in order to graduate. This student will need continual encouragement to see growth and progress as

successes which lead to the successful completion of his high school diploma. Recognition of growth through assessment structure can be one very important piece in a "Growth Enhancement Toolkit" that would not only assess and categorize, but also encourage and classify.

We need teachers who see their LEP student progress as success; we need administrators who see their LEP student progress as success; we need professional educational assessment developers who see LEP student growth and progression as success; and most importantly, we need our LEP students to be encouraged to see a continual growth and progress in knowledge and understanding as a measure of success.

According to the Texas Intercultural Development Research Association (IDRA), "Texas public schools are failing to graduate one out of every three students. Attrition rates as an indicator in a school holding power index show that the 2008-09 rate was 31% overall, and up to nearly 40 % for Black students and Hispanic students." 1.

According to the Census Bureau, between one in four and one in every five Hispanic students in the entire United States drops out of high school before receiving a high school diploma or GED. 2.

Improving school holding power is critical for our increasingly diverse public school population. Success breeds encouragement. Let's implement assessment systems that **reward a LEP student's continual academic growth**, thus encouraging him to complete his public high school education.

Thank you,

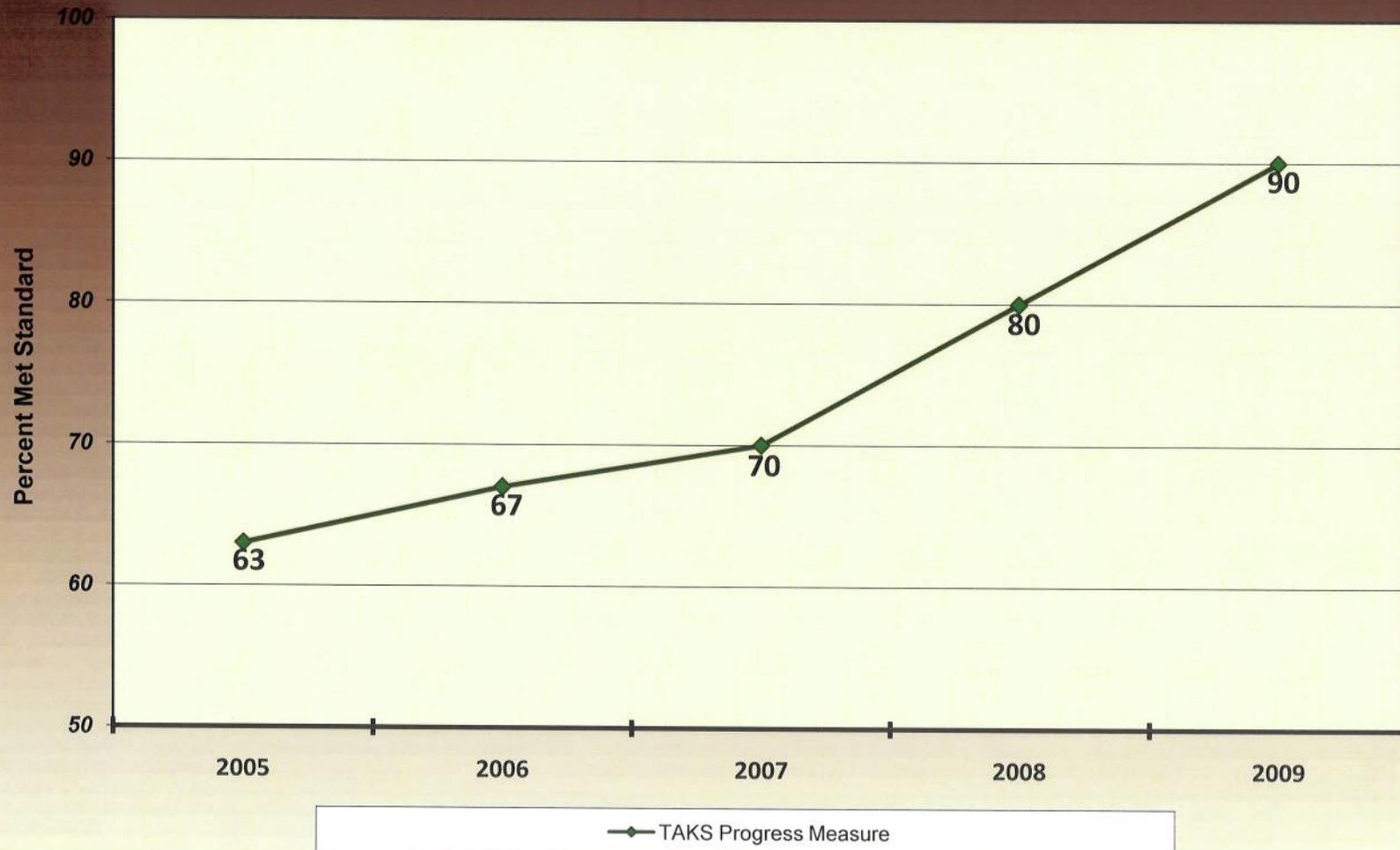
Julie Conde,
Director, Accountability/ESL
Responsive Education Solutions

SOURCES:

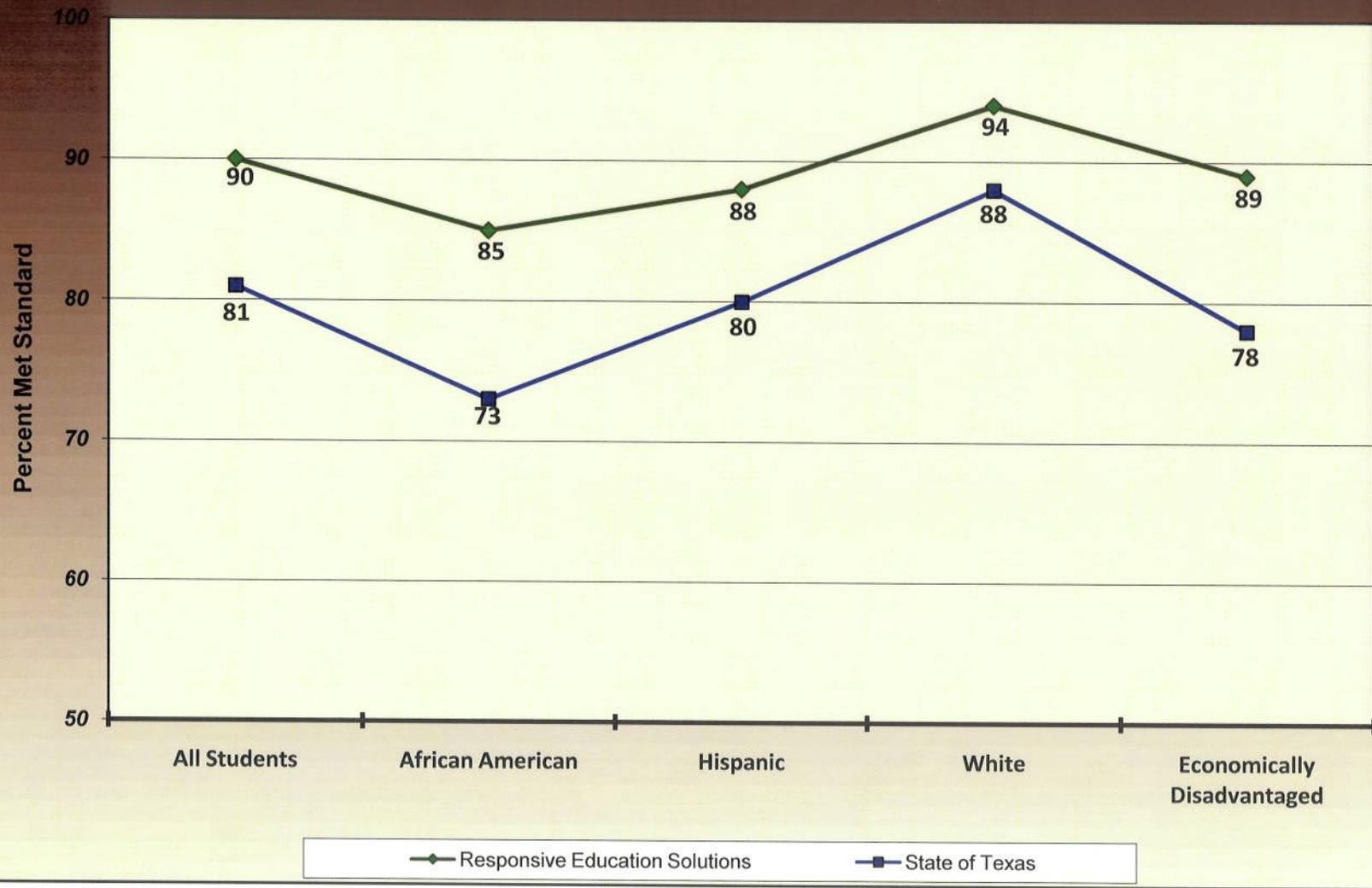
1. Roy L. Johnson, M.S.
Texas Public School Attrition Study, 2008-09
"Overall Attrition Rate Declines, But Gaps Persist Among Racial and Ethnic Groups"

2. U.S. Department of Commerce, Census Bureau, Current Population Survey (CPS), October 1994–2007.

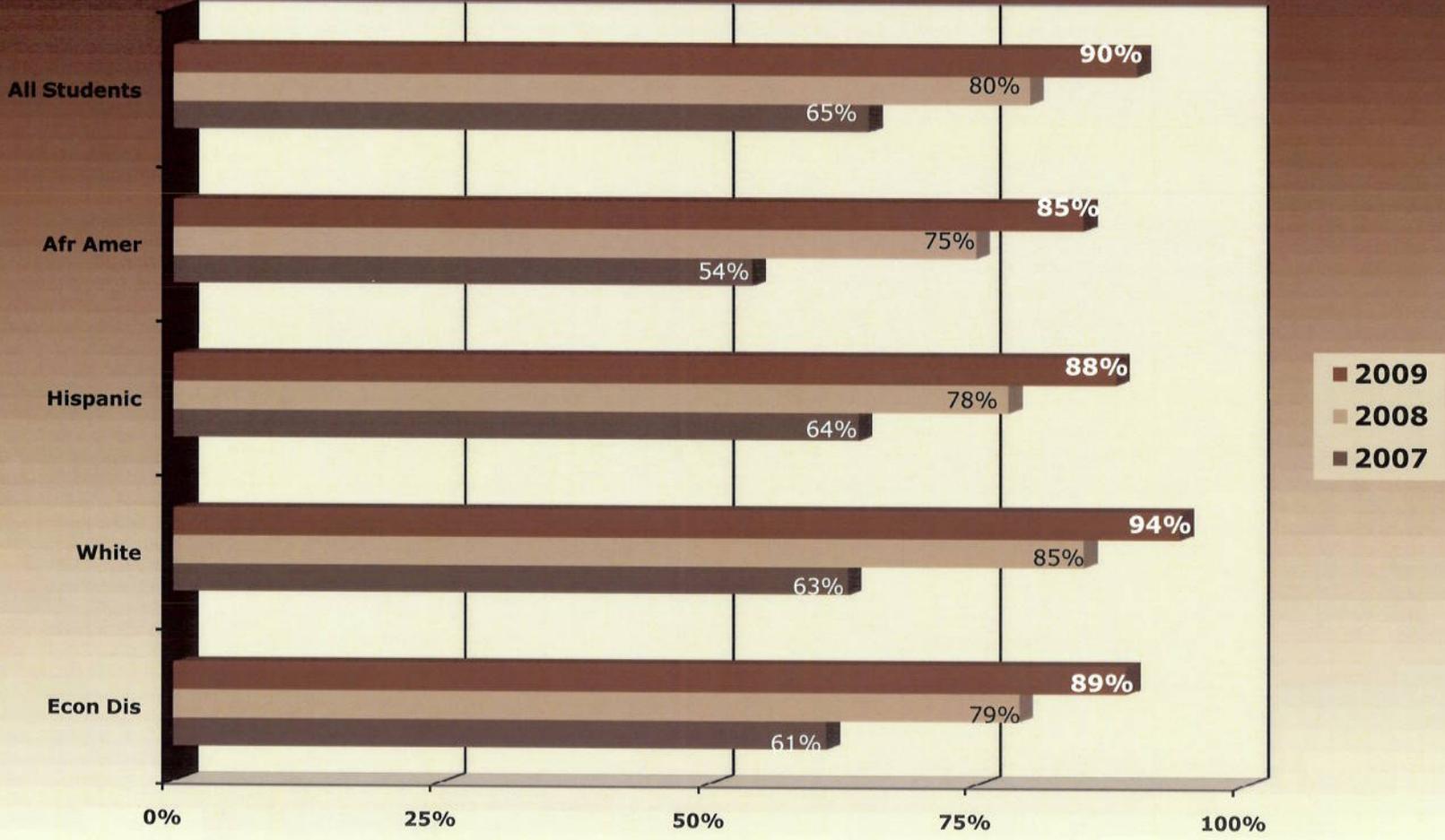
Responsive Education Solutions TAKS Progress Measure 2005 - 2009



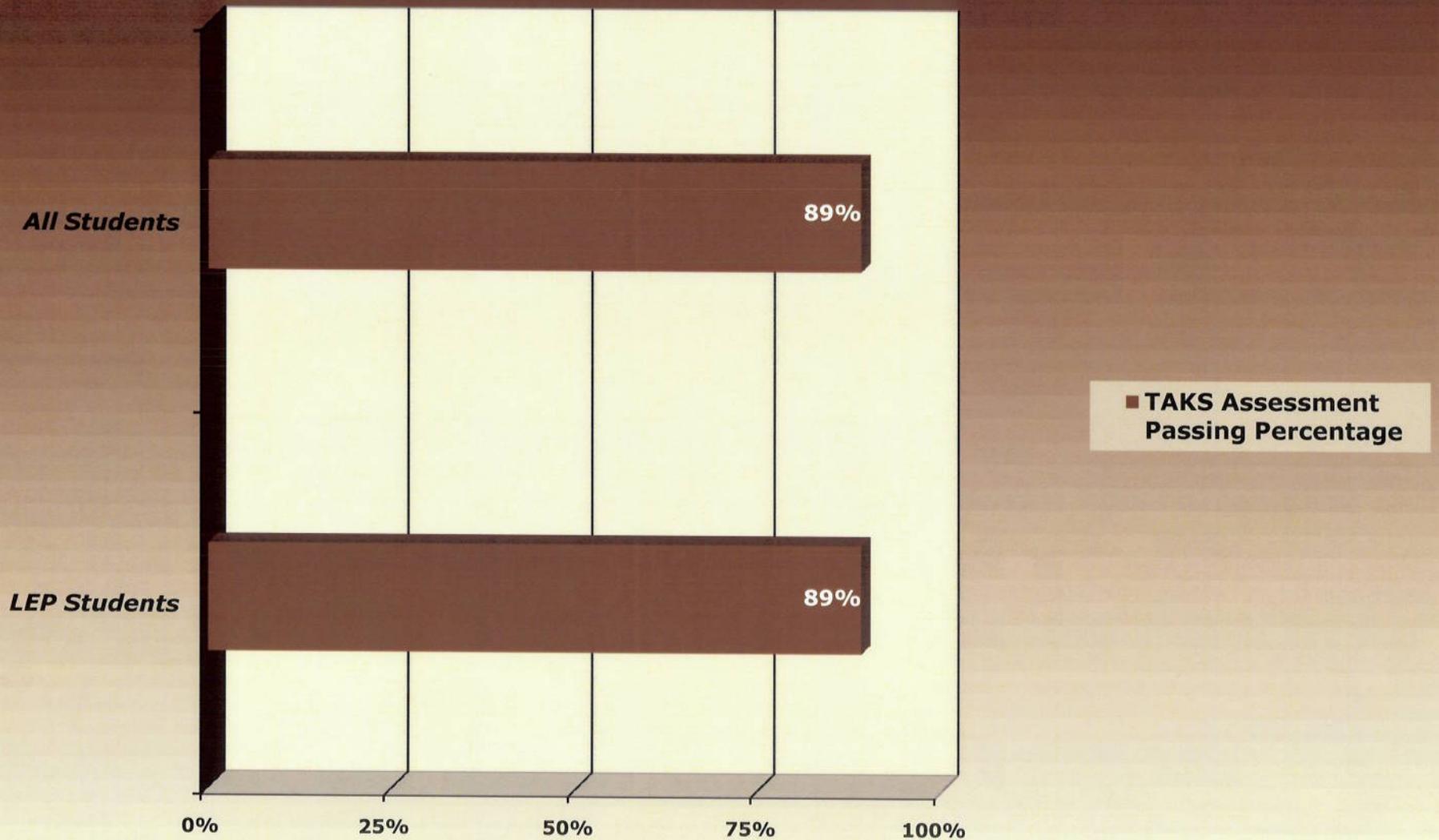
Responsive Education Solutions 2008-09 Statewide Rating Comparison



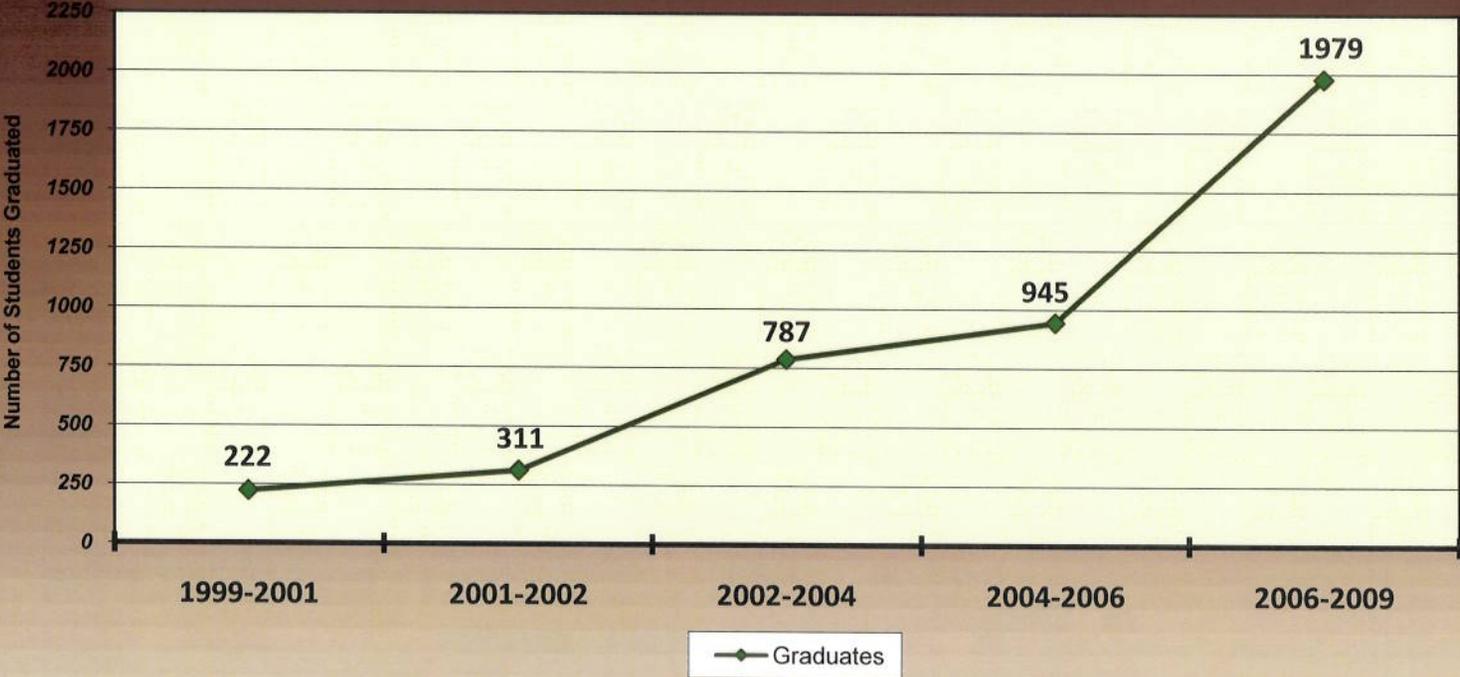
Responsive Education Solutions - District TAKS Progress Indicators 2007 through 2009



Responsive Education Solutions Premier High School Student Population 2009 TAKS Passing Percentage



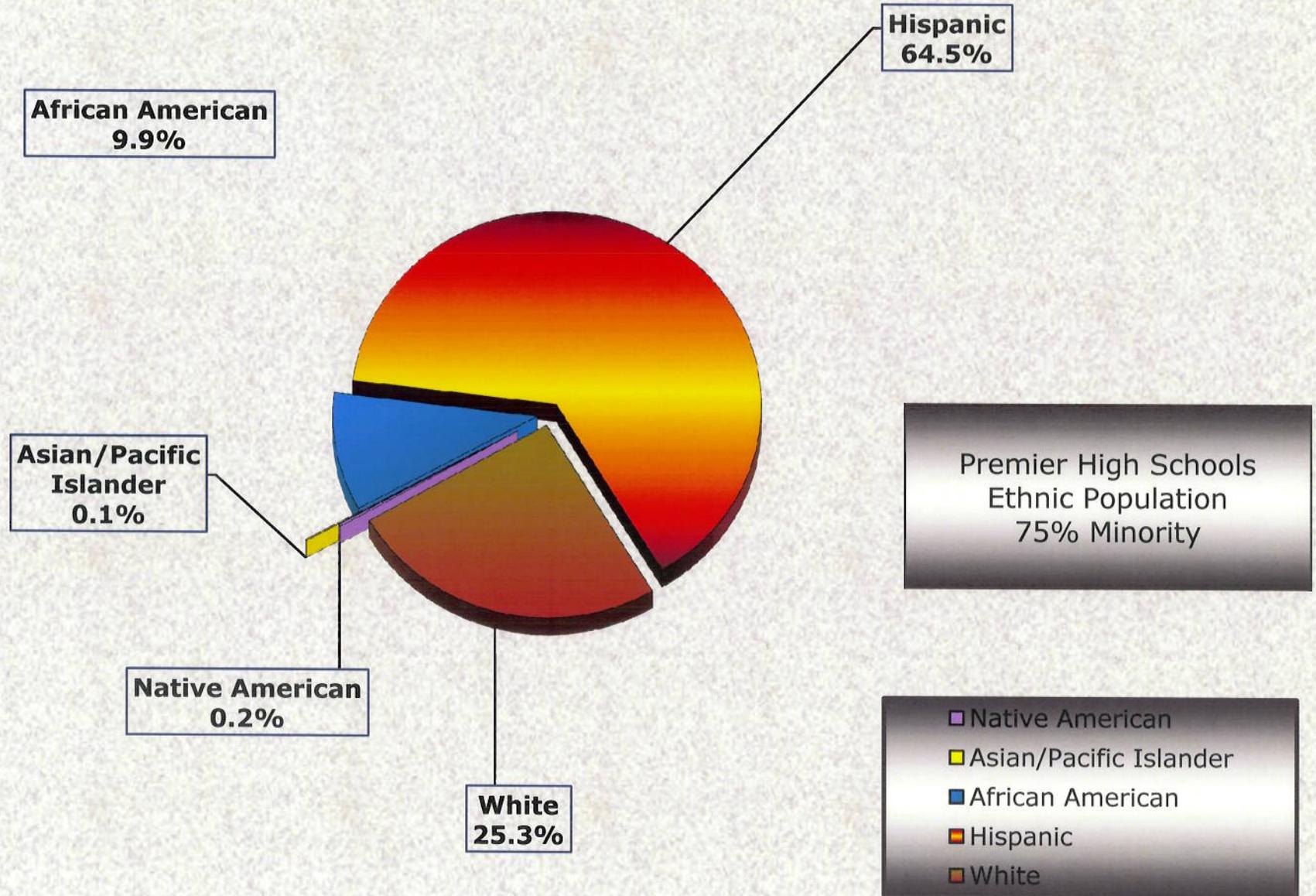
Responsive Education Solutions Graduates 2000 - 2009



**Total Graduates
4,244**

Responsive Education Solutions

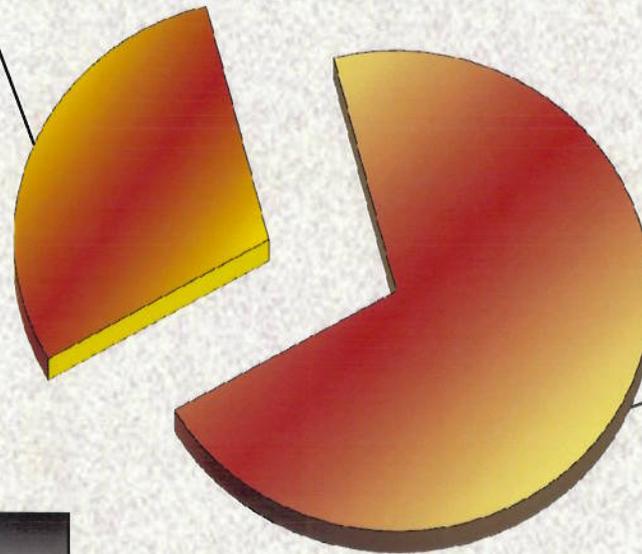
Premier High Schools Ethnic Distribution 2008-09



Responsive Education Solutions

Student Population by Age 2008-09

**18 years of age
or older -- Not
required to
attend school
29.4%**



**Under 18 years of
age -- Required to
attend school
70.6%**

*Total Students Attending in
2008-09
6520*
*Total Students 18 years of age
or older before end of year
1915*

Developing Assessment Systems that Support High-Quality Learning

Linda Darling-Hammond
Charles E. Ducommun Professor of Education
Stanford University

One part of a vision of what assessment systems should look like and do may be rooted in a set of principles that are shared by the systems of high-achieving nations, such as Finland, Singapore, Canada, Australia, Hong Kong (China), and the United Kingdom. There are several key themes that emerge from an examination of these systems:

1) **Assessments are grounded in a thoughtful, standards-based curriculum and are managed as part of a tightly integrated system of standards, curriculum, assessment, instruction, and teacher development.** Large nations like Canada, China, and Australia manage curriculum and assessments at the state or provincial level, while small nations like Singapore and England – which have school populations about the size of Kentucky and California, respectively – have national systems managed by a Ministry of Education. Each of these jurisdictions has undertaken a careful process of developing standards (generally described as curriculum expectations) and curriculum guidance, often in the form of syllabi, to guide teachers' instruction in the classroom, as well as professional development that is organized around the curriculum.

- Curriculum guidance is lean, but clear and focused on what students should know and be able to *do* as a result of their learning experiences. Assessment expectations are described in the curriculum.
- Curriculum and assessments are organized around a well-defined set of learning progressions along multiple dimensions within subject areas. These guide teaching decisions, classroom-based assessment, and external assessment.
- Teachers and other curriculum experts are involved an extensively vetted curriculum development process, and in the process of developing assessments grounded in the curriculum standards. These guide professional learning about curriculum, teaching, and assessment. Thus, everything that comes to schools is well-aligned and pulling in the same direction.

2) **Assessments include evidence of actual student performance on challenging tasks that evaluate standards of 21st century learning.** Curriculum and assessments seek to teach and evaluate knowledge and skills in authentic ways that examine a broad array of skills and competencies and generalize to higher education and multiple work domains. They emphasize deep knowledge of core concepts within and across the disciplines, problem solving, collaboration, analysis, synthesis, and critical thinking. As a large and increasing part of their examination systems, high-achieving nations use open-ended performance tasks and school-based, curriculum-embedded assessments to give students opportunities to develop and demonstrate higher-order thinking skills: the abilities to find and organize information to solve problems, frame and conduct investigations, analyze and synthesize data, and apply learning to new situations. The curriculum and assessment systems evaluate students' abilities in projects, group work, open-ended tasks, and oral presentations, as well as examinations that include

essays and open-ended tasks and problems, as well as selected response items, usually given at the end of a course or year.

3) **Teachers are integrally involved in the development of curriculum and the development and scoring of assessments** for both the on-demand portion of state or national examinations and local tasks that feed into examination scores and course grades. States invest in extensive moderation of the scoring process to ensure consistency and to enable teachers to deeply understand the standards and to develop stronger curriculum and instruction. The moderated scoring process is a strong professional learning experience, and officials believe teacher involvement drives the instructional improvements that improve student learning, as teachers become more skilled at their own assessment practices and their development of curriculum to teach the standards. The assessment systems are designed to increase the capacity of teachers to prepare students for the demands of college and career in the 21st Century.

4) **Assessments are structured to continuously improve teaching and learning**. Assessment *as, of, and for* learning is enabled by several features of assessment systems:

- The use of school-based, curriculum-embedded assessments provides teachers with models of good curriculum and assessment practice, enhances curriculum equity within and across schools, and allows teachers to see and evaluate student learning in ways that can feed back into instructional and curriculum decisions.
- Close examination of student work and moderated teacher scoring of both school-based components and externally developed open-ended examinations are sources of ongoing professional development that improve teaching.
- Developing both school-based and external assessments around learning progressions allows teachers to see where students are on multiple dimensions of learning and to strategically support their progress.

5) **Assessment and accountability systems use multiple measures to evaluate students and schools**. High-achieving countries use multiple measures to evaluate skills and knowledge needed for the demands of this dynamic, technological era. Students engage in a variety of tasks and tests that are both curriculum-embedded and on-demand, providing many ways to demonstrate and evaluate their learning. These are combined in reporting systems at the school and beyond the school level. School reporting and accountability is also based on multiple measures, including student achievement measures as one indicator among many. Other indicators often include student participation in challenging curriculum, progress through school, graduation rates, college-going, citizenship, safe and caring climate, and other indicators of school success and improvement.

6) **Assessment and accountability systems are used for information and improvement**. In most of these systems, student assessments are used to inform course grades, colleges, and employers, supports for individual student learning, and to shape curriculum improvement. The tests are typically not used to determine student graduation from high school; they set a higher standard linked to college and career expectations. Outcomes are publicly reported, and the

information is taken into account in a well-designed set of systems that focus on continual improvement for schools, including changes guided by school inspections and professional development supports organized by the Ministry or Department of Education.

Applying these lessons, and those from states that have previously developed assessment systems that have many of these qualities, as well as new knowledge from the leading edge of assessment development, we can imagine a systemic approach to transforming assessment of learning in the United States. In this new system:

The Federal Government would:

- Revise NAEP, using the new blueprints already established, to reflect the standards and more intellectually ambitious assessments of knowledge and skills
- Support research on the design, outcomes, and consequences of curriculum and assessments
- Allow, encourage, and fund the use of performance assessments for state assessment systems under ESEA, as well as the use of diagnostic and adaptive assessments that can better evaluate student performance over time.
- Support and fund initiatives to infuse knowledge of assessment and learning into pre- and in-service professional development.

States – working within Consortia -- would:

- Create Common Core Standards – mapped across the grade spans in a set of learning progressions around key dimensions of learning -- to serve as the basis for state curriculum and assessment efforts.
- Adopt and augment the standards as appropriate to their context.
- Create and deploy a curriculum framework that addresses the standards—drawing on exemplars and tested curriculum models.
- Build and manage an assessment system that includes both on-demand and curriculum-embedded assessments that evaluate the full range of standards and allow evaluation of student progress. Consortia of states might create joint assessments and an Assessment Bank of performance tasks linked to the standards that can be used as part of both on-demand tests and curriculum-embedded assessments.
- These would be accompanied by rubrics that embody the standards, and clear examples of good work, benchmarked to performance standards.
- Create an oversight / moderation / audit system for ensuring the comparability of locally managed and scored assessment components.
- Ensure that teacher and leader education and development infuse knowledge of learning, curriculum, and assessment.
- Implement high-quality professional learning focused on examination of student work, curriculum and assessment development, and moderated scoring.

Districts and schools – perhaps also working in networks or consortia – would:

- Examine the standards and evaluate current curriculum, assessment, and instructional practice in light of the standards.

- Evaluate state curriculum guidance, and further develop and adapt curriculum to support local student learning, select and augment curriculum materials, and continually evaluate and revise curriculum in light of student learning outcomes.
- Design, select, and incorporate formative assessments into the curriculum, organized around the standards, curriculum, and learning progressions, to inform teaching and student learning.
- Participate in administering and scoring relevant portions of the on-demand and curriculum-embedded components of the assessment system, and examining student work and outcomes.
- Help design and engage in professional development around learning, teaching, curriculum, & assessment.
- Engage in review and moderation processes to examine assessments and student work, within and beyond the school.

How Might A High-Quality Assessment System Operate?

Drawing from successful practices in the U.S. and abroad, a new assessment system might operate as follows.

Develop Curriculum Frameworks: When the Common Core standards have been released, vetted, and adopted, consortia of states would work with curriculum and assessment experts to develop (or adapt from previously successful work) curriculum frameworks mapped to the standards and learning progressions. There has been enormous investment in the United States in high-quality curriculum, for example through NSF and other organizations at the national level, and in many states and districts. Other English-speaking nations have also developed high quality curriculum materials linked to standards and learning progressions that should be evaluated in this process. This effort would inventory and cull from efforts with a strong evidence base of success in building out curriculum frameworks around which states can organize deeper curriculum development at the local level, state and local assessment development, instructional supports, and professional development.

Create a Digital Curriculum and Assessment Library: The results of this effort should ultimately be made available on-line in a digital platform that offers materials for curriculum building and, eventually, model syllabi for specific courses linked to the standards, formative and summative assessment tasks and instruments, and materials for training teachers and school leaders in both strategies for teaching specific curriculum concepts / units and assessment development and scoring. Assessment tasks linked to specific standards could be accessed from an Assessment Task Bank, like that recently developed in Hong Kong, so that they are available both for large-scale and classroom use. In addition, as described below, an electronic scoring platform should also be developed and made available across the states.

Develop State and Local Assessments: Initially, the state consortium would work to create a **common reference examination, which includes selected-response, constructed response and performance components** aimed at higher-order skills, linked to the Common Core standards for grades 3-8, like the NECAP assessment recently developed by a set of New

England states. This assessment would be designed to incorporate more analytic selected-response and open-ended items than many tests currently include and would include strategically selected curriculum-embedded performance assessments at the classroom level that are part of the summative assessment, while also providing formative information.

These curriculum-embedded components would be developed around core concepts or major skills that are particularly salient in evaluating students' progress in English language arts and mathematics. Exemplars to evaluate and build upon are already available in many states and in nations like England that have developed a set of "tests and tasks" for use in classrooms that help teachers evaluate students' learning in relation to well-described learning progressions in reading, writing, mathematics, and other subjects.

Curriculum-embedded components would link to the skills evaluated in the "on-demand" test, allowing for more ambitious tasks that take more time and require more student effort than can be allocated in a 2 or 3-hour test on a single day; these components would evaluate skills in ways that expect more student-initiated planning, management of information and ideas, interaction with other materials and people, and production of more extended responses that reveal additional abilities of students (oral presentations, exhibitions, and product development, as well as written responses).

In the context of summative assessments, curriculum-embedded tasks would be standardized, scored in moderated fashion, and scores would be aggregated up to count as part of the external assessment. Curriculum-embedded assessments would also include marker tasks that are designed to be used formatively to check for essential understandings and to give teachers useful information and feedback as part of ongoing instruction. Thoughtful curriculum guidance would outline the scaffolding and formative assessment needed to prepare students to succeed on the summative assessments.

A design much like this one was developed by the New Standards project in the 1990s, and has been implemented in states like Vermont, Kentucky, and Maine that have tied a set of performance tasks to a reference examination in English language arts and mathematics.

All components of the system would incorporate **principles of universal design** that seek to remove construct-irrelevant aspects of tasks that could increase barriers for non-native English speakers and students with other specific learning needs. In addition, designers who are skilled at developing linguistically supportive assessments and tests for students with learning disabilities would be engaged from the beginning in considering how to develop the assessments for maximum access, as well as how to design appropriate accommodations and modifications to enable as many students as possible to be validly assessed within the system.

The emphasis on evaluating **student growth over time** and on tying standards to a conception of learning progressions should encourage a growth oriented frame for both the "on-demand" examination and the more extended classroom assessments. Ideally, the reference exam would incorporate computer-based adaptive testing that creates vertically scaled assessments based on the full range of learning progressions in ELA and math. This would allow students to be evaluated in ways that give more accurate information about their abilities and their growth over

time. This approach should not preclude evaluation of grade-level standards, which could be part of any students' assessment, nor should it preclude a significant number of constructed response, open-ended items, as the technology for machine-scoring structured open-ended items is now fairly well-developed. As described later, strategic use of partial teacher scoring for these items would also be a desirable element of the system to support teachers' understanding of the standards and assessments, and their planning for instruction.

The emphasis on evaluating student growth should also inform the development of the curriculum-embedded elements of the system, which should be selected or developed to strategically evaluate students' progress along the learning continuum. Centrally developed tasks administered and scored by teachers with moderation (see below), using common rubrics, would be part of the set of reported examination scores. Existing tools like the Developmental Reading Assessment and the Primary Learning Record, which evaluate student progress along a learning continuum in ways that can inform both instruction and reporting, should be examined as well for their contribution to the classroom-embedded component of the assessment system.

In sophisticated state systems, it may be possible to begin to incorporate information about student learning that teachers develop from their own classroom evidence, linked to the standards and learning progressions and guided by the curriculum frameworks. This is the primary approach to assessment before high school in countries like Finland, England, New Zealand, and Australia. This approach is likely to be most productive of more sophisticated and adaptive teaching and well-supported student learning. This could be an optional aspect of the Consortium's work for states and communities with interest and capacity.

At the **high school level**, the Consortium might explore one or both of two options for assessment:

- **Course- or syllabus-based systems** like those in England, Australia, Singapore, Hong Kong, Alberta (Canada), as well as the International Baccalaureate. Generally conceptualized as end-of-course-exams in this country, this approach should become a more comprehensive course assessment approach like that pursued in these other countries. Such an approach would include within-course performance assessments that count toward the examination score, as well as high-quality assessment end-of-course components that feature constructed response as well as selected response items. Within-course performance assessments would tap central modes of inquiry in the disciplines, ensuring that students have the opportunity to engage in scientific investigations, literary analyses and other genres of writing, speaking and listening; mathematical modeling and applications; social scientific research. Such an approach might require an ELA and math assessment at a key juncture that evaluates an appropriate benchmark level for high school standards, and then, as in high-achieving nations, allow for pursuit of other courses/ assessments that are selected by students according to their interests and expertise. These could serve as additional information on the diploma for colleges and employers.
- **Standards-driven systems** that might include a more comprehensive benchmark assessment in ELA and mathematics complemented by collections of evidence that demonstrate students' abilities to meet certain standards within and across the disciplines. This set of

assessments would allow more curriculum flexibility in how to meet the standards. Systems like these are used in some provinces in Canada and Australia, in states like Rhode Island, Wyoming, Nebraska, and New Hampshire, and in school organizations like Envision Public Schools, New Tech High, Asia Society schools, and the New York Performance Standards Consortium. Sometimes these sets of evidence are organized into structured portfolios, such as the Technology portfolio in New Hampshire and the broader Graduation portfolios in these sets of schools that require specific tasks in each content area, scored with common rubrics and moderation.

- **A mixed model** could combine elements of both course- and standards-driven models, allowing some demonstrations of proficiency to occur in any one of a range of courses (rather than a single, predetermined course) or even outside the bounds of a course, like the efforts by some states to allow students to pass courses via demonstrations of competence rather than seat time (e.g. NH, OH). Such a system could also include specific components intended to develop and display research and inquiry skills that might also be interdisciplinary, such as the Project Work requirements in England, Singapore, and the International Baccalaureate, and the Senior Project requirements in Pennsylvania and Ohio.

Develop Moderation and Auditing Systems for Teacher-Scored Work: State consortia would develop protocols for managing moderation and auditing systems and training scorers so as to enable comparable, consistent scoring of performance assessments. In other nations' and states' systems that include these features routinely, procedures have been developed to ensure both widespread teacher involvement – often as part of professional development time – and to create common standards and high levels of reliability in evaluating student work. A range of models are possible, and the consortium would serve as a resource to individual states in developing and implementing strong, efficient approaches.

Provide Time and Training for Teachers and School Leaders: To implement an integrated system of curriculum, assessment, and instruction, time must be set aside for teacher development and participation in the system. Creative use of existing professional development days and incentives provided by recertification requirements (e.g. continuing education units) can be part of this commitment. In order to secure benefits for the quality of teaching and learning, states will need designate concrete commitments to support teacher engagement in curriculum and assessment development, scoring, and analysis.

Use Technology to Support the Assessment System: Technology should be used to enhance these assessments in a number of ways: by delivering the assessments; in on-line tasks of higher-order abilities, allowing students to search for information or manipulate variables and tracking information about the students' problem-solving processes; in some cases, scoring the results or delivering the responses to trained scorers / teachers to assess from an electronic platform. Such a platform may also support training and calibration of scorers and moderation of scores, as well as efficient aggregation of results in ways that support reporting and research about the responses. This use of technology is already being used in the International Baccalaureate assessment system, which includes both on-demand and classroom-based components.

In order to gain the efficiency and cost benefits of machine scoring and the teaching and learning benefits of teachers' moderated scoring, a mixed system would be developed where computer-based scoring is incorporated on constructed response tasks where useful – though teachers would score some of these tasks for anchoring and learning purposes – while other tasks that require human scoring engage most teachers in scoring to support improvements in instruction.

COMMENTS BY THE
ASSOCIATION OF TEST PUBLISHERS AND THE
ASSOCIATION OF AMERICAN PUBLISHERS

In Response to
74 Fed. Reg. 54795
October 23, 2009
Race to the Top Fund

The Association of American Publishers (AAP) and the Association of Test Publishers (ATP) file these comments in response to the notice published in the Federal Register on Oct. 23, 2009 (74 Fed. Reg. 54795). In its Notice, the United States Department of Education (“ED” or “Department”) proposes to fund “grants to consortia of States for the development of common, high-quality assessments aligned with an applicant consortium’s common set of K-12 standards that are internationally benchmarked and that build toward college and career readiness by the time of high school completion.” These comments are submitted timely by the due date of December 2, 2009.

Both Notices deal with the provision of \$350 million in funding for Race to the Top assessments, authorized under the American Recovery and Reinvestment Act of 2009; the October 23 Notice sets forth a number of questions directed towards the design and development of “assessment systems” to meet the needs identified in the initial Notice.

The AAP School Division is the principal trade association of the K-12 educational publishing industry in the United States. The Division’s Test Committee is comprised of many of the nation’s major test publishers and assessment organizations. The committee’s mission is to foster awareness of the role of testing in education, to promote appropriate use of assessments in education and to advocate public policy conducive to sound testing practices.

The ATP is the international trade association representing some 175 publishers and developers of assessments used in a variety of settings, including virtually every educational purpose under the responsibility of the Department of Education. The membership of the ATP includes both for-profit and non-profit companies and has served as the “Intelligent Voice for Testing” in providing input to the United States Congress, state legislatures, and federal and state agencies in their efforts to examine issues surrounding testing and the use of tests. Many ATP members provide testing products and services that would be affected by the proposed rules.

These comments are submitted on behalf of both trade organizations and their shared and separate members, which together compromise virtually all of the entities who provide testing products and services to the states under the No Child Left Behind Act and related laws and regulations administered by the Department. The U.S. testing industry supports the proposed Race To The Top (RTTT) assessment initiative, and we believe it will provide funding to support the development of testing systems that all stakeholders in our education system believe are desirable.

Our members have many decades of experience in developing and implementing complex assessment systems in all 50 states and the nation’s 15,000 school districts.

Those testing companies work closely with their SEA clients to ensure that statewide testing programs are implemented and operated in accordance with all federal and state regulations. The U.S. testing industry is comprised of educators, researchers, psychometricians, and technologists with extensive experience in developing and administering technically sound assessments that are used for many different purposes. We hope the Department recognizes the industry as a resource to be utilized as it shapes this important initiative.

We would like to address several topics raised by the Department pertaining to the design and development of proposed assessment systems.

I. Innovation

The US testing industry prides itself on the ability to innovate. We hope and expect that the RTTT assessment initiative will enable greater implementation of the innovations the U.S. testing industry has developed. For example, during the past two decades the industry has responded quickly and energetically to each education reform movement that has been enacted. The industry has been very responsive to the needs and demands of its customers. In the process, the industry has pioneered:

- Performance-based and portfolio assessments, in addition to constructed response and essays.
- Formative and interim assessments, as identified in the proposed RTTT notice.
- Technology-based student assessments administered online in addition to using paper and pencil, with technology-based scoring and assembly.
- Vertical scaling and growth measures, which in fact preceded the emphasis on alignment of standards in the No Child Left Behind Act and the Improving America's Schools Act as well as the current focus on growth indicators.
- Tests that provide both normative and criterion-referenced interpretations of student performance.
- International benchmarking.
- Tests of college and career readiness.
- Assessments for English language learners and students with disabilities adhering to universal design principles.
- Extensive and sophisticated data and reporting systems that allow teachers and principals – and parents – to monitor student performance and target interventions and resources to meet individual student needs.

One recent area of special innovative merit that the testing industry has undertaken is to develop a comprehensive set of operational best practices for statewide testing programs. These best practices, which have been developed jointly by the

Association of Test Publishers (ATP) and the Council of Chief State School Officers (CCSSO), complement the *Standards for Educational and Psychological Testing*, which address psychometric properties of tests and the technical aspects of measurement and assessment. AAP members who are members of the ATP have worked on this project with the CCSSO and the AAP fully supports these efforts. These Best Practices cover every element of statewide programs, from the RFP stage to program management, to item banking, to administration and test security, to scoring and reporting of test data, as well as including initial best practices for online assessments and the assessment of special populations (topics that will need to be updated in the future). Following a two-year development process by a joint Working Group of ATP and CCSSO members, the final draft of the Best Practices is being posted online by the CCSSO this week for a 60-day public comment period. We expect that many different state test stakeholders will provide their input and reactions to the draft document, which will be reviewed and considered by the ATP/CCSSO Working Group for inclusion in the final document.

Test publishers have accomplished all of these innovations by working in close collaboration with the nation's states and school districts. In the case of statewide tests that are required to meet federal accountability requirements, publishers have developed the tests in direct response to state RFPs that set out detailed descriptions of what the state is seeking and what they wish to include in their assessment systems. Unfortunately, funding constraints often limit the scope of assessment systems and the implementation of innovations.

There is a great deal of discussion about the "next generation of assessments." Much of that next generation is now available, but in most cases there has not been sustained funding for such assessment systems. We hope that through RTTT the Department will not only fund assessment innovations, but equally will foster them through policies that are not overly prescriptive.

II. Continuous Improvement

Another major topic of interest for the testing industry is for the RTTT funding to grow the capacity for continuous improvement. Over time, the testing industry has created and implemented extensive quality assurance systems. Quality assurance methods adopted by the testing industry include clearly defined scoring procedures and systems, reliable scoring technologies, ongoing training of personnel and constant oversight of the test scoring process. The operational best practices mentioned earlier will further augment quality assurance measures.

Fundamentally, the testing industry has endorsed the concept of multiple measures since the early days of education reform. In its testimony before the House Committee on Education and Labor on June 7, 1990, the AAP called for harmonizing strong technical quality with the need for multiple measures of what students can do and for ensuring that teachers and school building leaders obtain useful information from those assessments in order to inform teaching and learning. Moreover, AAP testified that the testing industry historically has been committed to the concept that assessment systems must be built based upon an identified purpose or purposes of each assessment in the system, and that assessments are built to measure specific identified content and must not be administered until all students have had a meaningful opportunity to learn that content. In a real sense, nothing has changed in the past 20+

years on this front – it is still vitally important that assessment systems be designed and developed around these principles.

Similarly, the ATP has advocated consistently for a multiple measures approach, well-identified test purposes, and the need to sequence test development, including in its work in 2000 with the Department’s Office for Civil Rights in developing guidance for use of tests for high stakes purposes. In a related vein, the ATP has urged Congress include professional training for teachers and principals on the use of assessments and assessment data as an allowable use of HEA funds – to ensure that those in leaders in the school and classroom receive more than in-service training in the use of assessments and assessment data.

III. High Quality Assessment Standards

The testing industry also firmly believes in high standards. Any federally funded assessment initiative should meet the highest psychometric standards in order to ensure validity, fairness, and reliability. It is imperative that formative and interim assessments, as well as summative assessments, meet these standards. Tests are used to make and inform decisions throughout the educational system. These decisions must be based on tests built by experienced professionals adhering to high standards. Furthermore, the testing programs and assessment systems also should be consistent with high technical standards, as well as the operational best practices.

IV. Preservation of Competitive Marketplace

The final issue upon which we comment is the need for competition. We strongly contend that any assessment initiative funded by the federal government allows for open competition. The current system is a highly competitive one where test publishers are constantly updating and improving their products and services in order to remain competitive. The results are innovation and lower costs to states and districts. We urge the Department to encourage fair and open competition through transparent procedures, and design the initiative so that the Department avoids a “winner take all” outcome.

Conclusion

The Association of American Publishers and the Association of Test Publishers appreciate the opportunity to provide these comments. We hope that our views will enable the Department to focus on sound, professional solutions for the use of Race to the Top funds in looking at state assessment issues. Regardless of how the Department determines to approve state grants for these funds, our members stand ready to work with their state partners and other experts to engage in a rational discussion about ways to come up with the best research and/or projects. We strongly believe that all such research, demonstration projects, or any other reports or outcomes funded by the Department, should be open to all states and publishers for future use in creating innovation and improving student achievement.

ASSOCIATION OF AMERICAN PUBLISHERS

Jay Diskey
Executive Director, AAP School Division
50 F St., NW, Suite 400
Washington, DC 20001
(202) 220-4549

ASSOCIATION OF TEST PUBLISHERS

Alan J. Thiemann
Legislative Counsel

LAW OFFICE OF ALAN J. THIEMANN
700 12th Street, NW, Suite 700
Washington, DC 20005
(202) 904-2467

BEFORE THE
UNITED STATES DEPARTMENT OF EDUCATION

RACE TO THE TOP FUND

Docket ID ED-2009-OESE-0006

RIN 1810-AB07 and 1810-AB09

COMES NOW, Measured Progress, Incorporated, and file these comments in response to the above-referenced Notice, published in the Federal Register on July 29, 2009 (74 Fed. Reg. 37,803). In its Notice, the United States Department of Education ("ED" or "Department") proposes to fund "grants to consortia of States for the development of common, high-quality assessments aligned with an applicant consortium's common set of K-12 standards that are internationally benchmarked and that build toward college and career readiness by the time of high school completion." Further, in its Notice published on November 18, 2009 (74 Fed. Reg. at 59,737 et seq.), the Department announced the priorities, requirements, definitions, and selection criteria it will use for the Race to the Top Fund. These comments are submitted timely by the due date of December 2, 2009.

Race to the Top Assessment Considerations: **Comments by Measured Progress, Inc.**

Introduction

Measured Progress appreciates the opportunity to provide input, which we have framed here as “considerations,” in response to the Federal Register notice on next generation summative assessments. We provide these suggestions in the spirit of supporting this effort and from the perspective of having worked closely with state departments and other entities to develop high quality, innovative, summative assessments.

Experience gained over 26 years has informed our ongoing work in Massachusetts, New Hampshire, Vermont, Rhode Island, Maine, Kentucky, Nevada, Montana, Utah, and New Mexico. In addition, we have worked with another contractor to develop end-of-course high school assessments in Georgia, and we deliver alternate assessment systems for six of the states above in addition to Florida, Washington, and New York. As a nonprofit, full service, assessment system provider, Measured Progress has established considerable credibility in the topic under consideration.

General Assessment Input

Many of the general requirements and required characteristics described in the notice are centered on issues currently being addressed in states’ assessment systems. As a publisher of customized assessments in response to RFPs from states, we offer that our preferred design is reflected in the Massachusetts Comprehensive Assessment System (MCAS), the New England Common Assessment Program (NECAP), and other programs that include a blend of a variety of item types (e.g. multiple-choice, short answer, constructed response, and extended response) tailored to the construct being measured. These formats are used to measure the extent to which students have met grade level expectations (criterion-referenced) related to common state content standards. It is worth noting that this is due to the fact that these programs were developed in collaboration with states around specific needs and requirements, and that design decisions arising from that collaboration were driven by the content standards and purposes of the assessment.

Releasing assessment items is strongly encouraged and limited only by states’ economic considerations. We have observed the benefits of using released items to foster awareness of the standards and familiarity with performance expectations, and we would strongly encourage expanded use of released items by schools and districts. Such use, along with examination of scoring rubrics and sample student work, are tremendously useful for professional development and for re-administration by local schools and districts.

Measured Progress promotes assessment literacy at all levels of the enterprise but especially focuses on LEA activities that build capacity in

Propose an assessment system (that is, a series of one or more assessments) that you would recommend and that meets the general requirements and required characteristics described in this notice. Describe how this assessment system would address the tensions or tradeoffs in meeting all of the general requirements and required characteristics. Describe the strengths and limitations of your recommended system, including the extent to which it is able to validly meet each of the requirements described in this notice. Where possible, provide specific illustrative examples.

effective school-wide and classroom-based assessment practices. This requires ongoing, focused, job-embedded, and systemic professional development. Our federal- and state-sponsored programs in Alaska (new principals' coaching and leadership), Michigan (classroom formative assessment, learning teams, and leadership), and Louisiana (effective classroom assessment practices and use of data) are excellent examples of collaboration in support of standards-based efforts to improve classroom instruction and student learning.

Since modeling good instructional practice is a primary design goal, we have consistently encouraged and developed systems to support the inclusion of open-response items. One way of expanding the role of assessment formats other than selected response, while building meaningful multiple measures into the design, is to include curriculum-embedded performance tasks that assess standards not so readily measured in current designs. These performance tasks would be scored by teachers, administered when appropriate at multiple intervals during the school year, and used within a structured audit process to strengthen the multiple assessment profile of student achievement.

Elaboration on this critical design element was provided in remarks by Stuart Kahl, our Chief Executive Officer, during the Denver hearing on general assessment issues. In these comments Dr. Kahl expanded on the value of multiple measures and teacher scoring, in the context of performance assessment.

Multiple Measures

We applaud the Department for its emphasis on multiple measures, a hallmark of good assessment practice. No testing expert, company, or user manual has ever failed to warn consumers that major decisions should not be based on the results of a single test. Nonetheless, despite the mention of multiple measures in NCLB, few, if any, states have done justice to the concept. For some, the term meant including two different item types in the same test. Many states have not even gone that far, due to the challenges of testing at all the required grades and meeting the timelines required by NCLB.

There is ample documentation of the impact that high stakes testing has on instructional practice. Therefore, we believe it is very important to use summative assessment design to model a process that yields actual student work so that teachers' tests in the classroom do the same. This allows and encourages them to evaluate that work to truly understand students' misconceptions and to modify instruction accordingly. We will continue to encourage states to include a healthy blend of constructed response items in their designs in support of this principle.

There is considerable discussion across the country of the possibility of additional interim, perhaps local, curriculum-embedded components being added to states' accountability assessment systems. We believe this is an

excellent direction. However, we believe that the Department should offer guidance as to what various components of accountability assessment should and should not be expected to accomplish.

Language in the Federal Register announcement about rapid turnaround and informing instruction can easily be misconstrued to mean having immediate implications for a classroom teacher while teaching a tested topic. We believe an on-demand, combined multiple-choice and constructed-response summative test is a valuable component of an accountability assessment program. However, such a general achievement measure cannot be expected to serve this more immediate formative assessment purpose. It could, however, affect teaching and learning through the use of its results to inform program improvement efforts, a longer term process.

Regarding a curriculum-embedded component of accountability assessment, a component that we would support, we believe the Department should make clear certain properties such a component should and should not have. A common complaint of local educators about end-of-year summative assessments is that they include items addressing content and skills that were taught six months earlier. They argue that tests students take during the course of instruction in a topic should count toward accountability results. We strongly disagree with this position. Schools should be accountable for seeing that students have retained important knowledge and skills. Thus, summative accountability testing should deal with retention, not short-term memory of students.

Taking this a step further, we believe that interim assessments that count toward accountability results should not cover material that can be tested via the more traditional on-demand summative measures. Many states have content standards that are not measured by their more traditional, on-demand summative tests – e.g., oral communication, research skills, media usage. These are the kinds of skills that curriculum-embedded performance assessment could effectively address. These assessments, to quote the Federal Register announcement, would elicit “complex responses and demonstrations of knowledge and skills consistent with the goal of being college and career ready.”

We believe the Department, in its solicitation, should make it clear that for purposes of accountability, interim assessments using traditional measures of knowledge and skills recently taught are not desirable since their results would not reflect what the students ultimately retain. Instead, they should tap important skills not readily assessed by the traditional, on-demand tests. (Note: There is a body of literature on how to conduct such performance assessments – i.e., how to ensure the quality and rigor of the assessment tasks and how to allow local scoring with centralized auditing to ensure scoring accuracy.)

If a goal is that teachers are involved in the scoring of constructed responses and performance tasks in order to measure effectively students' mastery of higher-order content and skills and to build teacher expertise and understanding of performance expectations, how can such assessments be administered and scored in the most time-efficient and cost-effective ways?

Teacher Scoring

The Federal Register announcement includes a requirement for assessment systems to involve teachers in the “scoring of constructed responses and performance tasks in order to measure effectively students’ mastery of higher-order content and skills and to build teacher expertise and understanding of performance expectations.” There is no question that involvement in such scoring constitutes one of the best professional development activities teachers can experience, and we commend the Department for including this requirement.

Given several of the requirements of high stakes, statewide testing, however, we recommend that teacher scoring not be “overdone.” If, for example, a state’s program includes an end-of-year on-demand assessment making use of constructed-response questions, we recommend the use of the testing contractors’ proven approaches to scoring – image scoring at contractors’ sites using experienced leadership and temporary scoring staff. Even though the scoring of images of student responses can be done on a fully distributed basis allowing anyone to participate in scoring from any location, maintaining scoring accuracy and meeting stringent timelines are more likely accomplished with the systems testing companies have established and operated for several years.

Occasionally, an article appears in the popular press finding fault with constructed-response scoring. These are written by individuals who are uninformed about what’s “under the hood” in these systems and the measurement quality they ensure. Oftentimes, the critics attack the qualifications of the scorers/readers. However, the systems, as they exist, apply high levels of expertise where it is needed, at the front end of the scoring process – in the development of the scoring rubrics and in the selection of student work corresponding to different score points for use in training and qualifying materials. This reduces the task of scoring to simple encoding or categorizing of responses, which many people can be trained to do effectively. After training, scorers must be qualified to score responses to each question by demonstrating an acceptable level of agreement between the scores they award to selected responses and the scores previously awarded by experts. Of course, scoring accuracy is monitored continuously during a scoring project by various forms of double scoring. The quality of the contractors’ scoring systems is well documented in the technical manuals for the assessment programs.

If, on the other hand, an accountability assessment program includes a locally administered interim component, such as a curriculum-embedded performance assessment, then clearly teacher scoring would be desirable. The scoreable products of such a component would be scored the same way as on-demand constructed responses, and in fact, products could include responses to follow-up constructed-response questions, along with reports, oral presentations, and other demonstrations of learning. A scoring audit process would also have to be implemented to ensure the quality of scoring. There would still be valuable training and generally the same quality of

professional development. Given the demand for multiple measures, including measures covering standards not easily assessed by on-demand tests, such curriculum-embedded components would provide the ideal opportunity for teacher scoring, thus addressing the two goals identified in the Federal Register announcement: measurement of higher-order skills and building of teacher expertise.

We recommend that the guidelines for Race to the Top assessment program funding refer to the “optimal combination of contractor and teacher scoring to complete scoring accurately and in a timely manner and to build teachers’ expertise.”

Technical Assessment Input

Measured Progress also firmly believes in high standards and seeks the best technical approaches for maintaining quality in every aspect of development and operations. Any federally funded assessment initiative should require the highest psychometric standards in order to ensure validity, fairness, and reliability. It is also imperative that all components of a summative assessment system meet these standards.

Further, we believe that future assessment systems should be consistent with both high technical standards and operational best practices. Measured Progress has taken a leadership role in a recent initiative to develop a comprehensive set of operational best practices for statewide testing programs. These Best Practices, which have been developed jointly by the Association of Test Publishers (ATP) and the Council of Chief State School Officers (CCSSO), complement the *Standards for Educational and Psychological Testing*, which addresses psychometric properties of tests and the technical aspects of measurement and assessment. These Best Practices, which cover every element of statewide programs from the RFP stage to program management, to item banking, to administration and test security, to scoring and reporting of test data, also include initial best practices for online assessments and the assessment of special populations.

Following a two-year development process by a joint Working Group of ATP and CCSSO members, the final draft of the Best Practices is being posted online by the CCSSO this week for a 60-day public comment period. We expect that many stakeholders will provide their input and reactions to the draft document. These will be reviewed and considered by the ATP/CCSSO Working Group for inclusion in the final document.

Taking into account the diversity of students with disabilities who take the assessments, provide recommendations for the development and administration of assessments for each content area that are valid and reliable, and that enable students to demonstrate their knowledge and skills in core academic areas. Innovative assessment designs and uses of technology have the potential to be inclusive of more students. How would you propose we take this into account?

Assessment of Students with Disabilities Input

The question of how to take into account the potential of innovative assessment design and technology to be inclusive of more students is one that has been a focal point of Measured Progress for over a decade. As mentioned earlier, we develop and administer a number of alternate assessment programs for states, and these programs utilize a variety of approaches to which innovative design and technology support have been applied. We have also been actively involved in numerous Enhanced Assessment Grants and General Supervisory Education Grants providing experiential and clinical data crucial to evaluating the effectiveness of such efforts. We believe this type of work should be supported and expanded, in concert with the development of next-generation summative assessment systems.

Dr. Sue Bechard, the Director of our Office of Inclusive Educational Assessment, presented testimony on this subject during the Atlanta hearings. She made three recommendations regarding features that Race to the Top proposals should include for the assessment of students with disabilities.

The first concerns assessment development. Race to the Top funding allows states the resources to develop a unified vision of the entire system, so a rigorous planning process involving all stakeholders should be required.

Race to the Top grants should require four components in a development plan:

- A clear articulation of the purposes of the assessment system and its components, which should be the same for all students and focused on improving teaching and learning. We strongly support the intent of NCLB to hold high expectations for students and require accountability for their learning, and we've learned that the best way to do this is to design a system that thoughtfully considers the diversity of all students in the planning phase.
- We applaud the considerations expressed in the common core standards initiative to include multiple measures of student performance. Since no one measure can serve all purposes, multiple measures, including outcomes beyond test scores, should be determined in the development phase. For example, decreasing drop-out and increasing post-school success in careers and college should be indicators to help determine the effectiveness of instructional programs.
- A strong evaluation plan should be articulated. We have anecdotal evidence that assessment of academic achievement has revealed the hidden potential of many students with disabilities. We also hear of instances where the curriculum is narrowed and students with disabilities are "blamed" for the poor performance of their schools. We know that assessment impacts curriculum and instruction, and studies to investigate both intended and unintended consequences need to be planned. Also, since transformation can only happen in the classroom, any proposal for

an assessment program should include investigations of the extent to which assessments are transforming instruction.

- Require that principles of universal design are applied to all components of the system, from the standards and descriptions of expected performance to assessment development and test construction. Development plans should describe how the assessments will provide multiple methods of presentation and expression and flexible options for engagement. For example, if a standard is stated in such a way that excludes students with certain disabilities (e.g., synthesize data, diagrams, maps, and other visual elements with words in the text to further comprehension), there should be a plan as to how students with visual impairment will be able to demonstrate expertise in that area. Or if reading standards specify demonstration of phonemic awareness and use of phonics-based decoding strategies, which is often seen in grades kindergarten through three, there should be a plan as to how hearing impaired students will be able to demonstrate how they are learning to read.

Second, two considerations regarding demonstration of knowledge and skills, validity, and administration of assessments for students with disabilities are pertinent to ensuring that valid assessments meet rigorous professional industry standards and provide real opportunities for all students to demonstrate what they know and can do.

- Race to the Top should provide opportunities to explore more meaningful ways to measure achievement of students with and without disabilities. In the current status model of summative assessment, it is difficult to measure achievement of students at the lowest and highest ends of the performance spectrum, as the preponderance of items are situated around the proficient/not proficient cut score to provide the greatest accuracy at that decision point for accountability purposes. Growth transcends grade level, and so should assessments designed to measure it. Measuring growth implies that there is an understanding of learning progressions or pathways that students typically follow as they learn and master key academic concepts. There are many gaps in the current research in this area for non-disabled students, and there is even less information regarding students with disabilities. There are many questions to be answered, such as: Are learning pathways the same for all students, but those with disabilities attain them at different rates? Do the information processing requirements for learning sophisticated academic concepts affect students differently depending on their disabilities?
- Proposals should describe precisely what academic content will be assessed. For example, we understand that when we present a student with a math problem that has text to provide a real-life context, we are also measuring to some extent the reading skills of the student. This may not be obvious if the student does not have difficulty with reading. If we are assessing students with disabilities that affect reading, we will not get

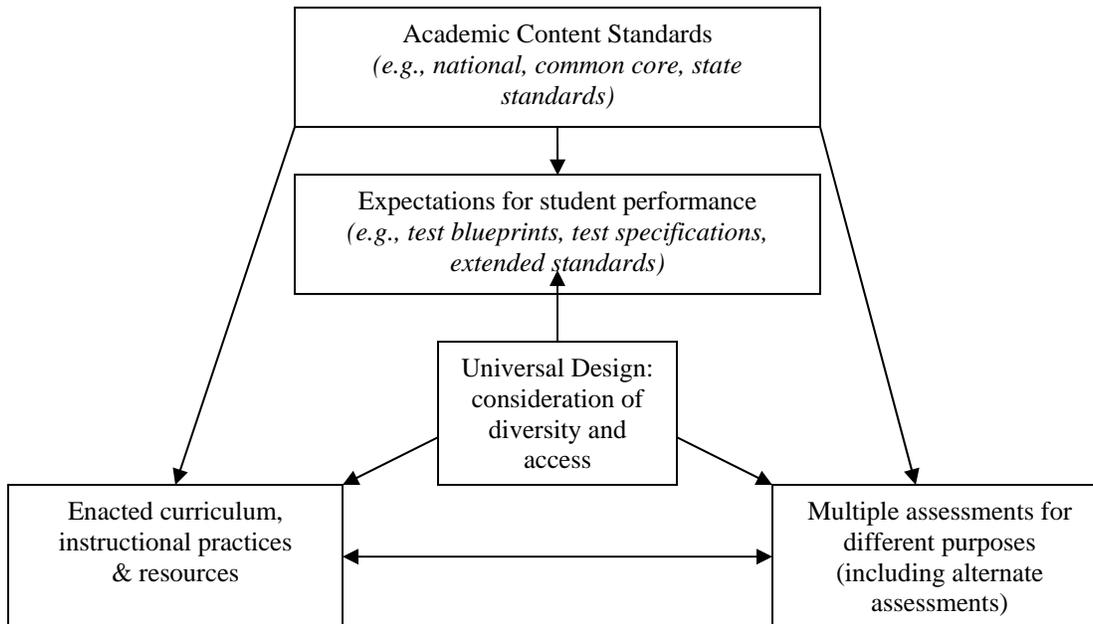
an accurate evaluation of math ability. In that case, we must find other ways to present the problem so that we are truly measuring the math skills we are targeting. A clear understanding of the relevant constructs will allow for the creation of multiple ways to present a problem or task and multiple ways for students to demonstrate what they know, enabling multiple views of growth in student learning.

Third, regarding innovative assessment design and technology, the promise of technology is not only about the speed of receiving results, but of more valid assessments for students with disabilities. Technology holds the promise of increasing access to assessments and of better measures of cognition and growth for all students, especially those with disabilities.

- Technology holds the promise of increasing access to assessment. Many students with disabilities use assistive devices and accommodations daily during instruction to access materials and demonstrate understanding. Assessments developed with technology have more opportunities to incorporate the same kinds of devices during assessment.
- Technology holds the promise of offering better measures of cognition and growth for all students, but especially for those with disabilities. Using technology for assessment will permit adaptive and scaffolding strategies to be employed. The delivery of items can be adaptive to accurately pinpoint the areas of competency and needs of individual students and to monitor each student's learning trajectory toward important milestones. These adaptive strategies must be cognitively-based, however, rather than based merely on item difficulties. By scaffolding assessment tasks, we can find out more about what the student actually knows and investigate the misconceptions they have when they are not successful. This will allow us to better understand the learning pathways for students at all levels of the performance spectrum.
- Finally, technology holds the promise of providing tools for educators, to deliver assessments for multiple purposes, for multiple audiences. It can be used to establish baseline performance data, to diagnose areas of need, to monitor progress, and to determine if students have reached performance milestones.

Race to the Top grants should allow for the time, the research, and the resources needed to develop assessments that are not burdened by high stakes, so that students can truly show what they know and teachers can determine better ways to teach.

A Unified System Design



Adapted from Flowers, C., Wakeman, S., Browder, D. & Karvonen, M. (2007). *Links for academic learning: An alignment protocol for alternate assessments based on alternate achievement standards*. Charlotte, North Carolina: University of North Carolina at Charlotte.)

Technology and Innovation Input

During the hearing on this topic in Boston, our president, Martin Borg, offered both general comments and specific recommendations on the subject of technology and innovation in assessment. These included recognition that while we've made good progress there is much that we can do to improve, and we have specific recommendations of how to go about it. We believe that there is a great potential for technology to promote individualization of instruction and assessment, which is especially important if we want to maximize and measure growth.

How would we recommend that different innovative technologies be deployed to create better assessments and why? Include examples: novel item types, constructed-response scoring, and uses of alternate input devices.

It makes sense to start with the student, the standards, and the measurement approaches before turning to technology, as technology is a means to an end. Novel item types, alternate input devices, and Web-based, distributed scoring networks for constructed-response items are readily available; the real key is to put them together in an intuitive interface that combines the different assessment methods and tools into a meaningful and interlinked whole. Some examples of proven, readily available technologies are:

- Adaptive testing
- Scenarios/games
- Machine-scored essays

In any design, successful implementation is based upon the notion that users get more out of the system than the work they need to put into it. Paying attention early in the process to the work styles and environments of teachers and principals will greatly improve the likelihood of adoption. Therefore, a large part of any design process should be usability studies at each level of user, providing a huge role for LEAs in the development program.

Measured Progress recommends that the Department fund opportunities to try out new approaches. We should move beyond mimicking a paper-based testing model to an assessment program designed from the ground up, taking advantage of today's rich technological environment. The following are illustrations of levels moving from computer-based testing toward the use of technology for maximizing student learning.

Level 1: Enhance paper-based testing with computers.

- It is common to use computer-based testing to streamline administration while preserving comparability with paper forms. The online environment is designed to closely resemble a test booklet experience.
- Machine scoring of essays essentially automates the existing paper-based approach.

Level 2: Leave paper-based testing behind and use only computer-based testing.

- This requires a higher level of integration with local infrastructure and expertise
- Adaptive testing provides individualized scores
- Scaffolded testing becomes feasible
- Scenarios and media-rich environments add depth to content
- Interactive problem solving is measured

Level 3: Use assessments to build learning relationships.

- Smart schedulers could track individual student proficiency and learning styles and match students with appropriate digitized content
- Specific teaching techniques could be recommended for a given student

Describe what a technology platform for assessment development, administration, scoring and reporting could and should offer. Platform should increase the quality and cost effectiveness of assessments.

A technology platform could and should have:

- A single platform that can offer summative, high-stakes testing using a variety of innovative and constructed-response items; fixed-form and adaptive testing can be used depending upon desired measure
- A pool of items, both multiple-choice and constructed-response, aligned to common standards with a library of tests available for interim assessment; teachers are able to build tests on the fly
- A library of classroom-administered performance assessments with embedded videos of what student work should look like and how it can be evaluated
- Considerable savings on scoring costs for interim assessments by simply expanding the distributed online scoring used by most test publishers. Built-in training and verification checks are already in place for these systems. Depending upon the measurement objectives, professional scorers, local teachers and /or a mixture of automated scoring could provide inexpensive and reliable results.
- A reporting center that measures and describes growth in a variety of ways, leveraging curriculum maps, student pathways, and standards maps, as well as values based on a scale
- Clear instructions and presentations that help teachers and other professionals understand the measures
- Embedded professional development at every level of reporting

How would we create this platform for summative assessments so that it could be easily adapted to be used by practitioners and professionals to develop, administer, and score high-quality interim assessments?

Much of the cost of administering online assessments comes from setting up the online testing environment. These costs can be reduced by using the same data—student i.d. numbers, student and teacher passwords, etc.—to power both formative and summative environments. The main differences between summative and interim tests concern test security and access to data. In summative, much of the administration is securely held at the state level; in interim tests, this is not as much of a concern. In fact, teachers need to see how students performed on specific items. Combining these systems is mostly a question of redefining user access to a single system.

Open source: If the federal government is funding the development of these systems, transferability is essential for the easy adoption of proven models. Measured Progress recommends that all innovative item development used in this effort becomes open source, so that item content *and* the item formats

and displays are transferable. This is easy with multiple-choice items, since no one owns how such a question is displayed. This may not be true with innovative items.

Common protocols and a single standard: To make sure that assessment systems interact with longitudinal data systems and student management systems, as well as other assessment platforms, the Race to the Top RFP should indicate a *single* standard. There are several to choose from.

Measured Progress remains firmly committed to technology and innovation in assessment as we craft a system that better informs student success and instructional accountability.

How would you recommend organizing a consortium to achieve success in developing and implementing the proposed assessment system? What role(s) do you recommend for third parties (e.g., conveners, project managers, assessment developers/partners, intermediaries)? What would you recommend that a consortium demonstrate to show that it has the capacity to implement the proposed plan?

Project Management Input

We are, to the best of our knowledge, the only organization to serve as a contractor for a consortium of states working together to meet the grades 3-8 and high school summative assessment requirements for NCLB. We have learned a great deal about what it takes to not only form a consortium, but how to make it work to the advantage of its members. Some characteristics of the New England Common Assessment Program (NECAP) that have contributed to its success are:

- Good will, trust, and a spirit of compromise are exhibited by all partners
- The states share a similar educational and assessment philosophy
- The states have relatively small student populations (although the largest was twice the size of the smallest)
- Geographic proximity has enabled a great deal of face to face time
- The states share some common goals:
 - Meeting NCLB requirements
 - Cost savings
 - Commitment to maintaining a high quality program
- Political will to make it work exists at the highest level (governors and commissioners)
- Cost sharing formulas have been developed, with some costs being distributed equally across the states (e.g. item development) and some costs distributed proportionally based on the numbers of students in each state (e.g. printing)
- The program is nearly identical in all states (this is essential if cost savings are to be realized)
- A program manager hired by the states, separate from the contractor, has been invaluable

- The timelines for building and maintaining a consortium need to account for the additional time it takes for more than one “client” to deliberate on the myriad of policies and procedures that go into a state assessment program

This program creates a wonderful opportunity. With widespread agreement that state assessment systems need to change and further agreement about the need for these systems to incorporate new components to accommodate multiple measures, including more complex and costly formats, the start-up costs associated with the development and implementation of new assessment systems would present an enormous challenge to states. The first year (or two) of any new program is always significantly more expensive than later, “maintenance” years, because of the additional planning, coordination, test development, logistics and analysis programming efforts required. The Race to the Top assessment program provides states the opportunity to secure funding for the start-up years of their new programs.

Elaboration on State Consortia

Near the beginning of the Federal Register announcement, the support of “one or more consortia” was mentioned. In other documents related to the program, “number of states” in a consortium was identified as a factor in funding decisions. We commend the Department for recognizing the benefits of consortia and encouraging their formation. However, we caution the Department against favoring large consortia for several reasons.

While there have been some relatively large state consortia in the past, they were focused on a limited population of students (English language learners) or a specific, well-defined course domain (algebra). NECAP is the only comprehensive assessment program serving a state consortium. NECAP has been very successful by all standards. However, that success did not come easily and there were a lot of factors contributing to it.

The original NECAP states were three small, like-minded, geographically compact states. A fourth recently joined the group. Their savings were substantial, allowing them to preserve quality, rather than diminish it because of a need to cut back on expenses during economic hard times. For example, they preserved their significant use of constructed-response questions requiring human scoring.

For small states, sharing the fixed costs equally, fixed costs being those for such things as program management, test development, analysis and report programming, was a tremendous benefit since for them, fixed costs were a large part of their overall program budget. The variable costs (printing, materials handling, shipping/receiving, human scoring) are those dependent on the number of students in a state. For very large states, these costs can be quite large, making consortia-related savings with respect to fixed costs relatively insignificant.

Savings with respect to variable costs are quite substantial for small states in a consortium because banding together creates economies of scale. For example, going it alone, a small state's constructed response scorers never get up to speed before they finish a question and start from scratch on the next one – not the case for large states. Thus, the large states already have economies of scale, so joining a consortium would offer more limited savings.

Geographic proximity of the NECAP states offered several advantages also. Management meetings of contractor and state staffs, test development committee meetings, and item and bias review meetings could be as often as needed, face-to-face, and low cost. The success of a consortium is all about relationships – the relationships needed to bear the larger burdens of reaching agreements, coordination, etc. With larger consortia, relationships are strained, with any one state's influence – and "ownership" – diminished. Also, as mentioned earlier, like-mindedness is critical. The more diverse the states in a larger consortium are, the more challenging the task of consensus building will be. Regarding the tests themselves, geographic proximity allows a regional flavor and greater relevance for reading passages and item contexts.

A letter report to Secretary Duncan from the Board on Testing and Assessment of the National Academies, dated October 5, 2009, makes a good case against the largest possible consortia (50 states). Decisions about federally mandated accountability assessments should not be based on a perceived need for comparability across states. There are too many obstacles to true comparability at both the national and international levels. Besides, NAEP gives us state comparisons that are as good as they're going to get. The problem with the percentages of proficient students being so variable across states and with many seemingly inconsistent with NAEP is that they show that there are some states that have set very low performance standards. All states performance standards should be high, not necessarily comparable. A national test is not needed to fix that.

In summary, we encourage the support of smaller consortia of states, perhaps 3 to 5 states, because of the "diminishing returns" associated with larger numbers of states joining forces, diminishing returns in terms of both cost savings (modest for larger states) and ease of management, consensus building, ownership.

Summary Recommendations

These closing points are offered for consideration with respect to next-generation assessment systems.

- Keep the focus on student learning, recognizing that
 - While summative assessments measure learning rather than create it, modeling good instructional practice is a primary design goal.

- Multiple measures of student learning are essential, whether the purpose is looking at student achievement, growth, or progress toward college and career readiness.
- The research on positive effects of formative assessment (assessment for learning) is about a process embedded in instruction.
- The goals of the larger educational endeavor and the standards that form its foundation should drive how student learning is measured.
- There is still much to learn about student learning, such that the development of measures of achievement will continue to be informed by advances in cognitive science.
- Base the assessment system design on the best knowledge available, to produce the highest quality instruments, methods, metrics, and reports.
 - Follow professional measurement standards and operational best practices, including the application of principles of universal design.
 - Inform the process with lessons learned from large scale assessment programs of the past and present.
 - Support ongoing research and development and create room for incubation.
 - Foster innovative methodologies and evaluate them based on well-documented high professional standards.
 - Acknowledge that there are existing state comprehensive assessment programs that are of high quality and for which longitudinal data should not be lost in the transition.
- Involvement of teachers and other stakeholders in a system redesign is essential.
 - In defining roles, take into account strengths as well as limitations, including time commitments.
 - Involve and inform various constituents with complementary elements of a coherent system.
 - Collaboration is strengthened by common purpose, while allowing for and celebrating diversity in approaches.
 - Inclusion of all students means removing barriers that may be created in the service of standardized accommodations (which opens a large opportunity for technology to help us drive toward personalization---striking an optimal balance in the process).
 - Federal, state, and local governance responsibilities matter.
- Credibility of future reports of improvement will be as high as the credibility of the measures.

- Include a rigorous process of validation and demonstration of reliability in qualifying next generation assessment systems.
- Transparency is essential when replacing, revising, or revamping systems used in any accountability function.
- An important aspect of credibility is using tools that were expressly designed to be used (rather than retrofitted) and are proven to be valid for the purposes to which they are applied, including measuring student achievement, growth, or progress toward college and career readiness.

Measured Progress appreciates the opportunity to provide these comments and believes that this initiative can have a major positive impact. The scope of our experience includes other important aspects described in the notice, such as High School Assessment and Assessment of English Learners. Our Project Management experience includes development, maintenance, and administration across a wide variety of programs as well as a major role in planning feasible development and implementation timelines. Our style is collaborative, and our clients are the best qualified to testify to the positive effect of our being “at the table.” We welcome the opportunity to support states pursuing Race to the Top funds to develop common summative assessments.



**RACE TO THE TOP ASSESSMENT PROGRAM
NOTICE OF PUBLIC MEETINGS AND REQUEST FOR INPUT**

**December 1 & 2, 2009
Denver, Colorado**

Questions on the Assessment of English Language Learners

The California Association for Bilingual Education (**CABE**) is a non-profit organization incorporated in 1976 to promote bilingual education and quality educational experiences for every second language learner in California. CABE has a 14 member Board of Directors, 5,000 members with over 50 chapters and a headquarters staff of 18 individuals, all working to promote equity, social justice and student achievement for students with diverse cultural, racial, and linguistic backgrounds.

CABE's members across California include parents, paraprofessionals, teachers, administrators, and researchers, who are committed to providing a voice for those who are silenced due to language, culture, or socioeconomic barriers.

Additionally, **CABE** has four statewide affiliates that work to further **CABE's** vision and mission as an advocacy oriented organization: **2-Way CABE**: implementation and technical assistance for quality two-way bilingual/dual immersion K-12 grade programs; **CAPBE** (California Association of Parents for Bilingual Education) a parent led affiliate with regional representative who work with parents at the local level in **CABE's** five regions; **CABTE** (California Association of Bilingual Teacher Education) who advocate for teacher preparation programs that fully equip new teachers with the skills, knowledge and attitudes to work with linguistically diverse background students; and **CASBE** (California Association of Secondary Bilingual Education).

CABE also recognizes and honors the fact that we live in a rich multilingual, multicultural, global society and that respect for diversity makes us a stronger state and nation.

CABE's vision "Biliteracy and Educational Equity for All" is based on the premise that in order to succeed and be powerful forces in their communities, students in the 21st century have to be: 1) Academically prepared; 2) Multilingual; 3) Multiculturally competent; 4) Technologically and information literate; and 4) Civically engaged and active advocates in their communities.

CABE's mission is "To promote and support educational excellence and social justice for all students in California;" Thus, on behalf of CABE, the following responses are submitted:

General Comments

To best determine the answers to the two questions posed by the Department, it is critical that the assessments and their variations be based upon an accurate English learner student profile through a strengths-based Theory of Action. This theory of action would account for the diversity of the English learner student population and mandates a variety of accommodations to specifically address the academic/content knowledge and language proficiency levels of ELs (e.g., native language testing, linguistic modification in English, etc).

The EL student profile should include indicators such as EL proficiency level, educational background in L1, and length of stay in US schools and program of instruction. Student profiles also would serve to simultaneously inform instruction and follow students to track developmental, vertical and horizontal progress for ELs throughout their schooling trajectories.

To uniformly apply the attributes of validity and reliability to each state's assessments, the Department should require that each state submit psychometric evidence from the test developers on the validity and reliability of the assessments administered to a wide-range of English learners.

Question 1

Provide recommendations for the development and administration of assessments for each content area that are valid and reliable for English language learners. How would you recommend that the assessments take into account the variations in English language proficiency of students in a manner that enables them to demonstrate their knowledge and skills in core academic areas? Innovative assessment designs and uses of technology have the potential to be inclusive of more students. How would you propose we take this into account?

We highly recommend the following:

1. First year beginning level students with little or no proficiency in English should be exempt from academic tests in their second language and the English proficiency test should serve as a proxy;
2. Recent immigrants (two years or less in US), speakers of indigenous languages, students with little or no schooling, students from war-torn countries with interrupted schooling and students without two consecutive years of educational experience in US (high mobility) should be exempt from taking academic tests in their second language for two years and the English proficiency test should serve as the proxy;
3. Assessments in reading/language arts need to be developed across the four language domains (listening, speaking, reading, and writing in L1* and L2) and across genres (narrative and expository texts);
4. There is a need to expand the types of performance-based assessments both by domain, by genres and by EL proficiency levels;
5. For elementary level students (grades 3-5/6) retellings (oral and/or written) in L1* and L2 be one pathway to assess students' comprehension and thus, allow students at different proficiency levels to demonstrate what they know and can do. The oral retell provides the opportunity for the teacher/school to gauge the ELD proficiency level simultaneously with reading comprehension. Scoring through valid and reliable instruments/rubrics such as running records, miscue analysis demonstrate growth and inform instruction;
6. Retelling can be captured by audio taping and can be scored by teams of teachers to ensure reliability in order to document growth;
7. Assessments need to inform instruction and go beyond filling in the bubble – performance based – e.g. writing in a variety of genre across all grade levels and content areas. Thus, assessment systems should include curriculum-embedded formative assessments as well as summative assessments;
8. Oral language development assessment needs to be embedded within the content standards. According to the National Literacy Panel for Language Minority and Youth, there is an absence of oral language development in instruction across all grade levels and content. "What gets tested gets taught;"

9. Linguistic complexity needs to be controlled in constructing tests for students beyond beginning levels of English proficiency and as they are developing English proficiency in all four language domains, e.g., especially for content area assessments in English;
10. The accommodations recommended by the *Technical Advisory Panel on Uniform National Rules for NAEP testing of English Language Learners* should be implemented by states to standardize the inclusion of English learners in federal accountability systems beyond on the NAEP testing;
11. Implement a temporary waiver of Annual Yearly Progress requirements while consortia engage in assessment reform;
12. Experts in English Learner education and assessment from all levels (universities, local and state education agencies and practitioners) should be actively included in the policy development and decision-making on assessment.

Question 2

In the context of reflecting student achievement, what are the relative merits of developing and administering content assessments in native languages? What are the technical, logistical, and financial requirements?

The relative merits of developing and administering content assessments in native languages are as follows:

1. L1* testing results in an accurate picture of what students know and can do for students who receive instruction in that language or for those who are already literate in their home language;
2. Given the national movement around world languages in preparation of a global citizenry, and that Spanish and Chinese are the top two world languages, L1 testing should align and support other initiatives promulgated by the federal government, e.g., World Languages and Strategic Language Initiatives;
3. Including native language assessments would reverse the punitive nature of the current accountability system by eliminating the practice of stigmatizing students, and labeling schools and districts as program improvement based on a single test that does not measure what many students really know and can do.

Recommendations for technical, logistical, and financial requirements:

4. Double test only in Language Arts (L1* and L2) and test in one language for content areas based upon language of instruction or preliminary assessment;
5. Use the native language tests from states that already have developed them. Also learn from their experience and build upon them –do not reinvent the wheel;
6. Use the new competition for consortia funds to do the developmental work on native language assessments and to develop tools and resources for the various accommodations;
7. Include experts in primary language assessments from all levels (universities, local and state education agencies and practitioners) in policy development and decision-making of native-language assessments.

L1*- indicates that students are proficient in their native language as evidenced by a home language survey and/or through instruction in the native/target language.

Respectfully Submitted,

Barbara M. Flores, Ph.D.
Coordinator of Bilingual MA Programs
CSU, San Bernardino
5500 University Parkway
San Bernardino, CA 92407

Director of Secondary/Higher Education
California Association of Bilingual Education
16033 East San Bernardino Road
Covina, CA 91722

Maria Quezada, Ph.D.
CEO & Executive Director
California Association of Bilingual Education
16033 East San Bernardino Road
Covina, CA 91722



Comments: Race to the Top Assessment Public and Expert Input Meeting: ELL Assessment

**Ellen Forte, President
edCount, LLC**

December 2, 2009

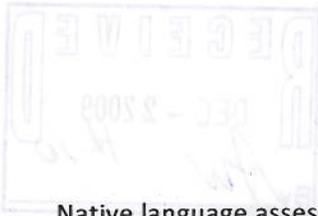
I'd like to thank the US Department of Education for convening this event and for allowing me the opportunity to share a few comments with this group. I hope to make four points in the next few minutes.

1. We have an obligation to provide better assessment opportunities for our ELs than those now widely used;
2. We must involve state and local educators in conversations about assessing ELs and not rely solely on advice from researchers;
3. We must not assume that better standards and assessments lead to better instruction; and
4. We must support efforts to integrate Titles I and III in part so that we apply the same rigor to the assessment of ELs as we have to the assessment of other students, including students with disabilities.

To the first point: we have an obligation to provide better assessment opportunities for our ELs than those now widely used. It's important to note that everyone here is passionate about high quality education for our English learners as for all other students. Where we differ is in our own life experiences and areas of expertise that contribute to our individual perspectives on how best to support ELs' education.

I come from the field of student assessment, where I have gained expertise in the development and application of rigorous tools to assess what students know and where I quickly learned that students who do not fit within certain parameters are generally cast aside, exempted from testing and often from the curriculum that testing was supposed to reflect. I also come from a family of educators who, albeit monolinguals, were specialists in communication and in the arts. From that background I acquired beliefs in the inherent right to express oneself and be heard, as well as in the responsibility of educators to find out what their students know and can do and to teach and evaluate based on that. Sometimes this means that you need to change what you ask, how you ask it, and what you recognize as a valid response. It means that no student is ever cast aside because you haven't the tools to connect with her.

But, that is what we continue to do. As an example, we continue to rely far too heavily on accommodations even though we lack strong evidence of the effectiveness of (a) most accommodations for any students and (b) any accommodations for some students. We know that accommodations are generally not chosen appropriately for individual students and that they are often not available at the time of testing. We even have evidence that some accommodations and some combinations of accommodations actually hinder students' performance. Accommodations are often likened to eye glasses for those whose eyesight is impaired. For ELs, however, eye glasses aren't going to work very well when the assessment is out of focus because language acquisition experts were themselves exempted from the standards and assessment development processes. We use accommodations because they are cheaper, not because they necessarily support score meaning for most users.



Native language assessments are also not necessarily the right option for many ELs because these students often lack academic literacy in their native language. The better answer is to rethink our assessment development processes and include consideration of ELs and of the interplay of cognition, language, text, and graphics from the outset.

Second: we must involve state and local educators in conversations about assessing ELs and not rely solely on advice from researchers. It is far too common to exclude those who actually work with ELs on a daily basis from the process by which policy decisions that have major implications for their students are made. Their involvement is especially critical when the teacher is an integral participant in the assessment process, as is the case for at least the speaking components of English language proficiency assessments and the accommodations and alternate assessment decisions for academic content assessments.

I must note here, however, that our teaching force as a whole does not have the expertise necessary to address adequately the linguistic needs of our ELs. This in-service deficit directly affects ELs' access to the curriculum and also means, in turn, that student teaching and mentorship opportunities for our pre-service teachers are extremely limited. We have work to do on building the capacity of our teaching force.

Third: we must not assume that better standards and assessments lead to better instruction. There are many layers between standards and classroom practice, including the dissemination of and professional development around the standards, teacher background knowledge and skills (i.e., do teachers know how to develop standards-based curricula and deliver instruction in relation to their standards for the students in their classrooms?), and teacher orientation (i.e., do teachers "believe in" the standards and in the relevance of their use to positive student outcomes?). Any assumption that a set of national standards or the adoption of the Common Core will translate into changes at the classroom level is entirely unfounded. The decision to exclude language acquisition experts in the development of the first set of Common Core standards was short-sighted and insulting. Again, presenting the problem in the first place is far better than applying a band-aid later on.

Finally, we must support efforts to integrate Titles I and III in part so that we apply the same rigor to the assessment of ELs as we have to the assessment other students, including students with disabilities. Inclusion of ELs in content assessments has only recently garnered much attention, and, as I noted earlier, we have yet to make significant progress in our fundamental understanding of how best to access what an EL knows and can do. Attention to the quality of English language proficiency assessments lags well behind that given to content assessments even in spite of some major improvements in recent years. The separation of policies and offices for Titles I and III has greatly contributed to this gap. We know that the majority of our ELs attend schools that receive Title I services and we know that some language support services targeted to ELs can also improve achievement among some low-achieving native-speakers of English. The best way to support our ELs is to embrace them and learn what they can teach us, not to isolate them.



**Talking Points: Catalina Fortino
United Federation of Teachers
Teacher Center Staff,
New York, NY,
On Behalf of the American Federation of Teachers,
To the U.S. Department of Education
Dec. 2, 2009**

My name is Catalina Fortino. I have been a teacher of English language learners for more than 20 years. As a practitioner, I am heartened that the U.S. Department of Education is paying close attention to the needs of our students and the educators who work with them every day. I welcome the opportunity to address you today.

Improving instruction and closing the achievement gap for ELLs will largely depend on the development and proper implementation of high-quality assessments. These high-quality assessments must align to standards, to curriculum and instruction and, crucially, align to the standards for English language proficiency. We need to have the ability to measure both English language proficiency and academic content knowledge. Only in this way can educators have accessible data to use for effective planning and students receive sound instruction.

I cannot emphasize enough how much we need ELL-focused reforms in schools around the country—in schools where ELL students are in the majority and in schools where only one or two ELLs attend.

The Race to the Top grants will be critical to those school reform efforts that include the development of improved assessments for ELLs. Current testing practices that assess ELLs' content knowledge in English are often not fair, not valid, and neither reliable nor appropriate.

Further, these testing practices make it very difficult if not impossible to distinguish between lack_of linguistic abilities in English or educational progress. Therefore improvements are greatly needed.

Research shows it takes students two to three years to become proficient in basic interpersonal communication skills, while it takes seven to ten years to acquire cognitive academic-language proficiency.

Schools are besieged by tests. Testing time reduces instructional time. Schools also lack resources. Not giving schools the resources they

need so much and the time, assistance and the preparation needed for their staffs to well-prepare for serving this group of students can have long-term detrimental consequences. This is especially true for those adolescent students on the threshold of graduation who will be going on to higher education or entering the workforce.

As to assessment, I know all too well the toll that a rigorous exam can take on ELLs who have not had enough time to learn the language.

I would like to add that in the early stages of language acquisition, research indicates that ELLs encode and decode text in English at a slower pace than they do in their native language. Furthermore, second-language processing demands very complex memory-recall processes, which may be compromised when an assessment is not matched to the student's level of English language proficiency.

While many factors must be taken into consideration to appropriately assess ELLs from preschool through 12th grade, the following actions are needed to help states improve their assessment practices and how they test students for English language proficiency and content knowledge:

- Statewide implementation of English language proficiency assessments that are aligned to English language proficiency standards
- Implementation of uniform, valid and reliable standardized tests of English language proficiency (such as the English language proficiency assessments developed by the WIDA—World-Class Instructional Design and Assessment—consortium of states. These particular assessments are research-based and aligned to English language proficiency standards that have been adopted by the states in the consortium)
- Ensuring that English language arts assessments are not used to measure English language proficiency
- Ensuring that content assessments are matched to a student's level of English language proficiency
- Ensuring that content assessments used for accountability purposes are also a valid, reliable and fair way to assess ELLs
- Ensuring that English language proficiency standards are aligned with state academic content standards
- Evaluating the current process involved in developing English language proficiency standards and assessments and making sure that the process is informed by research and best practices

- Evaluating the current process involved in developing and implementing the two types of assessments that ELLs take—English language proficiency and content assessments—and making sure the staff who is responsible for administering the exams has the preparation and resources to do it effectively

I would like to add: If content tests that are not matched to a student's level of English proficiency are used in high-stakes decisions, the results of ELLs who have not reached full proficiency will not be valid. Their scores would be at least as much a product of their language level as of their content knowledge.

The Race to the Top grants have the potential for schools to examine their current practices in educating ELLs and to implement assessments that are responsive and fair, making it possible for our students to succeed—not only to achieve academically but to become responsible citizens in our democratic society.

Thank you.

The Maryland State Department of Education welcomes the opportunity to participate with a broad consortium of states to develop a comprehensive, high quality assessment system to support the Common Core standards. Indeed, our goal would be to collaborate with all fifty states to maximize both effectiveness and efficiency. We hope that consortium efforts produce a unified and coherent assessment system, to include both formative and summative assessments that will support teachers in designing instructional experiences to increase student learning. We suggest that the competition for the assessment consortium explicitly call for both levels of assessments for reading and mathematics. Further, we suggest an emphasis in the competition on a high quality assessment system that is “unified and coherent.”

Considering the complexity of the task of designing both formative and summative assessments in reading and math, we respectfully suggest that instead of using a formula for the monetary awards requiring at least 50% of funds to go to LEA’s, that 100% of the funds go to the state consortium. District engagement in both the assessment development process and the pilot testing can be included but a more centralized management system might be more effective for the initial assessment development and pilot tasks. Without question, teachers, schools, districts and state education agencies must work together and we are confident they will in producing a quality assessment package in support of the Common Core Curriculum. Thus, we suggest the following alternative course:

Involve representatives from SEAs and LEAs in all aspects of assessment development, but the coordination of these activities would emanate from consortium leadership. SEAs would be able to identify LEAs best positioned to:

- a. Engage in pilot summative testing with different populations;
- b. Pilot the use of formative assessments;
- c. Investigate how best to involve teachers in scoring activities at the school level;
- d. Implement feedback systems for LEAs to provide meaningful, real-time data to assessment design teams;
- e. Create networks of schools and school districts across America to share best-practices in use of the unified assessment system.

The development of a technology “platform” required to implement the broad-based assessment design envisioned for the Common Core Standards represents a daunting task. Clearly, we believe that a single platform should serve the needs of ALL schools across the country. A single platform must serve multiple needs:

- Adapt to differing infrastructure landscapes encountered throughout the U. S.;
- Allow for the use of novel item types and assessment tasks that involve higher order thinking;
- Enable classroom teachers to evaluate and score student responses in an efficient and cost-effective manner;
- Provide assessment information to teachers in a timely manner that can inform future instruction based on student performance.

The complexity of a well designed technology platform that meets these needs requires significant cost. Thus, this supports our call for a centralized funding formula for development.

Maryland applauds the required characteristics of the assessment system, especially the “complex responses and demonstrations of knowledge and skills consistent with goal of being college and career ready” and “varied and unpredictable items types and content sampling.” Both of these characteristics are essential in measuring the Common Core Standards. We support the notion of teacher involvement

in scoring constructed responses and ask the Department to consider additional attention to guidance on how to accomplish this element while (1) meeting the timely assessment results to stakeholder groups and (2) ensuring inter-rater reliability in scoring and (3) ensuring time and quality of training to teachers to ensure their involvement is meaningful and ensures accurate results.

With seventeen years of experience in statewide summative assessments, including performance assessment measures, Maryland applauds the opportunity for multi-state consortia to engage in a more efficient use of funds to develop a comprehensive assessment system. Our annual \$50 million budget for reading and math summative assessments has served us well for the past 6 years but we are anxious to participate in developing an assessment system that will help us transition to quality measures of student learning to help us benchmark our student's growth against international standards.

Finally, we look for further guidance from USDE regarding the extensive work that will be necessary to move from Common Core Standards and Indicators at grades 3 – 8 and high school to the determination of specific assessment limits that will be necessary to guide assessment work. If state consortia will tackle this task, then clearly, costs escalate and the need for coordination at a central level becomes critical.



National Association for Gifted Children
1707 L Street, NW, Suite 550
Washington, DC 20036
(202) 785-4268
www.nagc.org

December 2, 2009

Honorable Arne Duncan
Secretary
U.S. Department of Education
400 Maryland Avenue, SW
Washington, DC 20202

RE: Race to the Top Assessment Program

Dear Secretary Duncan:

In recent months, the U.S. Department of Education's Race to the Top program and proposed reforms have generated overwhelming interest and action in determining the best path forward for improving student achievement in this country. The National Association for Gifted Children (NAGC) applauds the administration for sparking this dialogue, and we vigorously support the intent of the Race to the Top Assessment program. Focusing the grants on the design and quality of assessment systems to support improved teaching and learning rather than on accountability policies is a critical development. However, to achieve the goals of the Race to the Top Assessment program, projects funded must encourage the development of assessments that accurately measure learning for *all* students, not just those around the mean. To best reflect this aim, we have several recommendations for the proposed framework outlined in the October 20, 2009, Race to the Top Assessment Program Executive Summary.

(1) In the *Design of Assessment Systems – General Requirements* section

Recommendation 1: Rather than assessments focusing solely on the extent to which each individual student is on track, at each grade level tested, toward college or career readiness, we recommend that the framework acknowledge and incorporate the reality that **some students' achievement surpasses grade-level proficiency. Enhanced assessments must be able to accurately measure these students' knowledge, skills, and abilities in order to be able to show student learning gains.**

Currently, summative tests only track and assess student proficiency and mastery of an age- and grade-specific standard. Improved assessments will capture the extent of a student's ability so that educators can accelerate his/her instruction appropriately.

Current assessments are also limited in their ability to validly and reliably measure student achievement the further away a student moves from the mean. In other words, the assessment becomes less reliable at the upper and lower spectrums of achievement. For our most highly able students, access to tests or test items that are above grade level would provide educators with more precise understandings of how to teach and challenge these students.

Recommendation 2: NAGC recommends that the expectations for the information gathered from the assessments should be expanded to include "**Determinations of a student's current academic performance and future academic potential.**"

We applaud the efforts to boost college and career readiness, but we believe that the next generation of assessments should provide more precise understandings of each student's absolute achievement rather than achievement confined to his or her age and grade-level placement. Assessments that are able to accurately measure a broad spectrum of achievement, especially above the mean performance, will allow for more appropriate instruction and planning for students who may be on a different learning and graduation trajectory than their age peers.

(2) In the *Design of Assessment Systems – Required Characteristics* section

Recommendation 1: NAGC suggests that the framework make clear that the new generation of assessments must be valid, reliable, and fair for all students, including those already at or above grade level.

Because the current federal focus in NCLB, as well as recent federal assessment pilot projects, is and has been on students performing below or near proficiency levels, it may not be clear to grant applicants that the next generation of assessments must be able to measure the performance of all students, including advanced students who typically score at the highest percentiles of their state tests. Clearly, the goal of using assessment data to plan instruction for students cannot be met for top students if the assessments cannot accurately measure their knowledge and skill levels.

(3) In the *Design of Assessment Systems – Desired Characteristics* section

Recommendation 1: NAGC recommends including **teachers who teach gifted and talented students into item # 1**. NAGC strongly supports the inclusion of teachers in scoring of constructed responses and performance tests in order to effectively measure student mastery of higher-order content and skills and to better understand performance expectations. We believe that it is critical that teachers with expertise in instructing students who are gifted and talented would ensure that assessments and instruction improve for these students.

Recommendation 2: NAGC supports item #4 on building the technology infrastructure, but recommends that the item notes that **the assessments should be calibrated to the performance levels of the individual students, including those performing at the top end of the achievement spectrum**.

(4) In the *Design of Assessment Systems – LEA-Level Activities* section

Recommendation 1: NAGC recommends **adding "gifted and talented students" to the list** of student populations that should be included in the pilot test of the new assessments to ensure that new assessments are able to measure learning growth for those students at the top end of the performance scale.

Recommendation 2: NAGC recommends adding the following to the list of activities for participating LEAs: **"Designing systems that support students who surpass proficiency levels with flexible modifications, informed by the new assessment data, to course and grade level placements and other strategies to meet their learning needs."**

(5) After the *Question on the Assessment of Students with Disabilities*

Recommendation 1: NAGC recommends adding a new section entitled **"Questions on the Assessment of Gifted and Talented Students."** NAGC strongly believes that gifted and talented students have unique learning needs that are not being met in many schools under current conditions. Directly incorporating gifted and talented students into this assessment reform package would be an important step toward meeting the needs of these students. This view of gifted and talented students is consistent with the newly adopted definition of "teaching skills" in HEOA, which recognizes that this population has unique learning needs.

Thank you for the opportunity to comment on these important reforms. We appreciate your efforts to improve this critical area for our country's students.

Sincerely,



Nancy Green
Executive Director



Date: 12/02/09
To: Department of Education
From: Partnership for 21st Century Skills
Re: Race to the Top Assessment Program - Comments
Submitted by: Valerie Greenhill, Director of Strategic Initiatives, P21

The Partnership for 21st Century Skills (P21) applauds the proposal to create an Assessment Program as part of the Race to the Top initiative. We are pleased to see the attention given by the Department to assessments that ensure career and college readiness. The necessity of measuring complex tasks and higher order thinking is expressed in the early documentation about the Race to the Top Assessment Program, and we strongly support this emphasis. Twenty-first century assessment systems—including all forms of measurement—must assess the key dimensions of 21st century learning; they must measure those skills now prized in a complex global environment.

Although the Common Core State Standards Initiative drafts have not been finalized, we are hopeful that the inclusion of problem solving, critical thinking and communication skills will remain prominent indicators within those standards. The need to define “career and college readiness standards” in ways that address *both* core academic subject mastery *and* essential college and career readiness skills is critical. Therefore, one of the key RFP criteria in the Assessment Program should be the measurement of these skills.

Following are some high-level principles to consider. Many of these points have been made by expert panelists at the Department’s public meetings (notably the presentations by Dr. Linda Darling Hammond and Dr. Jim Dueck, but also many others). Please also accept as a background resource the accompanying document authored by the Partnership for 21st Century Skills, “State Implementation Guide for Assessment: a 21st Century Skills State Implementation Guide” – (available at <http://tinyurl.com/yfj4tvj>).

High Level Comments:

1. The assessment of 21st century skills should be a key criterion of the RFP.

It is imperative that the college and career readiness skills—as we expect them to be articulated in the Common Core State Standards, e.g., critical thinking, problem solving, communication, collaboration in English Language Arts and Mathematics—be a key criterion for funding in the RFP. In other words, while it is important for the Assessment Program to have as criteria the need to measure core academic subject mastery, it is *equally important* to also require new assessments to measure the skills required of 21st century citizens.

2. Develop assessments that measure core academic subject mastery and higher-order thinking skills

Assessment systems must measure core academic subject knowledge and understanding among all students. Students who can think critically and communicate effectively *must build on a base of core academic subject knowledge*.

In addition to core subject mastery, assessment systems should also produce meaningful data on whether students can demonstrate the ability to:

- Think critically
- Solve problems
- Communicate
- Collaborate
- Analyze information
- Use technology
- Innovate
- Think creatively
- Be globally competent
- Be financially literate

3. Emphasize performance-based assessments

Assessments should include evidence of actual student work on complex, authentic tasks that demonstrate higher-order thinking skills within a core subject. Rich tasks and open-item responses are necessary in this type of assessment system.

4. Emphasize curriculum-embedded assessments



It is important to develop curriculum-embedded assessments that produce meaningful information about a student's ability to think critically, problem-solve, communicate and collaborate. These assessments should inform and strengthen classroom practice, because they support inquiry-based approaches to student learning that, in turn, enable core academic subject mastery and higher-order thinking skills among students.

5. Prioritize the inclusion of practitioners

For curriculum-embedded, performance-based assessments to be effective measures of student outcomes, practitioners must be considered as central to the process. Practitioners should be involved in the development and implementation (scoring) of assessments. Practitioners should also be provided with ample opportunities for professional learning around the use of such assessments.

6. Develop policy frameworks that support the new assessments

We realize that this Assessment Program focuses solely on the development of new assessments and is not designed to attend to current accountability requirements in ESEA. However, we feel it is important to underscore an obvious point: assessments must operate in an aligned system that includes standards, professional development, instructional practice, learning environments. National and state policies must support each of these in a systemic fashion. The Partnership for 21st Century Skills currently works with a network of 14 partner states that are committed to integrating 21st century skills statewide; each of these state initiatives has the potential to be strengthened by effective national policies that promote career and college ready teaching and learning systems.



Assessment: A 21st Century Skills Implementation Guide



Produced by



**PARTNERSHIP FOR
21ST CENTURY SKILLS**

To succeed in college, career and life in the 21st century, students must be supported in mastering both content and skills. This Implementation Guide presents state leaders, policymakers and/or district and school leaders with assessment tactics and examples to assist in statewide 21st century skills initiatives. The Partnership for 21st Century Skills has issued five brief, user-friendly guides, one for each of the P21 support systems:



1. Standards
2. Assessment
3. Professional Development
4. Curriculum & Instruction
5. Learning Environments

It is worth noting that these support systems are not merely ends, but means to a greater goal—to help children develop the cognitive, academic, emotional and physical competencies they need to succeed in 21st century life.

The Partnership recognizes that taking an aligned, comprehensive approach across all five support systems is a significant challenge for all educators. The Implementation Guides have been developed to help support this difficult work. While not every recommendation and example will apply to every state, we hope the resources will help jumpstart efforts to produce more capable, successful 21st century students and citizens.

All 21st century skills initiatives must focus on:

1: Core Academic Subject Mastery

It is important to note that no 21st century skills implementation can be successful without developing core academic subject knowledge and understanding among all students. Students who can think critically and communicate effectively *must build on a base of core academic subject knowledge*. For this reason, core academic subjects are a bedrock component of the P21 Framework for 21st Century Learning. All 21st century skills can and should be taught in the context of core academic subjects.

2: 21st Century Skills Outcomes

In addition to core subject mastery, the Partnership asks every state, district and school the following question: are schools helping students become...

- Critical thinkers?
- Problem solvers?
- Good communicators?
- Good collaborators?
- Information and technology literate?
- Flexible and adaptable?
- Innovative and creative?
- Globally competent?
- Financially literate?

To learn more about the Partnership's state initiatives, the Framework or the Implementation Guides, please visit www.21stcenturyskills.org.

Rationale

Our nation faces serious questions in regards to our educational system. The purpose of this document is to provide you with perspective on the key issues to consider—as a policymaker, as state leader, as a district or school administrator—to ensure that you are planning for the future and building strategies that will solidify the success of our students, not only in school and work, but in life.

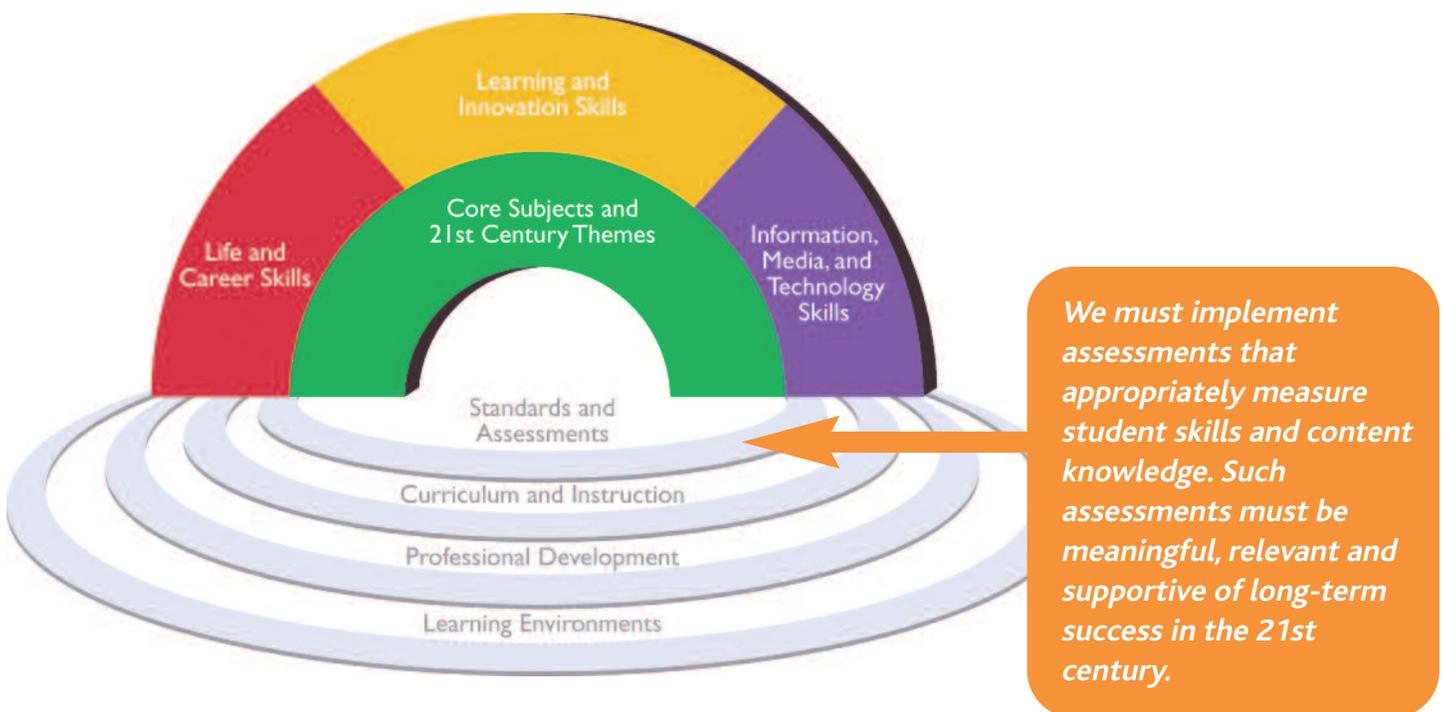
Most K-12 assessments in widespread use today—whether of 21st century skills and content or of traditional core subject areas—measure knowledge of discrete facts, not the ability to apply knowledge in complex situations. High stakes assessments alone do not generate evidence of the skill sets that the business and education communities believe will ensure success in the 21st century.

Vision

Twenty-first century accountability systems—including all forms of measurement—must assess the key dimensions of 21st century learning; they must measure those skills now prized in a complex global environment. There is growing consensus that our education systems should pursue measurement of student outcomes that are:

- Performance-based
- Embedded in curriculum
- Based on a common evidentiary model of cognition and learning

Each of these approaches inherently supports the measurement of 21st century skills.



Guiding Recommendations, Promising Directions

The following action steps can be taken to move states, districts and schools towards ensuring that our nation's students will be prepared for success in the 21st century.

Guiding Recommendations	Promising Directions
<p>#1: Build measurement of 21st century skills into large-scale summative assessments.</p> <p>Assessments should incorporate broader use of performance-based measures that focus on higher-order thinking and measure skills such as:</p> <ul style="list-style-type: none"> • Critical thinking • Problem solving • Communication skills • ICT literacy • Information literacy • Media literacy <p>The assessment development process should be collaborative, involving not only assessment experts, but practitioners, education leaders and, where appropriate, outside vendors who provide assessment-related services and products.</p>	<ul style="list-style-type: none"> • Migrate summative assessments from the rote memorization to higher levels of emphasis on higher-order skills like critical thinking. <i>Promising Practice: West Virginia is revamping its summative assessments to incorporate higher-order thinking skills.</i> • Explore how information technology can be incorporated into the country's "gold standard" for assessment. <i>Promising Practice: Problem Solving in Technology-Rich Environments (TRE) project.</i> The National Assessment of Educational Progress (NAEP) tested scientific inquiry skills, such as the ability to find information about a given topic, judge what information is relevant, plan and conduct experiments, monitor one's efforts, organize and interpret results, and communicate a coherent interpretation.¹ • Engage students in problem-solving tasks that align with core subject standards. <i>Promising Practice: Calipers Project (NSF).</i> With a focus on physical science standards related to forces and motion and life sciences standards related to populations and ecosystems, Calipers engages students in problem-solving tasks, such as determining the proper angle and speed to rescue an injured skier on an icy mountain.² • Develop standards-based, balanced approaches to assessment that allow students to demonstrate their knowledge through real-world tasks and building portfolios. <i>Promising Practice: The Ohio Performance Assessment Pilot Project</i> is designed to support the initial research, development and pilot testing of a standards-based, balanced assessment approach, allowing students to demonstrate their knowledge and skills through various real-world tasks and activities, the building of portfolios and other exercises. The pilot program uses multiple measures to evaluate students. By monitoring each school's program and receiving feedback from teachers and administrators, Ohio will begin to develop measures that offer a more comprehensive assessment of academic progress.³ • Develop evidentiary based assessments of 21st century skills that leverage performance data for continuously improved learning. <ul style="list-style-type: none"> • <i>Promising Practice: the College and Work Readiness Assessment (CWRA)</i> measures how students perform on constructed response tasks that require an integrated set of critical thinking, analytic reasoning, problem solving and written communication skills. • <i>Promising Practice: UCLA's IMMEX (Interactive Multi-Media EXercises)</i> http://www.immex.ucla.edu/

¹ Tucker, Bill. Beyond the Bubble. Rep. Feb. 2009. Education Sector Reports. http://www.educationsector.org/usr_doc/Beyond_the_Bubble.pdf.

² Ibid.

³ Ohio Performance Assessment Pilot Project." Ohio Department of Education. 30 Dec. 2008. 19 Feb. 2009 <http://www.ode.state.oh.us>.

Guiding Recommendations	Promising Directions
<p>#2: Globally benchmark summative assessments. We must ensure that U.S. students are being measured for their mastery of 21st century skills in ways that allow comparisons with students from other countries. To compete in a global economy, our students must demonstrate excellence on a global scale, not just a local or national scale.</p>	<p>Although they are not fully inclusive of 21st century skills in all cases, PISA and TIMSS are the best examples of this as of the publication of this document.</p> <ul style="list-style-type: none"> • Program for International Student Assessment (PISA) assesses high school students ICT literacy through establishing current skill and testing through various activities. Performance is assessed based not only on the ability to complete tasks, but also the manner in which tasks are completed. http://www.pisa.oecd.org • Trends in International Mathematics and Science Study (TIMSS) provides reliable and timely data on the mathematics and science achievement of U.S. 4th- and 8th-grade students compared to that of students in other countries. http://nces.ed.gov/timss/
<p>#3: Build 21st century skills into formative assessment strategies. States and districts should provide teachers with rubrics and checklists—along with the necessary professional development—to assess student mastery of 21st century skills in ways that impact, inform and improve learning in real time.</p>	<ul style="list-style-type: none"> • Use rubrics to evaluate 21st century skills. <ul style="list-style-type: none"> • <i>Promising Practice: Catalina Foothills School District</i> in Arizona has a series of rubrics used to assess students in real time. Rubrics evaluate 21st century skills such as critical thinking, productivity, and self-direction. • <i>Promising Practice: Lawrence Township</i> of Indiana currently uses rubrics to evaluate interactive communication and self-direction. • <i>Promising Practice: New Technology High School</i> has implemented rubrics for evaluating peer collaboration and teamwork, work ethic and written communication. • Develop innovative performance-based measurements. <i>Promising Practice: The North Carolina Business Committee for Education</i> and the Center for 21st Century Skills are currently entering the second year of work with the N.C. Science, Mathematics and Technology Center and Dr. John Bransford of the University of Washington to develop and pilot a multimedia online interactive scenario-based biology assessment.
<p>#4: Create an aligned accountability system; all assessment strategies should align with 21st century skills standards, professional development and curriculum and instruction. The goal here is to create an <i>aligned system</i> that enhances student learning and satisfies accountability requirements; for example, combining large-scale and classroom assessments using curriculum embedded performance tasks allows educators at every level to understand how students are progressing and <i>why</i>, and to use this information to enhance student learning in real time.⁴ Assessment strategies that measure 21st century skills must be developed in concert with standards, curriculum, instruction and professional development approaches.</p>	<p>Develop valid, reliable assessments aligned to 21st century skills whose results can be used to inform instruction and ensure accountability. <i>Promising Practice: West Virginia</i> is developing a new assessment program to create valid and reliable assessments that 1) are aligned to the 21st century skill descriptors and state content standards and objectives, 2) inform instruction, 3) promote school improvement and 4) produce results that can be used to calculate school, county and state accountability.</p>

⁴ Darling-Hammond, Linda. *Powerful Learning: What We Know About Teaching for Understanding*. San Francisco: John Wiley & Sons, Inc., 2008. pps 210-2-11.

Guiding Recommendations	Promising Directions
<p>#5: Consider ICT literacy assessment as a starting point. ICT literacy assessment, both formative and summative, provides an effective starting point for many states due to the fact that commercial testing products are already available.</p>	<p>Assess student abilities to navigate, critically evaluate and make sense of information available through digital technology. <i>Promising Practices:</i></p> <ul style="list-style-type: none"> • <i>ETS iSkills Assessment</i> http://www.ets.org/ictliteracy/ • <i>[U.K. specific:] Key Stage 3 ICT Literacy Assessment, Great Britain</i> • <i>Learning.com's TechLiteracy Assessment</i> • <i>PISA ICT Literacy Assessment</i>
<p>#6: Encourage and fund research and development around 21st century skills assessment. State departments of education, state universities, colleges of education and like institutions should focus efforts on a rigorous agenda to work on and have major core competence in assessment of 21st century skills. They should strive to build fundamental centers of excellence around the assessment of 21st century skills, including new item types and uses of technology.</p>	<p><i>Promising Practice: Assessment and Teaching of 21st Century Skills</i> is an international, collaborative effort sponsored by Cisco, Intel and Microsoft intended to provide: clear, operational definitions of 21st Century skills, solutions to technical psychometric problems that confront those seeking to develop tests of these skills, strategies for delivering assessments using ICT, and classroom-based strategies for helping students develop the skills. http://www.atc21s.org/</p> <p><i>Promising Practice: The Educational Testing Service's Cognitively Based Assessment of, for and as Learning (CBAL)</i> is a technology-based research project in Portland, Maine. In schools with one-to-one laptop programs, the project focuses on the research and development of a cognitive model for how students read and develop reading skills.</p>
<p>#7: Create open repositories for assessment items and rubrics that help measure 21st century skills. State departments of education should become recognized as centers of excellence for measuring 21st century skills, creating open repositories for sharing assessment items, rubrics and promising practices.</p>	<ul style="list-style-type: none"> • Align skill assessment rubrics with business expectations for workplace readiness. <i>Promising Practice: New Jersey</i> is incorporating 21st Century Knowledge and Skills into the protocol established by the NJ Performance Assessment Alliance Project. • Collect and review existing assessment tools to formulate state best practices. <ul style="list-style-type: none"> • <i>Promising Practice: Massachusetts</i> is reviewing rubrics for evaluating high school graduation projects from several other states with the goal of developing their own rubrics based on state standards and frameworks. These will be shared with schools in order to ensure that even these first-stage assessments meet high standards. • <i>Promising Practice:</i> in 2004 the ECS National Center for Learning and Citizenship started collecting, judging and coding existing assessment instruments for civic education. The Campaign for the Civic Mission of Schools and the Center for Civic Education have contributed resources to support the creation of this draft database. http://www.ecs.org/Qna/splash_new.asp • Develop high-quality rubrics for self-direction, critical thinking, information literacy and other skill areas. <ul style="list-style-type: none"> • <i>Promising Practice: Catalina Foothills School District (Tucson, AZ)</i> and <i>Lawrence Township ISD (IN)</i> have developed a number of high-quality rubrics focused on specific 21st century skill areas. These can be located on Route 21 (http://www.21stcenturyskills.org/route21/).

Resources

In addition to the listings above, The Partnership for 21st Century Skills has compiled the following list of resources to provide you with background knowledge, models and promising practices in the various areas of assessment, as well as a list of key expert contacts.

Education Sector

Bill Tucker and Elena Silva

<http://www.educationsector.org>

Microsoft/Cisco/Intel Assessment of 21st Century Skills Project

Bob Kozma

<http://www.atc21s.org>

The New Technology Foundation

James Popham, Director of Strategic Planning <http://www.newtechfoundation.org>

Bob Pearlman, Strategy Consultant for Education Reform <http://www.bobpearlman.org/>

Route 21: P21's online database that includes district-created rubrics for assessing 21st century skills.

<http://www.21stcenturyskills.org/route21/>

The School Redesign Network

Ray Pecheone, Director

<http://www.srnleads.org/>

The University of Washington

John Bransford, Professor of Education <http://education.washington.edu>

A complete updated list of available references, including reports, state initiatives, white papers and more are available at www.21stcenturyskills.org.

Free White Paper on 21st Century Skills Assessment

Download "21st Century Skills Assessment" from the Partnership for 21st Century Skills website at http://www.21stcenturyskills.org/documents/21st_century_skills_assessment.pdf.

About the Partnership for 21st Century Skills

The Partnership for 21st Century Skills has emerged as the leading advocacy organization focused on infusing 21st century skills into education. The organization brings together the business community, education leaders and policymakers to define a powerful vision for 21st century education to ensure every child's success as citizens and workers in the 21st century. The Partnership encourages schools, districts and states to advocate for the infusion of 21st century skills into education and provides tools and resources to help facilitate and drive change.

To learn more about 21st century learning and state actions to date, visit www.21stcenturyskills.org.



Student Growth Data for Productivity Indicator Systems

Edward Haertel
School of Education
Stanford University

Exploratory Seminar:
Next Generation K-12 Assessment Systems
Educational Testing Service
Princeton, New Jersey
December 7, 2009

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author. |



Steps to consider

- defining growth for an individual
- defining and comparing growth for groups
- group growth measures in productivity indicator systems

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author. 2

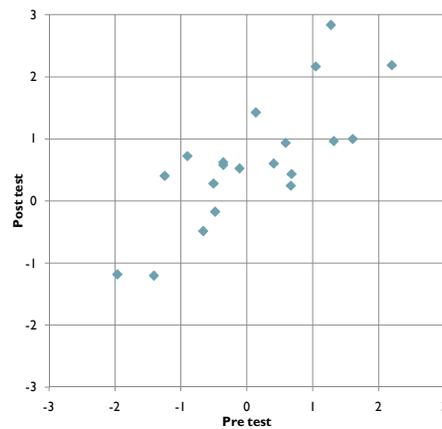
Defining growth for an individual

- Abstract problem:
 - define a **mapping from a vector** (representing achievement, maybe additional attributes, at two or more time points) **onto a single dimension**
 - simplest case: $(\text{pre}, \text{post}) \rightarrow (\text{post} - \text{pre})$
 - observations may represent > 2 time points
 - observations at one time point may be vector-valued
 - observations may include more than test scores

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

3

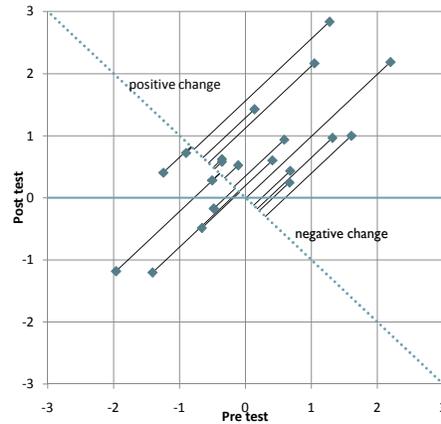
Gain score is mapping from (pre, post) to (post - pre)



Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

4

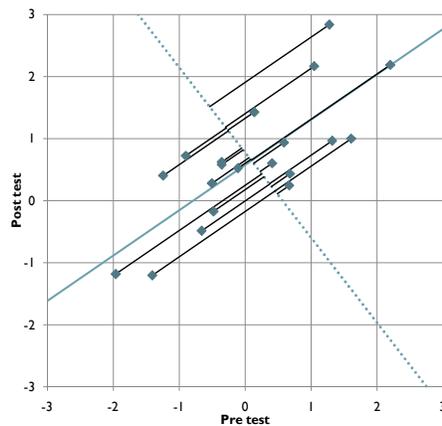
Gain score is mapping from (pre, post) to (post - pre)



Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

5

Residual from regressing post on pre is another option



Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

6

Defining growth for an individual

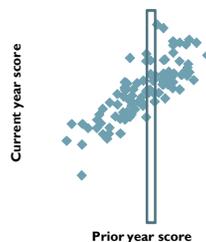
- Substantive problem: create a suitable **index of year-n performance relative to expectation**
 - What information is individual growth score supposed to capture?
 - Should year-n expectation be the same for all students? *Then we don't need growth measures*
 - Should year-n expectation be the same for all students with the same prior-year score? *Then we could use Student Growth Percentiles (SGPs)*
 - Or is controlling for prior-year score not sufficient?

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

7

Student Growth Percentiles

- Construction
 - Each student's SGP score is the percentile rank of that student's current-year score within the distribution for students with the same prior-year score



Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

8

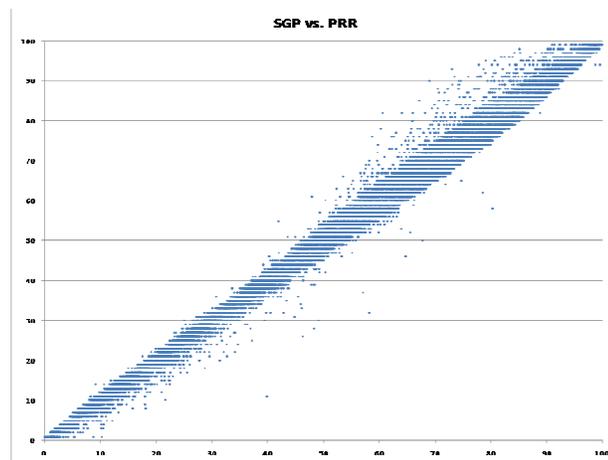
Student Growth Percentiles

- Interpretation
 - How much has this student grown relative to others who began at the “same” (prior-year) starting point?
- Advantages
 - Invariant under monotone transformations of score scale
 - Directs attention to distribution of outcomes, versus point estimate

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

9

Is anything really new here?



Thanks to Andrew Ho and Katherine Furgol for this graphic

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

10

Controlling for prior-year score is not sufficient

- First problem—Measurement Error: prior-year achievement is imperfectly measured
- Second problem—Omitted variables: models with additional variables predict different prior-year true scores as a function of
 - additional test scores
 - demographic / out-of-school factors
 - much to be said here

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

11

Controlling for prior-year score is not sufficient

- Third problem—Different trajectories: students with identical prior-year true scores may have different expected growth depending on
 - individual aptitudes
 - out-of-school supports for learning
 - prior instructional histories

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

12

Better measurement scales are not enough

- Consider “4 inches growth in height”
 - interpretation and evaluation depend on child’s age as well as prior-year height
 - measurement without context is almost useless, despite “ideal” measurement scale

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

13

Performance relative to expectation

- What do we wish to adjust away?
What do we wish to capture?
- Fundamental problem
disentangle determinants of year-n achievement for which [teachers / schools / “the system”] should be held accountable from those beyond the [teacher’s / school’s / system’s] control
 - no bright line
 - what should be in or out will vary with intended use / interpretation

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

14

In or out?

- District leadership
- School norms, academic press
- Quality of school instructional staff
- Early childhood history; medical history
- Quality of schooling in prior years
- Parent involvement
- Assignment of pupils (to schools, to classes)
- Peer culture
- Students' school attendance records

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

15

Defining growth for a group

- Easy to characterize the time of an individual runner
- harder to characterize a group of runners
- harder still to characterize a group's change over time or to compare groups
- default approach is
 - first to map individuals' vectors to scalars, then to summarize those scalars
 - not the only choice, but enough for today

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

16

Some obvious options

- mean growth (unweighted average)
- proportion above cut score
- vector of proportions above successive cut scores
- weighted average
 - e.g., to credit “gap closing”

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

17

Individual growth definition and method of aggregation interact

- We could define “performance relative to expectation” in a manner that sets higher growth expectations for lower-performing students, then take unweighted average

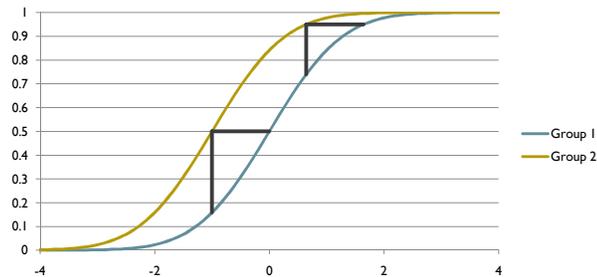
Or

- We could define “performance relative to expectation” to set equal growth expectations for all students, then take weighted average

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

18

Comparing groups



Holland (2002) describes “vertical” and “horizontal” gap measures. Vertical (e.g., % proficient) are widely used; horizontal would be better. But his main message is that no one number can be relied upon to do a good job of summarizing differences between distributions

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

19

Pointing toward a solution

- Recognize this as a problem of policy capturing
 - it is not solely a technical problem
 - don't let policy makers get away with insisting on technical answers to values questions
 - role of technical expert is to clarify choices and associated trade offs and to facilitate decisions
 - best answer will vary from case to case
 - will depend on scope of application, data available, level of aggregation, stakes, ...

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

20

Pointing toward a solution

- Define individual growth scores relative to expectation so that (as best possible) equal numerical values represent equivalent perceived success / goodness
- Take an average
 - unweighted or weighted average of individual growth scores
- Goal is scalar index of group performance to support intended uses and interpretations

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

21

Footnotes

- Keep it simple
- Use continuous measures
 - Avoid cut points and categories (e.g., “basic”)
- Use a single average
 - Avoid multiple conjunctive criteria
- Use weighting to create incentives to allocate instructional resources differently
 - e.g., extra weight for members of underserved groups

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

22

Footnotes

- Separate creating growth scores from creating descriptive / evaluative categories
 - System may require “ambitious but attainable” targets, but determining those targets should be a separate step from defining the (continuous) group performance scale
 - Scale may be kept fixed even as targets change over time

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

23

Group growth measures as productivity indicators

- Two broad categories of uses
 - Evaluating programs
 - multiple districts, schools, perhaps classrooms each assigned to one of two or more conditions; conditions compared to one another
 - Evaluating units
 - large number of individual districts, schools, perhaps classrooms compared to one another
- For each category of uses, group outcome summary alone has limited utility

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

24

Simplest case for evaluation of programs

- Valid (if impoverished) design would just compare outcomes across treatments
 - t-test or one-way ANOVA
 - regards treatment as “black box”
 - if individual growth is properly defined (as argued here), then no individual-level covariates should be required
 - group-level covariates may be helpful; may be required with nonrandom assignment
 - **caution: achievement is not unidimensional**

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

25

Simplest cases for evaluation of units

- Arms-length, “good job” / “bad job” accountability plan
 - use rewards and sanctions to create incentives for improvement
 - publicize relative success of different schools and promote school choice to mobilize market pressures for improvement
 - note: does not increase understanding of “what works”

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

26

Simplest cases for evaluation of units

- Use group outcomes as one factor in resource allocations
 - problem of paradoxical incentives (rewarding failure?)
 - no theory of action for wise use of additional resources (ask for an improvement plan?)
- Study outliers
 - logic of old “effective schools” research
 - no statistical warrant for retrospective inferences

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

27

Group growth measures as part of indicator systems

- “Productivity indicator systems” seem to call for much more
 - Constructing individual growth measure required analysis and measurement of achievement influences **outside** the school’s control
 - Analysis of achievement influences **within** the school’s control is even more important

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

28

Group growth measures as part of indicator systems

- Stronger program evaluation models would capture (for example) the relevant features of classroom instruction
- instructional history trajectory \neq achievement trajectory
 - students' instructional histories will affect their readiness to profit from a specific instructional intervention at a specific point in time
 - e.g., obviously, math preparation for physics course

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

29

Group growth measures as part of indicator systems

- Indicator systems
 - Systematic, with logical connections
 - Designed for a specific purpose
 - may be narrow or broad
 - need not attempt to be comprehensive
 - Quantifies and **connects** variables relevant to purpose
 - diagnosis
 - monitoring
 - evaluation
 - explanation

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

30

Group growth measures as part of indicator systems

- Design dictated by theory of action
 - may cover multiple levels of system
 - focus may be on one or two levels
 - single subject or multiple subject areas
 - relevant variables will differ by school type
 - EL instructional practices
 - school campus safety
 - elementary / middle / high schools

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

31

Left for another day...

- Utilization of information from productivity indicator systems
 - getting relevant information to the right people at the right time
 - requires attention to larger, technologically supported social system
 - understanding and using information
 - to recognize problems
 - to frame possible solutions
 - to reach better decisions

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

32

Conclusions

- The promise of growth measures is as useful building blocks for studying educational productivity
- Growth measures alone will offer little insight
- Individual growth measures, group summaries, and indicator systems built for one purpose may be suboptimal for other purposes

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

33

Thank you

34



Ellen Haley
President

20 Ryan Ranch Road
Monterey, CA 93940-5703
(831) 393-7757 Tel
(831) 393-7243 Fax

December 2, 2009

Office of Elementary and Secondary Education
U.S. Department of Education
400 Maryland Avenue, SW, Room 3E108
Washington DC, 20202

Attention: **Race to the Top Assessment Program Public Input**

CTB/McGraw-Hill is submitting these comments in response to the Notice published in the Federal Register on October 23, 2009 (74 Fed. Reg. 54795) to assist the Department of Education (DOE) in carrying out its proposed assessment initiative under the Race To The Top program for encouraging and supporting states to develop and implement comprehensive, balanced assessment systems.

We appreciate the opportunity to comment, and we offer our assistance and expertise to work with the DOE as it embarks on this critical task of transforming education in America. We support the Common Core State Standards Initiative to ensure every American student graduates high school with the skills and knowledge needed to thrive in the 21st century. With CTB's capability and experience, and based on what we hear during our work with states and districts, education scholars and professionals as well as researchers across the country, we submit these comments. We also are submitting our 21st Century Assessment whitepaper. As members of the Association of American Publishers and the Association of Test Publishers, we endorse their joint comments filed in response to the Notice.

Assessment Framework and High-Quality Assessment Design and Development

The development of effective assessment systems, which incorporate formative, interim, and summative assessments, is critical in transforming and improving education. The DOE is correctly focusing on looking at proposals from states for systems, not just isolated assessments. Assessment systems must take into consideration today's sophisticated and rapidly-changing technologies; evolving understanding of learning; the challenges of delivering multiple forms of assessments to meet multiple purposes; ensuring consistency with recognized technical, psychometric, and operational standards and guidelines; meet the needs of students with disabilities and English language learners; and report results so that they are actionable for teachers, administrators, families, and especially students. All of this must be done in a fair, timely, and cost-effective way.

It is clear that assessment systems must be developed following the highest professional and technical rigor, using the best available technologies and supported by defensible research and statistics. This is a challenging responsibility and one we do not take lightly. CTB/McGraw-Hill applies our extensive experience and expertise to ensure test quality. Whether developing items for a customized state assessment or for a consortium of states assessing common content standards, the process is complex and involves multiple stages and must be done by testing experts.

The DOE should ensure that any proposal for assessment systems, and the assessments therein, is developed with strong evidence of the professionalism and technical experience of the developers. Furthermore, there should be evidence that the proposal, at a minimum, meets the current requirements for demonstrating technical quality. The proposal should clearly state how the program will use the assessment information; how it would determine whether that information assisted in improving instruction and student learning; and outline the plans for making corrections to the assessment system based on the results. Information alone is necessary but not sufficient. An assessment system also should identify

how the information is used for action – in professional development, to target interventions, and to provide support for students, educators, and parents. An evaluation system is essential to determine if the assessment system worked as intended to provide for continuous feedback that leads to improved student learning.

We also would like to commend the DOE for inclusion of norm-referenced tests as a component in an optimally balanced assessment system. By design, CTB/McGraw-Hill's TerraNova provides vertical alignment and vertical learning progressions, which in our experience are foundational for evaluating student growth. Because it aligns with state standards and the frameworks of the National Assessment of Educational Progress (NAEP), TerraNova generates normative information that can provide standards/criterion-based and competency-based inferences about student progress. It would provide a solid foundation for any assessment system to meet the general and required characteristics defined in the Notice.

One technical recommendation that we offer is that both "development" and "acquisition" by states and LEAs of instructional materials, assessments, and professional development should be allowable activities. While both "acquisition" and "development" are used together in the Notice, we note that in the last bullet on p. 54798, column 3, only "development" is included. It should read "Acquisition or development of formative or interim assessments that align with State summative assessments as part of a comprehensive assessment system."

We make these comments based on CTB/McGraw-Hill's extensive experience in "operationalizing" and implementing stringently developed and complex assessment systems that are valid, reliable, and flexible. Our innovative assessment solutions are developed in close collaboration with educators, scholars, and technology experts to meet the highest psychometric standards. We have several client studies demonstrating improved student performance due – in part – to the data provided by our assessment solutions and the dedication of instructional leaders to utilize the data in developing their instructional approaches. But as our long history in the assessment industry has shown us, simply developing effective assessments is only part of the overall solution. Any successful assessment system includes continued maintenance and enhancement in content and technology and, of course, professional development.

Assessment Technology and Infrastructure

Assessments in the reformed educational system that the DOE is seeking must perform multiple tasks through the use of multiple assessment types and formats, item types, and modes of administration.

Digital and other technologies facilitate and are a requirement of any modern assessment system. While technology is a necessary component of a balanced assessment system, states are at varying points along the continuum toward the optimum use of technology, and most state systems are not sufficient. Affording flexibility in shaping the Department's proposal should recognize this technology dilemma. However, we strongly endorse an increased focus on the application of technology. Although technology affords a variety of benefits, it cannot, in and of itself, overcome poor assessment development and design. It does, however, provide benefits by being able to measure student learning across the complete spectrum of standards with shorter instruments and facilitates the use of multiple measures. In addition, technology can enable the use of performance-based or constructed-response item types with rigorous, valid, and fair scoring rubrics. It provides an opportunity to incorporate a broader range of items including "scaffolded" items, open-ended responses and performance simulations. It allows for the measurement of 21st century content and skills and is itself the application of 21st century learning.

In addition, the use of technology offers increased access to the assessment process for more students in several ways. For example, a computer-based system allows for a broader range of accommodations for students with disabilities and English language learners that could more closely resemble the instruction they receive without compromising the overall construct. Items could be personalized and "scaffolded" to ensure that the assessment achieves alignment with the appropriate cognitive levels while increasing access by providing a ladder between lower and higher levels of complexity.

With quality reporting at the student, class, school, district, and state levels, administrative leaders can make more informed decisions about the classroom curriculum and professional development needs in a timeframe that makes a difference for differentiated instruction. With the multiple measures, reporting progress by standards acquires increased reliability and validity through the increase in data points for the overall assessment. The use of technology also offers the opportunity for more immediacy in feedback and rapid turnaround for accountability purposes – technology frees test administration and the return of student score results from other time constraints.

Improved utilization of technology would be facilitated by greater commonality of definitions and other technology-related standards. The lack of such consistency can be a limiting factor in the implementation of innovative assessment systems. Any new assessment program must consider the implications of how any limitations in the consistency of standards can be addressed.

The DOE should seek information on the technology infrastructure and resources that will be available. In particular, a baseline of technology resources and availability, and a timeline for any implementation of technology would be useful information for both states' planning and DOE evaluation. Infrastructure capabilities have major implications for operations, but also for learning opportunities for students. Additionally, the DOE may request assurances that applicants will equalize the access to technology and the technology infrastructure across all schools and districts to ensure the program efficacy and equity.

Professional Development

The Notice focused more on teacher evaluation and teacher scoring as design elements and LEA activities to support the transition to a new system. These again are necessary but not sufficient for an effective assessment system. The need for better professional development in understanding and using the information from an assessment is one of the strongest and clearest points of consensus for improving any assessment system. It should be a system requirement that applicants identify how professional development would be infused throughout the assessment system and how assessment results will be integrated into professional development.

Further, it is clear that a primary purpose of an assessment system is as a professional support resource for teachers. For example, technology-based formative and interim assessments are powerful tools for improving education. They provide educators with student and classroom performance data that can be used to identify learning gaps and consequently to adjust a teacher's approach, content, and pace of instruction. Because the assessment data is available immediately, instructional intervention can be made quickly to improve student comprehension and academic progress. CTB/McGraw-Hill strongly supports the use of not only high-quality, measurement-based summative assessments, but also similarly designed formative and interim assessments for teacher decision-making. With this in mind, the DOE may be served best by ensuring that any proposals address teacher capacity and professional development in such a way as to empower teachers to utilize the assessment results "nimble and aptly" in their classrooms, transforming data into actionable information. To avoid the dilemma of "data rich – information poor," the DOE needs to support the provision of resources and interventions for educators to allow differentiated instruction that furthers academic progress.

Teacher scoring and the use of technology for scoring both have benefits. Through technology, assessments can be scored efficiently and effectively. By combining technology with some teacher scoring, professional development and teacher engagement benefits can also be realized. The requirements for a final proposal for a balanced assessment system should recognize both of these objectives and provide sufficient flexibility such that teacher scoring be included as a "desired" but not required characteristic, enabling broad flexibility for implementation.

Operational Considerations – Opportunity to Learn

In structuring the assessment program that would be supported by Race To The Top funds, the DOE should carefully address the fundamental question of "how is this going to work?" The answers will range from the articulation of the purposes, to addressing the structure and governance of the consortia, as well as the sequencing of content standards, professional development, supporting instructional materials,

and performance standards. Standards and assessments do not make a curriculum. Teachers need an opportunity to teach and students need an opportunity to learn before new assessments should be implemented and consequences attached – whether for state or federal accountability or for teachers making judgments about individual children.

We also underscore that it is essential that **all** students have access to the highest quality instructional materials, as a matter of equity and fairness to ensure that all students have an equal opportunity to learn. The Notice addresses this in part through the list of types of LEA activities.

As many expert panelists have advised the DOE, there are many operational considerations that must be articulated in seeking proposals. Our experience is that the greater the clarity of purpose, the more that the procedures and timelines are detailed, the clearer the lines of responsibilities and authority, and the better the articulation of how the parts are intended to work together in a continuous improvement loop, the greater the prospects for success.

CTB/McGraw-Hill looks forward to contributing our expertise and services to The Race To The Top Initiatives.

Conclusion

CTB/McGraw-Hill has an 82-year record of educational assessment innovation and excellence starting with being the first to introduce objective, standardized, achievement tests. Our innovation continues today with technologies that include online formative assessment, automated test assembly, technology-embedded assessment solutions, and artificial intelligence. CTB/McGraw-Hill continues this leadership today in collaboration with our partners in education serving more than 17 million students in all 50 states and in 53 countries.

The Race To The Top program offers a unique opportunity to lay the foundation and support the transformation of education. As pioneers and experts in innovative assessment design, development and implementation, and as a leading and nationally recognized assessment company working with states and districts in the United States, we have offered our comments to The Race To The Top initiative. This is an opportunity to foster sound, balanced assessment systems aligned to common content standards. We support this effort, as it speaks to the very foundation and expertise of CTB/McGraw-Hill. We therefore look forward to working with the Department of Education and the states in developing the solutions and to participating in the implementation of the Race To The Top assessment program, which we hope will define the future of assessment in the United States.

Sincerely,

A handwritten signature in blue ink that reads 'Ellen Haley'.

Ellen Haley

Assessment in the 21st Century: *Preparing Students for the Global Workplace*

A Vision for Innovative Assessment and Reporting in the 21st Century

Education today is undergoing profound change. This transformation, driven by education technology, new forms of educational resources, and accountability measures, is creating entirely new ways of teaching and learning. In turn, these changes are driving focused, efficient delivery of educational resources, delivering true data-driven instruction targeting individual student success.

Assessment: A Critical Educational Tool

A wholesale shift is taking place in the skills required to participate and succeed in the “knowledge economy” of the 21st Century. This has tremendous implications for educators, as well as their students, worldwide. Educational assessment and reporting is evolving correspondingly, and is turning out to be very different from the static testing model of yesteryear.

Test scores today aren’t simply recorded and used as grading or advancement tools; they are no longer viewed as outputs; they are inputs. Deconstructed scores provide dynamic, revealing information about a student’s study habits, abilities, potential for growth, and subject mastery, as well as guideposts for effective classroom instruction and professional development of teachers. The same assessments also provide systems information: Are the standards correctly calibrated? Is instruction succeeding in upholding the performance standards? Are students well prepared? Is the funding well spent? As such, assessment generates information upon which policy decisions can be made. Educational assessment is a foundation activity in every school, every district, and every state—a vital component towards raising standards and achieving educational excellence.

The value of assessment lies in the information it provides:

- Detailed, personalized performance data to drive student progress
- Student results that correlate with, and guide, instruction
- Summative combined with formative assessment that creates a continuous, dynamic learning environment – with real-time feedback that bridges learning gaps and informs instruction.

Foundations of a Balanced 21st Century Assessment Program

The principles of a balanced 21st Century assessment program are straightforward – they should:

- Measure student achievement and mastery of skills

Educational assessment is a foundation activity in every school, every district, and every state – a vital component towards raising standards and achieving educational excellence.

- Provide information to enhance instruction
- Enable evaluations of the effectiveness of instruction
- Provide data and tools to improve student progress
- Engage parents in student learning
- Monitor educational systems for public accountability
- Help prepare students to effectively compete in the global economy

Technology: The Key to 21st Century Assessment

The future of education lies in digital technology and the development of more effective and efficient methods of teaching, assessment, and reporting. This will be particularly important as agencies begin to incorporate 21st Century skills, such as critical thinking, complex problem solving, visual literacy, and real world literacy. States are increasingly requesting online solutions to bring the benefits of technology into classrooms, to provide prompt feedback, reduce the turnaround time for student reports, save instruction time, and ultimately save costs. Much of the technology already exists to make an efficient, universal, and yet customized digital-based system a reality. Importantly, technology also allows learning environments to be extended outside the classroom, and can motivate students to want to

learn. Tech-savvy students want on-demand information, and will not be satisfied with static materials. Today's "digital natives" are at ease using the latest hardware and software, and are eager to incorporate their own learning technology, under their control, into their lives.

The impact of education technology should not be underestimated. It is the key component that will motivate students to take control of their own learning and keep them engaged in classroom activities. In addition, the right use of technology will ultimately lower costs of assessment construction, administration, scoring, and reporting, while providing valuable resources for intervention, instruction, professional development for teachers, and parent engagement in student learning.

The impact of education technology should not be underestimated. It is a key component towards motivating students to take control of their own learning and keeping them engaged in classroom activities. In addition,

technology will ultimately lower costs of assessment construction, administration, and reporting, while providing valuable resources for intervention, instruction, professional development for teachers, and parental engagement.

As technology becomes ever more readily available in classrooms, several factors will ensure greater access:

- Technology should be reasonably priced and reliable.
- Assessment systems must work with the level of technology in the classroom.
- Funding for IT support and operational costs must be included – this is as important as computer hardware and online access.

The impact of education technology should not be underestimated. It is a key component towards motivating students to take control of their own learning and keeping them engaged in classroom activities.

By implementing a comprehensive computer and online system of assessment and reporting, educators can:

- Identify and eliminate gaps in individual student learning as they are developing
- Allow students to carry their assessment records with them from district to district
- Implement much faster turnaround of student results
- Be accurate, timely, and nuanced enough to account for the numerous variables that can have an impact on how a student learns and retains knowledge

The effective use of education technology in classrooms will take the combined and coordinated efforts of everyone who has a stake in the education infrastructure, including teachers, parents, and students; business leaders in both the education publishing and technology industries; and legislators at all levels of government.

Classrooms need access to robust, wireless, high-bandwidth Internet access. This will occur as the price of storage and bandwidth continues to decrease. In addition, a smart software agent can guide knowledge navigation and create a more intuitive and highly customized learning experience. Most critical, the user interface that students will employ must be identified and developed.

Will it be a laptop, a PDA-like device, or something completely different? One possibility is a hand-held, notebook-sized device designed specifically for education with wireless Internet connectivity, intelligent software working in the background, the ability to display text and images and play music, and which can be interfaced via voice, stylus, or keyboard.

The Use of Multiple Data Sources to Inform Instruction

No single assessment can determine whether or not all educational goals are being met. More than one type of assessment is necessary to tell educators what students know and can do. Similarly, no one assessment provides complete information regarding one student's progress. The consideration of multiple data sources in educational assessment is the keystone to valid, fair, and reliable information about student achievement. Assessments provide a partial insight into, and reflection of, a student's abilities and progress. A test score is a proxy for gauging academic knowledge, and provides an estimate of measure for a complex underlying construct. Ultimately, the scores must be interpreted and then used to support achievement.

Student response devices, or “clickers,” are a recent technological development for administering assessments. Using clickers is simple and easy: each student in a class is given a clicker (a handheld device similar in size and layout to a television remote control) that they use to enter answers to test questions. Questions can either be on paper forms or displayed on a projection screen. The teacher has a small receiver that plugs directly into a computer's USB port. Test responses are instantly uploaded by a wireless connection, and software aggregates the results and uploads the data to the assessment program.

Acuity UnWired™ is the new clicker integration software for the popular Acuity® assessment. Prior to the formal release, Acuity UnWired was used in a series of pilot projects in five states. Sixty fourth graders and 75 sixth graders from several classrooms in Raleigh County, West Virginia, participated in a pilot project. Students used clickers to complete Acuity benchmark tests consisting of 30-40 multiple choice questions in math and reading language arts.

All of the teachers involved in the pilot project agreed that their students were more engaged and interested in tests when using the clickers. Teachers appreciated the immediate feedback and instant reports that they could generate, and viewed these features as decided advantages over traditional test administration. They observed that the most valuable part of the pilot program was knowing what they would need to reteach or reemphasize, as soon as the test was complete. All said they could and would use the results to make real-time adjustments to their instruction.

“The students were enthusiastic about using the clickers, and I knew immediately where certain students were struggling in their studies,” said Michelle Woods, a 4th grade teacher at Bradley Elementary School.

“It is so important to have real time data so that teachers and students can use it for learning purposes,” said Sandra Sheatsley, Principal of Bradley Elementary School.

The Use of Multiple Assessments to Guide Instruction

No single assessment can determine whether all educational goals are being met. More than one type of assessment is necessary to tell educators what students know and can do. Similarly, no single assessment provides complete information regarding one student's progress. The consideration of multiple data sources in educational assessment is fundamental to valid, fair, and reliable information about student progress and achievement. Assessments provide a partial insight into, and reflection of, a student's abilities and progress. A test score is a proxy for gauging academic knowledge, and provides an estimate of measure for a complex underlying construct. Ultimately, the scores must be interpreted and then used to support achievement.

In a comprehensive assessment system, all aspects of educational attainment should be evaluated, including the 21st Century themes of global awareness, systems thinking, ethics, visual literacy, and communication skills.

Components of 21st Century Assessment Program

Students demonstrate mastery of key standards through assessments that provide personalized diagnostic feedback to address their individual learning needs. Assessment programs must include multiple assessment measures so that educators have thorough, comprehensive data to base instructional decisions on. In a comprehensive assessment system, all aspects of educational attainment should be evaluated, including the 21st Century themes of global awareness, systems thinking, ethics, visual literacy, and communication skills. *Assessments must utilize available technology to enhance curricula and instruction and engage students in their learning.*

Key components of the ideal 21st Century assessment program should include:

Summative Assessment

These tests summarize what has been learned over time by:

- Measuring student achievement and including reports ranging from individual or aggregated performance on what students know on specific content standards and skills
- Showing academic progress over time
- Comparing student performance and growth over time and across jurisdictions
- Monitoring educational systems for public accountability
- Helping provide information to enhance instructional practices
- Enabling evaluations of the effectiveness of instructional practices

Standardized Achievement Tests

These assessments can be used alone or embedded into state summative assessments to give educators, students, and parents a more complete view of achievement. Embedding a standardized test into a state test has multiple benefits:

- Shows how students are performing on state standards, as well as how they are performing in comparison to national peers
- Provides a stable vertical measurement scale across grades which aligns with the growth model concept some states are adopting
- Saves states money by reducing test development costs

Formative Assessment

Formative assessment - encompassing interim, diagnostic, predictive, and benchmark assessment - provides teachers, students, and parents with ongoing, targeted information on academic strengths and weaknesses in order to improve student learning.

Formative assessment – encompassing interim, diagnostic, predictive, and benchmark assessment—provides teachers, students, and parents with ongoing targeted information regarding academic strengths and weaknesses in order to improve student learning.

Successful formative assessment:

- Provides targeted data to guide teaching and student learning, and identifies appropriate interventions on the student and/or classroom level
- Provides research-based predictions of student performance on state tests
- Aligns to current state content standards
- Guides curriculum development and teacher professional development

Case study: Technology-Based Assessment Improves Student Performance

The Park Hill School District in Kansas City, Missouri, with over 10,000 students in 15 schools, began using Acuity Assessment in 2006. The District has evaluated Acuity's effectiveness by tracking student performance on the Missouri Assessment Program (MAP) over time. Overall, the number of students passing the MAP increased by approximately 10 percent following the implementation of Acuity.

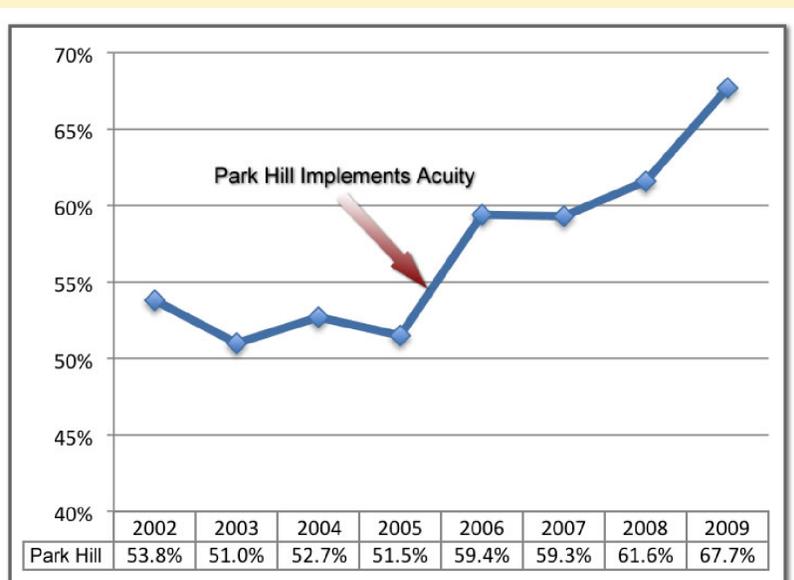
“Our state test scores have improved since we began using Acuity, and I am confident that Acuity has facilitated the improvement of our instruction,” said Jeff Klein, Ph.D., Executive Director of Research, Evaluation, and Assessment for Park Hill Schools. “Our focus has moved toward an emphasis on growth rather than just the end-of-year score. Teachers want to see how much a student improved over the course of a year, not just where they ended.”

While teachers had long used classroom assessments to gauge student learning, the District lacked the means to consistently measure student performance against the Missouri Grade-Level Expectations or to predict student performance on the MAP. In 2006, the District began using Acuity to help bridge that gap. Acuity allows classroom teachers to diagnose students' strengths and instructional needs, while predicting student performance on state assessments. Acuity integrates predictive and diagnostic assessments, reports, instructional resources, item banks, and item authoring - all aligned to state standards, and designed to improve student achievement.

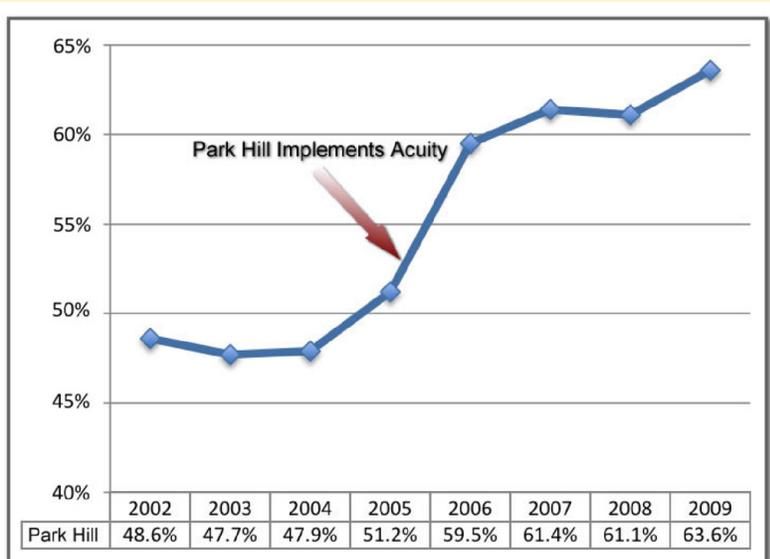
“We were tired of being surprised by MAP scores at the end of the year,” added Dr. Klein. “We would try to use MAP data to target instruction for incoming students in the Fall but, during the year, teachers didn’t have a sense of whether they were making a difference toward those end-of-year standards-based outcomes.”

Test scores improved markedly following the introduction of Acuity in 2006: the percentage of students passing the MAP test increased, in both communication arts and mathematics, by 10 percent or more (see the figures below, noting the jump in 2006). The district also reports a sharpened focus on Missouri Grade-Level Expectations and student learning.

Percentage of Students Proficient in MAP Communication Arts



Percentage of Students Proficient in MAP Mathematics



“Overall, Acuity has helped Park Hill take the next step into the world of standards-based education,” said Dr. Klein. “It has helped teachers move from a focus on teaching to a focus on learning. As a result of Acuity, our teachers are not only more knowledgeable about Missouri’s standards, but can also measure progress toward proficiency on these standards.”

21st Century Skills Assessment

“I’m calling on our nation’s governors and state education chiefs to develop standards and assessments that don’t simply measure whether students can fill in a bubble on a test, but whether they possess 21st century skills like problem-solving and critical thinking and entrepreneurship and creativity.”

President Barack Obama, March 10, 2009, from “Remarks by the President to the Hispanic Chamber of Commerce on a complete and competitive American education.”

Assessments must not only measure the conventional goals of core subjects, but should incorporate new skills necessary for today’s “knowledge economy”. Educators today are exploring the best ways to teach and assess these “21st Century skills”, that include:

- Critical Thinking and Problem Solving
- Information, Media, and Technology Skills
- Creativity and Innovation
- Communication and Collaboration

Case Study: Assessment of 21st Century Skills in West Virginia’s Global21 Program

The integration of 21st century skills into educational curricula is increasingly recognized as essential to the advancement of K-12 education. All students today need new competencies, knowledge, and expertise to master the multi-dimensional demands and abilities required of them in the 21st century.

The West Virginia Department of Education is among the first states in the nation to incorporate formally 21st century skills into their learning plan and assessment development. Global21 is West Virginia’s new learning platform, combining a solid foundation of content knowledge in core subjects with three broad standards, which are defined in Policy 2520.14 (<http://hwvde.state.wv.us.policies/>) to promote 21st century performance skills:

- *Standard 1. Information and Communication Skills*
- *Standard 2. Thinking and Problem-solving Skills*
- *Standard 3. Personal and Workplace Productivity Skills*

The fundamental components of Global21 are standards and assessments. West Virginia, working with CTB/McGraw-Hill, has developed the nation's first assessment program that specifically addresses 21st century skills. This partnership has incorporated new student mastery requirements by constructing more rigorous depth of knowledge item types in the development of WESTEST 2. Also, the online writing prompts of WESTEST 2 were designed around the three skills of the Standards defined above.

Acuity benchmark/formative assessments, Writing Roadmap™ and the WESTEST 2 online writing allow students to utilize and demonstrate their content knowledge and technology skills by successfully accessing and using these electronic assessments tools as per the Global21 vision. All of these formative assessment tools allow West Virginia to balance the state's summative assessment system by funding formative assessment tools that provide feedback to students and inform instruction on a regular basis during the instructional year.

Effective reports and tools:

- *Enable real-time data-driven decisions that lead to school improvement, enhanced student performance*
- *Allow comparisons of results over time and across individuals and groups*
- *Forge a connected community of educational leaders sharing ideas and best practices*

Assessments for Special Needs Student Populations

These assessments should effectively measure the progress of special needs student populations and ensure they receive the instruction and remediation required to support learning. The interconnection of assessment and instruction for these populations is critical to ensure their ongoing progress and ability to succeed in today's accountability environment. These assessments include:

- Assessments designed for students with cognitive disabilities that prevent them from participating in general classroom assessments, (e.g.). performance task assessments, or portfolios of work which are collected throughout the school year and assessed at year's end
- English language proficiency assessments and resources that enable more learners to reach their education goals
- The incorporation of universal design principles to the extent practicable

Individualized Reporting and Instructional Tools/Resources

Individualized reports and instructional tools based on assessment data help each student achieve his or her personal best, provide teachers with the information needed to address student strengths and areas requiring improvement, and help parents stay engaged in their students' learning. Effective reports and resources:

- Provide clear explanations of scores, state standards, and curriculum goals
- Are individualized for each student
- Are offered online to provide immediate access to student results and information about state assessment programs
- Optimize interactive capabilities and include links to instruction and remediation
- Target effective instructional strategies and best practices and enable teachers to differentiate instruction
- Combine Web-based reports from multiple data sources into easily understood formats
- Enable real-time data-driven instructional decisions that improve student performance
- Go beyond static data displays and provide dynamic, actionable information
- Allow comparisons of results over time and across individuals and groups
- Forge a connected community to share ideas and best practices

Professional Development

Effective professional development for teachers is a critical component of a successful assessment program. As with other aspects of 21st Century assessment, technology will play a large role in strengthening professional development. Effective professional development programs educate teachers on new learning tools and methods, and enable them to:

- Analyze assessment data at the student, school, and district levels
- Transform data analyses into enhanced curricula, and target instruction at individual and group levels
- Adjust instructional styles to meet specific student needs
- Acquire new knowledge to expand their skills base
- Effective professional development programs include:
 - On-site programs
 - Teacher-led online programs
 - Web-based modules that teachers can access anytime, to provide targeted information to plan assessment administrations

Effective professional development for teachers is a critical component of a successful assessment program. ... As with all 21st Century assessment components, technology will play a large role in strengthening professional development offerings.

Parental Engagement

A key component of student success is parental involvement. An ideal assessment program encourages this, and also provides parents with support, in a practical and convenient online format - especially important in the era of working-parent households. Key parental engagement solutions include:

- Personal learning plans based on the analysis of students' specific needs and strengths
- Family-friendly home activities, planning tools, resources, and advice
- Guidance for acting on assessment information at home - a proven way to improve student performance
- Activities to prepare for state standards mastery
- Online accessibility from any computer or handheld device
- Community information and resources for parents and students, to strengthen learning and the home-school connection

21st Century Classroom – The Assessment and Reporting Vision in Practice

What will the ideal balanced assessment and reporting system look like? We envision an integrated system in which targeted, engaging, and differentiated instruction is informed by data from multiple assessments, enabled by technology and bolstered by professional development. A personalized, motivating learning environment, accessible online and with immediate feedback, will inform individualized learning, including critical thinking and problem-solving skills. 21st Century Skills are not easily assessed using traditional technologies and will be better measured with the use of innovative test items, real-world simulations, and computer-adaptive assessments focused on assessing student performance

based on real-time student responses. Assessments will be scored immediately with Artificial Intelligence, economizing precious teacher and instructional time. Scores will be aggregated and disaggregated to enable comparison and trend information—from the individual to the district, state, national, and international levels with sub-group analysis.

Online reports will be customized for different users of the information with anytime/anyplace access—an impossible task but for today's technology.

Assessments will be delivered in several ways, using a range of high- and low-tech methodologies, including but not limited to teacher observation, group discussions, student portfolios, paper-and-pencil assessments with scanned scores, computers and hand-held devices, clickers (student response devices), smart phones, touch screens, and new technologies yet to be developed. Technology would mitigate the need for many special accommodations, and broaden the opportunities for those that are still necessary.

Assessments will be delivered in several ways, using a range of high- and low-tech methodologies, including but not limited to teacher observation, group discussions, student portfolios, paper-and-pencil assessments with scanned scores, computers and hand-held devices, clickers (student response devices), smart phones, touch screens, and new technologies yet to be developed.

Students will take tests at varying intervals, depending on the purpose. Interim, benchmark, or formative assessments will be used, with high predictive validity for performance on summative assessments, and will also provide targeted interventions. Assessment of students' soft skills such as flexibility and adaptability, self direction, social skills, productivity and accountability, and leadership and responsibility, would provide information on college and workplace readiness. Individual student study guides will be provided in paper and/or electronically. Parents will play an even greater role in students' education through access to online reports and support resources. These will help families review students' areas of strength and those requiring improvement, as well as provide links to tutorials and other resources, and school and community contacts, enabling parents to help their students truly succeed.

Summary

The power of assessment to improve instruction is greatest when the assessment occurs naturally within the learning environment, and the feedback is targeted and immediate. Effective 21st Century assessment programs may include standardized tests that present a view of student performance based on national comparisons; formative classroom assessments that identify learning gaps and provide opportunities for remedy; performance assessments that allow students to show their ability to do in-depth work; and summative assessments that gauge student mastery of learning standards. Assessment data become actionable through reports that not only provide assessment results but also present that data in ways that help educators as well as students know what is effective and what could be improved; and parent-friendly reports and resources that enable families to participate in the learning process. Technology enables real-time assessment administration, scoring, reporting, and remedial interventions in ways previously impossible.

Assessments and related resources not only capture data on learning and achievement but are fundamentally changing the way teachers teach, students learn, and parents engage.

Assessments and related resources not only capture data on learning and achievement, but are fundamentally changing the way teachers teach, students learn, and parents engage. They provide essential information to guide genuine education reform and improvement, and they allow us to provide students with the quality education they need to compete and thrive in the world economy.

Assessment for Teaching and Learning

Margaret Heritage
CRESST/UCLA

**Exploratory Seminar: Next Generation K-12 Assessment
Systems**

ETS, December 7-8, 2009

Copyright © 2009. All content is protected and cannot be reproduced without written
consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



Overview

- 1- Assessments to Lead Learning
- 2- Descriptions of Learning
- 3- An Instructionally Useful Assessment
Framework
- 4- Supporting Teachers

Copyright © 2009. All content is protected and cannot be reproduced without written
consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



Assessments to Lead Learning



“The only good kind of instruction is that which marches ahead of development and leads it; it must be aimed not so much at the ripe as at the ripening functions.”

(Vygotsky 1986, p. 188)

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



The only good kind of assessment is that which identifies potential development – “the ripening functions” – so teachers can march ahead and lead development.

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



How Do Teachers Lead Development?

- Structure new experiences that build on and extend previous learning
- Plan interactions between and among teacher and students to:
 - support engagement with, and learning from the experiences
 - make connections between prior and new learning to extend learning
 - provide feedback
- Support metacognitive activity and self-regulation

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



Current Assessment

- Assessments measure present performance (past-to-present model)

“what has already matured to the present day”
(Vygotsky, 1933/1935, p. 120).

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



Current Assessment

- Assessments are static

“the examiner presents items, either one at a time or all at once, and each examinee is asked to respond to these items successively, without feedback or intervention of any kind. At some point in time after the examination is over, each examinee receives the only feedback her or she will get: a report card or a set of scores. By that time the examinee is studying for one or more future tests”
(Sternberg & Grigorenko, 2002, p. vii).

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



Problems for Teachers (and Learners)

- Develop new learning that builds on the present state

“overcoming the present state of being through a process of relying on presently existing psychological functions in the service of developing novel ones”

Valinsler & Van der Veer, 2001, p.38

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



Problems for Teachers (and Learners)

- Insufficient feedback for teachers and students

*Got it – move on to new topic
Didn't get it – reteach*

- No indication to teachers about what is within students' reach

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



Two Models of Assessment

PAST-TO-PRESENT

PRESENT-TO-FUTURE



Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



Assessment to Assist Learning

- Retrospective:
 - Fix the point of what students can do on their own
- Prospective:
 - an indication of the “zone of nearest development, i.e., those processes in the development of the same functions, which, as they are not mature today, still are already on their way, are already growing through, and already tomorrow will bear fruit” (Vygotsky,

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



-
- Independent performance to determine a person's actual level of development does not cover the whole picture of development
 - Responsiveness to mediation provides insight into an individual's future development
(Vygotsky, 1998)
 - Future development is what teachers and students must concern themselves with

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



Dynamic Assessment

- Accumulated knowledge is not the best indication of ability to acquire new knowledge
- Individuals function at less than 100% of capacity
- Best test of any performance is a sample
- When obstacles that could mask performance are removed, greater ability than was suspected is often revealed

(Haywood & Tzurial, 2002).

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



Present-to-Future Models

- How does the student respond to assistance?
 - Present students with a task, then a +1 assist, a + 2 assist and so on until they reach the point when they could go no further with assistance
 - Results would characterize the region of tasks between what the learner could accomplish alone and what could be accomplished and ultimately mastered with assistance

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



Zone of Nearest Development



- Structure new experiences that build on and extend previous learning
- Provide students with feedback about their learning and what to do to move forward
- Support metacognitive activity and self-regulation

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



Supporting Learning IN the ZPD

- Assessments include a second component of questions/probes/tasks for teachers to use while supporting learning in the ZPD
- Provide teachers with the scaffolding and formative assessment strategies they can use to keep move learning forward in the ZPD.
- Strategies can be used to provide feedback and support metacognitive activity and self-regulation.
- Upper boundaries of learning potential will change as the student moves to independent competence
- Teachers will need to determine when to administer the next assessment to determine the ZPD

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



Descriptions of Learning



Descriptions of Learning

- Common Core Standards provide information of what is expected at the end of each grade level
- A conceptually coherent view of learning?
- Knowledge, skills, and conceptual understanding developed together in a mutually reinforcing way?
- Can they lead to the development of instructionally useful systems?

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



What Informs (Should Inform) the Standards?

- Characteristics of knowledge and thought at advanced stages of learning and practice
- Tightly connected schemata
- Component parts of an idea
- Attention must be paid to both the development of connected ideas, and to the underlying subcomponents of those ideas that interact to create networks of schemata, so learners reach “mastery of the connexity and structure of a large body of knowledge” (*Bruner, 1960, p.3-4*).

“Conceptual power derives from taking a complex cognitive phenomenon and analyzing it into its underlying components” (Anderson, Reder, & Simon, 1998, p. 247).

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



Learning Progressions

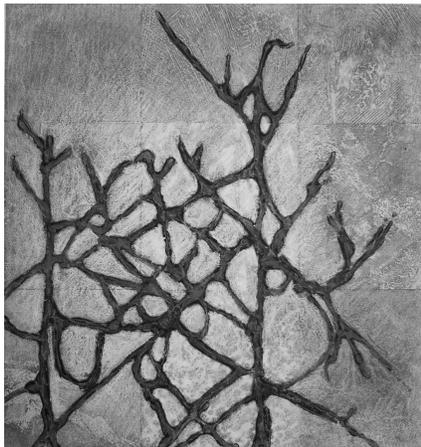
- Teachers need a trajectory of how learning develops
- Learning progressions lay out the important concepts and skills of the domain in a connected network that represents how competence in a domain develops
- Progressions describe how competence from rudimentary understandings and skills progressively increases in sophistication

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



Developing Progressions



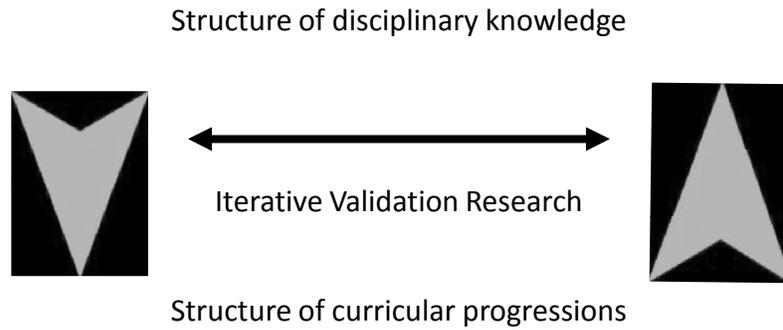
- Not necessarily linear
- Represent learning in one domain that supports another
- Language: to support thinking and language development

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



Developing Progressions

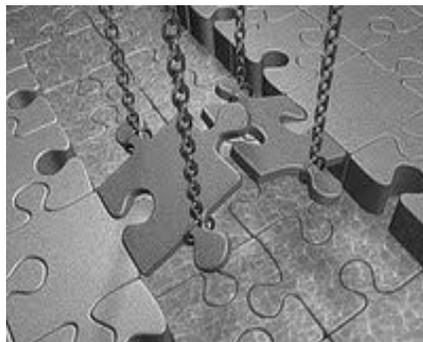


Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



Bridging Cognition and the Classroom



- Translate what is known about expertise
- Framework to integrate curriculum, instruction and assessment
- Deepen teachers' knowledge

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



An Instructionally Useful Assessment Framework



Learning Progressions and Assessment

- Assessments mapped to learning progressions in line with cognitive and information processing research
- Focus on assessing the development of schemata and subschemata and skills
- 3 levels of assessment operating within different assessment cycles
- All levels are complementary (built on the same model of learning) and all support teaching and learning

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



An Assessment Framework

Focus	Purpose	Cycle	Description	Scoring
Milestone performance (schemata/skills)	Accountability Consolidation of learning	<i>Long:</i> a period of instruction spanning months or even longer than a year, depending on complexity of concept/skills	Use of knowledge and skills in novel situations (assessing networks of schemata) Extended opportunities to represent knowledge and skills Digital collection of artifacts, including audio, video	Performance descriptors for teacher moderation *-levels to show the extent of consolidation In addition, expert sampling for accountability
Developing subschemata/skills	To determine the students' ZPD	<i>Medium:</i> Frequent assessment based on teacher judgment of when to administer or embedded in curricular units	Assisted performance +1, +2 etc. Use of technology to administer	Teacher and student determine outer limit of performance
Scaffolds/probes linked to medium cycle.	To keep learning moving forward in the ZPD	<i>Short:</i> used while students are learning	Assisted performance decreasing	Teachers and students interpret performance based on what the students are able to do unassisted and where they still need assistance. Teachers make adjustments to teaching and learners to learning along the way. Interaction between teacher and students and among students.
Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.				

Assessment Framework

- Provide teachers with the information they need to assist learning
- Permit students to be involved with their teachers in assessing and monitoring their own learning
- What about affect, motivation, self-regulation?

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



Implications for Teachers (and Learners)

- Investment in developing teacher knowledge and skill:
 - ✓ Knowledge about what it means to develop competence in a domain/across domains
 - ✓ Understand teaching/learning is functionally interdependent with the developmental processes that are emerging
 - ✓ How to use assessment information integrated into instruction to structure experiences in the ZPD

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



Supporting Teachers

- Vastly better descriptions of learning
- Significant changes in the content of pre-service and in-service programs and ongoing professional communities
- Assist teachers to develop models of how students' thinking and skill develops in a domain/across domains

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



Supporting Teachers

- Including students in the process of assessment and learning
- Cost-effective and time-efficient assessment
- Teacher moderation will require strategies for training teachers, especially with regard to understanding performance criteria and comparability of judgment
- Contractual changes – more time for reflection and planning

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

National Center for Research on Evaluation, Standards, & Student Testing



National Center for Research
on Evaluation, Standards, & Student Testing

UCLA | Graduate School of Education & Information Studies

VISIT US ON THE WEB
cresst.org

mheritag@ucla.edu

Race to the Top Assessment - Written Input

Kentucky Department of Education
Commissioner Terry Holiday
December 2, 2009

Kentucky supports the creation of a Common Core Assessment System, and we look forward to being a partner in the work. In March 2009, Kentucky legislators revamped our assessment system with specific requirements that start in the 2011-12 school year. With Kentucky's Senate Bill 1 as our reference, we would like to see a nationally-based model that provides for the following:

- a balanced assessment system aligned to support an emphasis on college and career readiness, including these components:
 - professional learning focused on formative classroom processes that help teachers create better instructional practices. Federal local education agency (LEA) pass-through dollars could be used to support development of professional learning for teachers
 - formative classroom tools, such as a bank of items linked to the Common Core Standards that may be used by teachers to create real-time feedback for their instruction
 - interim (benchmark) tests to provide curriculum feedback over the course of the school year and matched to both the common core standards and the summative common core test; interim assessments should be online for quick turnaround of results
 - a high-quality summative Common Core Test that measures knowledge and higher-order thinking skills by using a variety of item types, such as constructed response, performance-based and multiple-choice
 - an immediate focus on grades 3-7 reading and mathematics, with science and social studies brought to operational levels as soon as possible (High school needs are best met by end-of-course testing models; thus we leave them out of the Common Core Assessment system.)
- a vertical scale incorporated into the summative tests that could be used for a longitudinal growth model
- a set of national user group norms or profiles that provide a way to compare state to state and to provide student and national/regional scores (Ideally, the national profile would be linked to show how scores would compare to international norms.)
- a commonality of scores across states
- a set of common core alternate standards and a national model for 1 percent alternate assessment and 2 percent modified assessment

- a method for creating an accountability system using results from the formative, interim, summative and alternate assessment to meet federal and state requirements
- a technology plan that calls for interim tests to be administered online in the first generation (computer adaptive testing has great potential) and discussion of how the summative test may move to online in its second or third generation as new technology and one-to-one student/computer access evolves in the next five to ten years
- a financial plan that uses federal stimulus money to develop and maintain the test components for the future, but calls for states to purchase the tests to support ongoing use
- a set of new NCLB guidelines that support the work of developing a national assessment system

Development of the balanced assessment system should use a consortium model similar to the Achieve ADP End-of-Course Algebra II efforts. The development of the first generation should be delivered for spring of 2012; however, we must begin thinking of future test generations that can take advantage of online opportunities to pose creative, real-life simulations for test problems.

What I have listed is ambitious, but it is the right time to move our country to a nationally supported balanced assessment system. It's the right time to help our teachers focus on common standards and methods to change instruction to meet students' needs, and it's the right time to agree that the variety of state summative assessments must be aligned into a national summative test so we make sure our students and teachers are focused on national standards. We look forward to this work!

Terry Holliday, Ph.D.
Kentucky Commissioner of Education

TH:kd



**CONSORTIUM FOR CITIZENS
WITH DISABILITIES**

December 1, 2009

Subject: Race to the Top Assessment Program

The Consortium for Citizens with Disabilities is a coalition of nearly 100 national consumer, advocacy, provider and professional organizations headquartered in Washington, D.C. Since 1973, CCD has advocated on behalf of people of all ages with physical and mental disabilities and their families. CCD has worked to achieve federal legislation and regulations that assure that the 54 million children and adults with disabilities are fully integrated into the mainstream of society. Students who receive special education supports and services account for 13.5% of public school enrollment. They are disproportionately minorities and 24% of students receiving special education live in poverty as compared to 16% of the general population (Source: Overview Of Findings From Wave 1 Of The Special Education Elementary Longitudinal Study (SEELS), June 2004.) Approximately 50 national organizations participate in the CCD Education Task Force.

CCD believes that the development of common, high-quality assessments aligned with a common set of K-12 standards provides an unprecedented opportunity for equity among diverse learners, including students with disabilities.

As the Department considers the development of the Race to the Top Assessment Competition, CCD urges the Department to focus on the following areas for the next generation of summative assessments:

1. Create assessments that are accessible to diverse learners.
2. Create better Alternate Assessments based on Alternate Achievement Standards (AA-AAS).
3. Do Not Fund the Development of the Alternate Assessment based on Modified Achievement Standards.
4. Require assessments that embed individual student accommodations and allow student control over the test environment.
5. Require research to support any testing accommodation considered as non-standard.
6. Require any "adaptive testing" be aligned with grade-level standards.
7. Require empirical analyses of test items including the study of interactions between specific items and specific student populations.
8. Create assessments that provide meaningful feedback to educators and families.

1. Create assessments that are accessible to diverse learners.

CCD believes the true solution is to design assessment systems differently from the start, creating them from the outset to be accurate for the widest range of students, including those with disabilities. Universal Design for Learning (UDL) provides the foundation for research-based guidelines for creating flexible and valid on-line, computer-based assessments (see *Universal Design for Computer-Based Testing Guidelines* Pearson Educational Measurement & CAST, June, 2009; <http://www.pearsonedmeasurement.com/cast/index.html>) building upon prior physical and sensory access-oriented Universal Design for Assessment work (Thompson, Johnstone, & Thurlow, 2002).

A UDL approach also offers guidance for enhancing student engagement and persistence. Flexibility in recruiting attention, sustaining effort and supporting self-regulation are all highly individualized and nearly impossible to address without employing the inherent transformability, discrimination and data collection of digital media. The proponents of computer adaptive testing often point to the “automatic” difficulty adjustments of that approach as enhancing student engagement by decreasing the challenge presented to them. This is the same rationale used to support the simplification of the curriculum for struggling students, identical to the “out of level” testing that results in moving students with disabilities further away from the mainstream curriculum. Universal Design for Learning seeks to maintain high achievement standards for all students through the use of customized scaffolds and supports that reinforce the importance of maintaining grade-level expectations for all learners.

While UDL was originally conceived for students with disabilities, CCD believes it is critical to recognize that UDL can benefit all students. UDL offers a way to design assessments that will accommodate flexible goals and needs for a variety of learners. By presenting material through several means, assessments that are based on UDL allow several types of learners to access the material and demonstrate their knowledge.

UDL offers ways to address multiple learning needs and provide a better picture of student’s abilities. An assessment can only be considered an accurate picture of a student’s knowledge and skills if it is designed to allow a student to most effectively demonstrate what they know. Funding grants which incorporate principles of UDL is essential to help reveal a more accurate picture of how all students perform.

Therefore, as the Department moves forward in considering what elements grantees should include in their application, CCD urges the Department to include UDL and utilize the Center for Applied Special Technology (CAST) and the National UDL Taskforce, as valuable resources.

CCD also urges the Department to fund innovative test delivery models particularly on-line or digital delivery systems. The advantages of online assessment include:

- immediate score reporting so test results can guide instruction
- decreased administrative burdens on school personnel
- increased security of testing materials, and
- more flexible test scheduling.

Additionally, online/digital assessment environments allow maximum flexibility for any additional individual accommodations required by students.

Digital technologies offer a flexible base for representing assessment items in multiple ways and with which the equivalence of underlying constructs can be maintained (Honey, Pansnik, Fasca, 2007; Rose, Meyer, & Hitchcock, 2005; Meyer & Rose, 2006). Digital multimedia can present the same underlying construct in different “surface” representations - text, audio, image, video, etc., thereby reaching a greater range of student. Further, the ease by which digital tools can discriminate one item from another can be used to provide each student with customized supports for construct irrelevant items while simultaneously diminishing those supports for the items actually being assessed.

Digital media can also allow students to express what they know in multiple ways. For response demands to be equivalent for all students (a prerequisite for test validity), students must be allowed to respond optimally, employing areas of strength. If students can respond in flexible and customizable ways, construct-irrelevant barriers can be significantly reduced.

2. Create Better Alternate Assessments based on Alternate Achievement Standards.

As you know, current federal regulations allow states to develop and administer alternate assessments based on alternate achievement standards (AA-AAS) for a limited number of students with the most significant cognitive disabilities. While this policy has been in place for some time, the consistency and availability of these assessments varies widely between states. A recent study by the National Center for Special Education Research, within the Institute Of Education Sciences, found that many states approach the AA-AAS differently (Cameto, R., Knokey, A.-M., Nagle, K., Sanford, C., Blackorby, J., Sinclair, B., and Riley, D. (2009). Some states use a portfolio or body of evidence to constitute the entire assessment. Others use techniques such as a rating scale/checklist, performance task/events, or multiple choice/constructed response assessments. The inconsistent approach to these assessments across states creates varying standards and expectations and fails to provide the information we need to accurately judge programs.

We also know from a new 7-state survey conducted by the National Alternate Assessment Center that 75 percent of the students participating in state AA-AAS are reading sight words and using a calculator to do basic math operations. This finding suggests that many students assigned to this assessment may, in fact, be capable of participating in more rigorous assessments.

3. Do Not Fund the development of Alternate Assessments Based on Modified Achievement Standards (AA-MAS) through this grant program.

Many students with disabilities can achieve grade-level work when given the right access to high quality instruction, with qualified teachers and appropriate accommodations for both instruction and assessment. In fact, we now know from data collected by the National Center on Educational Outcomes (NCEO) and through grant-funded work among several states that students with disabilities perform across the proficiency range on state assessments (performance achieved without full and equitable access to instruction in the general curriculum by qualified teachers). IDEA eligibility does not and should not pre-determine that a student will perform below grade level. In fact, several studies confirm that students without disabilities perform well below grade level. Thus, many states are opting to not develop the AA-MAS.

4. Require assessments that embed individual student accommodations and allow student control over the test environment.

Researchers have developed systems of online testing environments that provide accommodations that adjust to individual student preferences on demand (such as those developed by Nimble Assessment Systems) as well as online accommodations decision-making tools (such as STELLA developed by Rebecca Kopriva and colleagues at the University of Wisconsin) that increase test validity. Research shows that accommodations delivered within a computer-based testing environment increase the consistency and integrity of accommodations and result in improved utilization by the student. Students should be provided with an optimum testing environment that allows maximum student engagement and persistence.

5. Require research to support any testing accommodation considered as non-standard.

Studies conducted on testing accommodations show that many states are currently implementing test accommodation guidelines that are not defensible through research. While test develop employing UDL coupled with online testing environments are sure to eliminate the need for many test accommodations required in traditional tests, some accommodations will continue to be needed by certain students. Common assessments based on a common set of standards provide for the development of a common set of test accommodations across states. Any accommodation considered to be construct-relevant—to impact the skill being measured by the test—must be supported by rigorous research evidence. The standardization of test accommodations across states will dramatically improve both the validity and comparability of test results, making test data more useful to educators, parents and policymakers.

6. Require any “adaptive testing” be aligned with grade-level standards.

While online testing environments hold great promise, they also offer opportunity to lower student expectations through “adaptive” approaches that adjust item difficulty based on student responses. Such approaches are not appropriate for summative assessments used for system accountability. While computer adaptive testing might be useful for formative assessment, its use in summative assessment would surely lead to a decrease in the challenge to certain students and a lowering of academic expectations for those students. The current ESEA testing requirements do not allow for “out-of-level” testing. This standard has resulted in the demise of a heretofore-widespread practice for students with disabilities. Today, schools are being held accountable for the performance of students with disabilities on general assessments with only limited exceptions. This advancement has resulted in improved access to the general curriculum, expanded learning opportunities and heightened expectations for millions of students. Therefore, any computer adaptive testing developed under this assessment program initiative for use as a summative assessment must be aligned to grade-level standards verifiable through rigorous peer review. No exceptions for diverse learners such as students with disabilities and English language learners should be permitted.

7. Require empirical analyses of test items including the study of interactions between specific items and specific student populations.

Test items should be analyzed to ensure that they do not disadvantage certain populations of students in their format and/or linguistic complexity. Research studies, such as cognitive labs, should be designed to investigate the interaction between students and test items. Interactions will differ within one broadly defined population of students (for example students with learning

disabilities); therefore reviewing items in the absence of their specific interactions with students is insufficient. For assessments to provide useful results, all learners and their specific needs must be included in test development procedures, the field-testing of items, and post-hoc analyses of item by student interactions.

8. Create Assessments that Provide Meaningful Feedback to Educators & Families

As the Department considers its grant proposal, CCD encourages the Department to place a strong emphasis on the importance of creating assessments that yield meaningful information for educators and families. Assessments should be tools that help inform instruction, identify areas of strength and weakness, and help inform decision making. However, assessments can only be effective if they are presented in a way that enables a student to accurately demonstrate their knowledge and skill. Educators need meaningful professional development to help them understand how to use assessment data to inform and drive instruction. Parents need to understand what complex scores show about how their child is learning, and educators must be able to describe results and help parents interpret this complex data meaningfully.

To this end, CCD encourages the Department to fund grants that included professional development and training. Considering how assessments can provide meaningful feedback to educators and parents from the first stage of assessment creation will help ensure their success.

Thank you for the opportunity to comment on this important initiative.

Sincerely,

American Association of People with Disabilities
Association of University Centers on Disabilities
Children and Adults with Attention-Deficit/Hyperactivity Disorder
Council of Parent Attorneys and Advocates
Council for Learning Disabilities
Disability Rights Education and Defense Fund
Easter Seals
Helen Keller National Center
Learning Disabilities Association of America
National Coalition on Deaf-Blindness
National Disability Rights Network
National Down Syndrome Congress
National Down Syndrome Society
School Social Work Association of America
The Advocacy Institute
The Arc of the United States
The National Alliance on Mental Illness
The National Center for Learning Disabilities
The National Parent Teacher Association
United Cerebral Palsy

Implications of current policy for educational measurement

Daniel Koretz
Harvard Graduate School of Education

Next Generation K-12 Assessment Systems
Educational Testing Service
Princeton, NJ
December 7, 2009

The policy context

- Half-century shift in functions of large-scale assessment
 - Monitoring and accountability ever more important
- NCLB and RTT are an intensification of this trend
- Repeated cycle:
 1. Proclamations of success from inflated scores
 2. Later, crisis of inadequate performance
 3. New iteration, usually more severe

Why is this a problem?

- Lack of persuasive evidence of large-scale positive effects
- Persuasive evidence of unwanted side effects
 - Degraded instruction
 - Gaming
 - Score inflation
- Large research literature showing similar problems in performance-based accountability systems in other fields



What should be done in response?

- How should educational accountability systems be improved?
- How should tests—and testing programs—be altered?
 - What has the field done so far in response?
 - What additional does the field need to do?



What the field has done to date

- Changes in test design
 - Attempts to test higher-order skills better
 - New designs to encourage certain styles of instruction and classroom assessment
- Statistical and psychometric innovations, e.g.:
 - Growth modeling
 - Standards-based reporting (largely a change for the worse)
- RTT seems to continue these directions



What the field has not done

- Core enterprise has been largely unaffected by:
 - Research in other fields on accountability systems
 - Research on problems of test-based accountability
- Has not adequately addressed implications of using tests for accountability:
 - How test design should be modified
 - How validation must be augmented
 - Whether routine operation of testing programs should be changed



Campbell's Law

- Basic framework, from G. Baker:

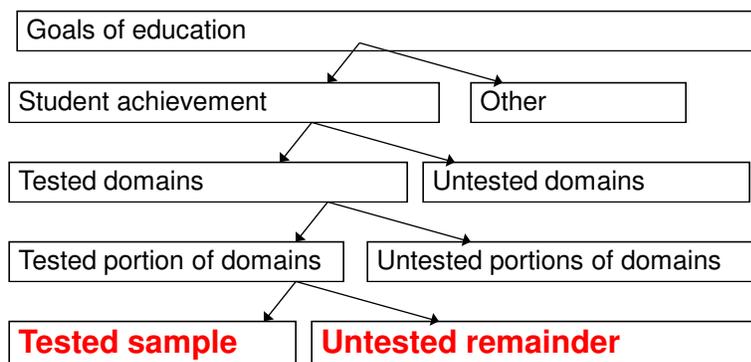
$$V = \mathbf{f} \mathbf{a} + \varepsilon$$

$$P = \mathbf{g} \mathbf{a} + \phi$$

- Where \mathbf{f} and \mathbf{g} are vectors of marginal products of actions
- The performance measure is incomplete: \mathbf{g} omits important actions
- How does this play out in test-based accountability?



Sampling in constructing a test



Two aspects of incomplete sampling at final stage

- Substantive: *predictable* recurrences of parts of the tested domain (target of inference)
 - Instructional response: reallocation
- Nonsubstantive: *predictable* recurrences of item style, response demands, substantively unimportant bits of content
 - Instructional response: coaching

Consequences of sampling

- Low stakes: modest
 - Measurement error (uncertainty)
 - (Usually) modest differences among tests (e.g., PISA vs. TIMSS)
- High stakes: very large
 - Incentives to focus on the tested sample, not the domain
 - Narrowed instruction, bad test preparation
 - Score inflation

↑ Algebra 1

J.1

J.2

J.3

J.4

J.5

J.6

J.7

2003S#17(o)

2003S#38(m)

2002F#37(m)

2000S#36(m)

Source: Quincy MA High School Math Dept.



HARVARD
GRADUATE SCHOOL OF EDUCATION

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Coaching: based on an incidental characteristic of test items

Whenever you have a right triangle—a triangle with a 90-degree angle—you can use the Pythagorean theorem....The sum of the squares of the legs of the triangle (the sides next to the right angle) will equal the square of the hypotenuse (the side opposite the right angle)....

Two of the most common ratios that fit the Pythagorean theorem are 3:4:5 and 5:12:13. Since these are ratios, any multiples of these numbers will also work, such as 6:8:10, and 30:40:50.

Princeton Review, *Cracking The MCAS Grade 10 Mathematics*



HARVARD
GRADUATE SCHOOL OF EDUCATION

12

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

NY grade 7 item, 2008

27 Which tool is **most** appropriate for measuring the mass of a serving of cheese?

- A ruler
- B thermometer
- C measuring cup
- D weighing scale



NY grade 7 item, 2009

9 Which tool would be **most appropriate** for Natasha to use when finding the mass of a watermelon?

- A scale
- B inch ruler
- C meter stick
- D measuring cup



An example of coaching (cheating?)

“The question on the review sheet for...[the] exam...reads in part:

‘The average amount that each band member must raise is a function of the number of band members, b , with the rule $f(b)=12000/b$.’

The question on the actual test reads in part:

‘The average amount each cheerleader must pay is a function of the number of cheerleaders, n , with the rule $f(n)=420/n$.’”

Strauss, V., *The Washington Post*, July 10, 2001, p. A09



What should be done?

- Make more than test scores count
- Limit predictable recurrences in tests
- Expand validation/ evaluation



Limiting predictable recurrences

- Reduce both substantive and nonsubstantive recurrences to improve incentives
- Reduce predictable recurrences overall
 - Reduce affordances for coaching, undesirable reallocation
- Introduce planned variations
 - Responds to unintended recurrences
 - Can provide an audit function (e.g., 'self-auditing assessments')

Expand validation/ evaluation

- Traditional validation is necessary but inadequate
 - Cross-sectional, so cannot evaluate inferences about gains
 - Largely completed before inflation occurs
- Need to institutionalize audits and ongoing validation
- Need direct monitoring of effects on schooling

Revisit linking

- Most models depend on linking scores over time
- Key assumption of NEAT linking may be untenable under high-stakes conditions
 - Failure of assumption would build score inflation into the scale
- Alternatives to NEAT linking may be increasingly impractical

What does growth modeling do to problem of Campbell's Law?

- Not much
- Problem of bad incentives leading to undesirable practices and score inflation is unaffected
- Specific forms of inflation may change
 - E.g., issue of persistence of coaching effects over grades

What do complex performance tasks do to problem of Campbell's Law?

- May exacerbate the problem
 - Tasks are memorable
 - Reduced number of tasks increases impact of predictability
 - Construct-irrelevant recurrences may exert more influence
 - May be harder to avoid gratuitous recurrences while maintaining acceptable comparability

Next steps

- Not arguing against further development of growth modeling, performance assessment
- But they do not address the core problems of Campbell's Law:
 - incompleteness, predictability, and corruptibility of performance measures
- Use of tests for accountability and incentives must become a core concern of the measurement community

Supplementary slides

Fallacy of the “test worth teaching to”

- A good test is not enough to prevent inflation
 - Inflation does not require bad material on the test
 - Bad test prep can undermine a good test
- Score inflation depends on what is emphasized and deemphasized in instruction
 - If teachers de-emphasize important content not tested, scores become inflated

From: Sheri Krause [skrause@wasb.org]
Sent: Wednesday, December 02, 2009 2:49 PM
To: Race To The Top Assessment Input
Subject: Race to the Top Assessment Program

On October 26, 2009, the Department of Education requested input on a possible Race to the Top program for the development of and implementation of high quality assessments based on common standards.

The Department's notice stated: *If the Secretary determines that it is not feasible to conduct this second program, the \$350 million designated for this program will revert to fund additional grants under the general Race to the Top program.*

On behalf of the Wisconsin Association of School Boards, I **strongly encourage the Department to maintain this second program focused on high quality state assessment systems** and not to allow these funds to revert to the general Race to the Top program.

In Wisconsin, leaders at the school, district and state levels are prepared to transform our current state assessments into a high quality system that builds toward college and career readiness by the time our students' complete high school.

Federal support will be critical for Wisconsin to provide a system of world-class assessments for our students. The goal of developing high quality assessments based upon common standards is worthy of a second distinct program.

Sincerely,
Sheri Krause

*Sheri Krause
Government Relations Specialist
Wisconsin Association of School Boards
122 W. Washington Ave., Suite 400
Madison, WI 53703
Phone: 608-257-2622
Fax: 608-257-8386*



**RACE TO THE TOP ASSESSMENT PROGRAM
NOTICE OF PUBLIC MEETINGS AND REQUEST FOR INPUT**

**December 1 & 2, 2009
Denver, Colorado**

Questions on the Assessment of English Language Learners

On behalf of Californians Together, a coalition of 23 state-wide professional, parent, and civil rights organizations focused on improving policy and practice for English Learners, the following responses are submitted:

General Comments

To best determine the answers to the two questions posed by the Department, it is critical that the assessments and their variations be based upon an accurate English learner student profile through a strengths-based Theory of Action. This theory of action would account for the diversity of the English learner student population and mandates a variety of accommodations and assessment practices to specifically address the academic/content knowledge and language proficiency levels of ELs (e.g., native language testing, linguistic modification in English, etc).

The EL student profile should include indicators such as EL proficiency level, educational background in L1, and length of stay in US schools and program of instruction. Student profiles also would simultaneously serve to inform instruction, and follows students to track developmental, vertical and horizontal progress for ELs throughout their schooling trajectories.

To uniformly apply the attributes of validity and reliability to each state's assessments, the Department should require that each state submit psychometric evidence from the test developers on the validity and reliability of the assessments administered to a wide-range of English learners.

Question 1

Provide recommendations for the development and administration of assessments for each content area that are valid and reliable for English language learners. How would you recommend that the assessments take into account the variations in English language proficiency of students in a manner that enables them to demonstrate their knowledge and skills in core academic areas? Innovative assessment designs and uses of technology have the potential to be inclusive of more students. How would you propose we take this into account?

We highly recommend the following:

1. First year, beginning level students with little or no proficiency in English should be exempt from academic tests in their second language and the English proficiency test should serve as a proxy;
2. Recent immigrants (two years or less in US), speakers of indigenous languages, students with little or no schooling, students from war-torn countries with interrupted schooling and students without two

- consecutive years of educational experience in US (high mobility) should be exempt from taking academic tests in their second language for two years and the English proficiency test should serve as the proxy;
3. Assessments in reading/language arts need to be developed across the four language domains (listening, speaking, reading, and writing in L1* and L2) and across genres (narrative and expository texts);
 4. There is a need to expand the types of performance-based assessments by language domain, content-area, by genres and by EL proficiency levels;
 5. For elementary level students (grades 3-5/6) retellings (oral and/or written) in L1* and L2 be one pathway to assess students' comprehension and thus, allow students at different proficiency levels to demonstrate what they know and can do. The oral retell provides the opportunity for the teacher/school to gauge the ELD proficiency level simultaneously with reading comprehension. Scoring through valid and reliable instruments/rubrics such as running records, miscue analysis demonstrate growth and inform instruction;
 6. Retelling can be captured by audio taping and can be scored by teams of teachers to ensure reliability in order to document growth;
 7. Assessments need to inform instruction and go beyond filling in the bubble – performance based – e.g. writing in a variety of genre across all grade levels and content areas. Thus, assessment systems should include curriculum-embedded formative assessments as well as summative assessments;
 8. Oral language development assessment needs to be embedded within the content standards. According to the National Literacy Panel for Language Minority and Youth, there is an absence of oral language development in instruction across all grade levels and content. “What gets tested gets taught;”
 9. Linguistic complexity needs to be controlled in constructing tests for students beyond beginning levels of English proficiency and as they are developing English proficiency in all four language domains, e.g., especially for content area assessments in English;
 10. The accommodations recommended by the *Technical Advisory Panel on Uniform National Rules for NAEP testing of English Language Learners* should be implemented by states to standardize the inclusion of English learners in federal accountability systems beyond on the NAEP testing;
 11. Implement a temporary waiver of Annual Yearly Progress requirements while consortia engage in assessment reform;
 12. Experts in English Learner education and assessment from all levels (universities, local and state education agencies and practitioners) should be actively included in the policy development and decision-making on assessment.

Question 2

In the context of reflecting student achievement, what are the relative merits of developing and administering content assessments in native languages? What are the technical, logistical, and financial requirements?

The relative merits of developing and administering content assessments in native languages are as follows:

1. L1* testing results in an accurate picture of what students know and can do for students who receive instruction in that language or for those who are already literate in their home language;
2. Given the national movement around world languages in preparation of a global citizenry, and that Spanish and Chinese are the top two world languages, L1 testing should align and support other initiatives promulgated by the federal government, i.e. World Languages and Strategic Language Initiatives;
3. Including native language assessments would reverse the punitive nature of the current accountability system by eliminating the practice of stigmatizing students, and labeling schools and districts as program improvement based on a single test that does not measure what many students really know and can do.

Recommendations for technical, logistical, and financial requirements:

4. Double test only in Language Arts (L1* and L2) and test in one language for content areas based upon language of instruction or preliminary assessment;

5. Use the native language tests from states that already have developed them. Also learn from their experience and build upon them –do not reinvent the wheel;
6. Use the new competition for consortia funds to do the developmental work on native language assessments and to develop tools and resources for the various accommodations;
7. Include experts in primary language assessments from all levels (universities, local and state education agencies and practitioners) in policy development and decision-making of native-language assessments.

L1*- indicates that students are proficient in their native language as evidenced by a home language survey and/or through instruction in the native/target language.

Respectfully Submitted,

Magaly Lavadenz

Magaly Lavadenz, Ph.D.
Director, Center for Equity for English Learners
Loyola Marymount University
mlavaden@lmu.edu

Member Organization of Californians Together

Shelly Spiegel-Coleman
Executive Director
Californian's Together



525 E. 7th Street, Long Beach, California 90813, 562-983-1333, 562-436-1822 fax

Comments on Presentation on Implications for Policy by Dan Koretz

Robert L. Linn

CRESST, University of Colorado at Boulder

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Shift in Assessment Uses

- Ever expanding use for accountability
 - Modest levels of accountability for schools at both state and national level before NCLB (mostly through the publication of results)
 - NCLB created sharp increase in use of sanctions for schools
 - RttT promises further increases with uses for individual teacher accountability

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Score Inflation

- Dan has studied the issues of score inflation more than anyone I know
- Has provided convincing evidence in the past that score inflation is a major problem that undermines the validity of inferences from assessment results
- Good examples of how scores get inflated

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Accountability Effects

- Dan argues effects are largely negative and distrusts gains
- Agree that gains shown on state assessments are exaggerated due to score inflation
- Magnitude of score inflation largely unknown

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Center on Education Policy

- Analyses of assessment results from all states with sufficient data to evaluate trends for last several years
- Found that increases in state assessments were much more common than decreases since NCLB
- Doesn't prove that NCLB improved achievement

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

CEP Results

- Generally positive gains on state assessments in last few years
- Gains also generally found on NAEP but the gains are smaller on NAEP than on state assessments
- Gains larger in mathematics than reading and larger for elementary and middle school than high school

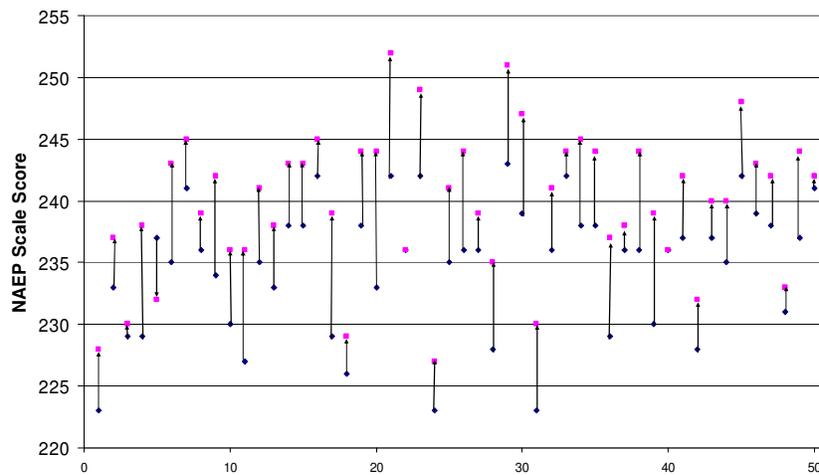
Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Number of States with Changes in State Means from 2003 to 2009 on NAEP Mathematics Assessments

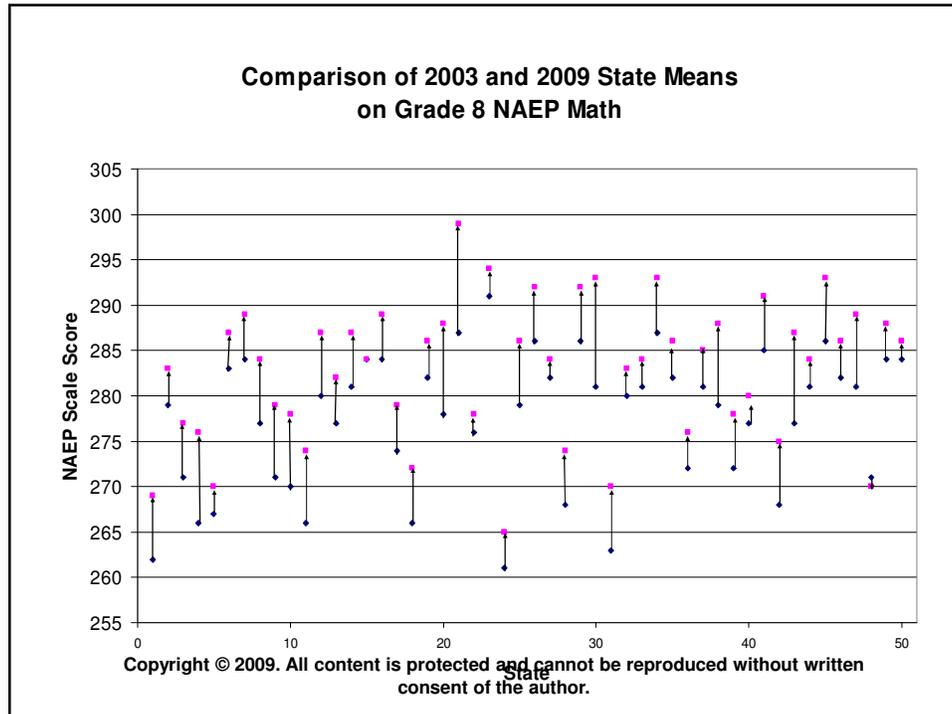
Change	Grade 4	Grade 8
Significant Increase	42	41
Non-significant Increase	6	7
No Change	1	1
Non-significant Decrease	0	1
Significant Decrease	1	0

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Comparison of 2003 and 2009 State Means on Grade 4 NAEP Math



Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.



Audit Functions

- NAEP has served an audit function for state assessment results
- Suggest that, while there is substantial inflation, of state test scores, there is also substantial improvement

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

Future Assessments

- Increase emphasis on growth in comparison to status
- Build audit function into ongoing assessments
- Improve linking designs to reduce inflation (e.g., use of links to more than one prior year)

Copyright © 2009. All content is protected and cannot be reproduced without written consent of the author.

A COMPREHENSIVE ASSESSMENT SYSTEM: TOUGH CHOICES FOR THE RTT ASSESSMENT COMPETITION

Scott Marion

National Center for the Improvement of Educational Assessment

**Expanded Comments based on my presentation to the November 12, 2009 Boston Race to
the Top Assessment Public and Expert Input Meeting**

December 1, 2009

Introduction

While USED made clear in the Notice that the focus of the comments related to the RTT assessment funds were not to be on accountability policies, I strongly urge USED to develop a clear conception of how the results from the proposed assessment system are to be used before considering all of the comments and crafting final rules and a potential Notice of Intent to Apply (NIA). Every assessment design discussion must begin with a clear explication of the purposes and uses of the proposed assessment system. The language in the Federal Register certainly implies certain purposes such as evaluating the “effectiveness of teachers and schools” as well as tracking the progress of individual students toward becoming college and work ready, but assessment designers should not have to guess at the nature of the accountability system being proposed. Assessment designers will better understand the challenges, constraints, and uses if USED (along with Congress and the White House) can be very explicit about the forthcoming accountability system. I have made some assumptions about the nature of the uses of the assessment system in my comments and responses to the questions, but I have tried to make these assumptions as clear as possible to help USED best interpret and use my responses.

Becoming crystal clear about the intended purposes and uses is also critical when designing assessment systems. Assessment design always involves trade-offs, especially when living in our current resource-constrained reality. A clear understanding of the proposed purposes and uses of the assessment system can serve as important touchstones when evaluating potentially competing design decisions. The set of “requirements” in the Notice appears to describe a “silver bullet” assessment system. In other words, as far as I know, there has never been a large-scale “summative” assessment system that has ever come close to fulfilling all of the ambitious requirements put forth in the Notice. Further and in case some are holding out hope that a comprehensive assessment system is a way to get around this dilemma, there is no evidence that any such system has ever been implemented. Without being pessimistic, I argue that something will have to give. Therefore, an explicit set of purposes and uses will serve as key touchstones during difficult discussions about design trade-offs. Again, I will try to make clear in my comments where I am privileging one design decision over another. For example, the rich performance tasks/constructed response items called for in the notice make year-to-year equating—another key requirement—much more challenging. Given the current technology, trying to balance these equally could mean doing neither very well. In the following paragraphs, I offer comments organized around the following key issues. For each of the sections below, I discuss the issue(s) and try to offer specific advice for drafting a NIA. The last section includes advice on drafting an NIA not found elsewhere in my comments.

- ✓ An explicit theory of action
- ✓ Purposes and uses
- ✓ Sound design principles

- ✓ My proposed design
- ✓ Innovation and timeframe
- ✓ Access and equity
- ✓ A note about psychometrics
- ✓ High schools
- ✓ Advice on the proposed NIA

A Preview of My Vision

The vision for a comprehensive assessment system that I outline below is a conceptually coherent system that incorporates explicit curriculum (or opportunity to learn (OTL)) components as a basis for building a valid assessment system. The system includes a cumulative end of year summative assessment component, interim performance tasks embedded in the curriculum units, and formative assessment tools and supports. The system also includes professional development focused on proper implementation of the standards, curriculum, formative assessment, and interpretation of the interim and summative assessment results. Finally, for too long we have considered our assessment reports as an afterthought. A reporting structure as comprehensive as the assessment system must be designed that facilitates decisions and actions to help reveal student and school strengths and weaknesses.

A Theory of Action

USED should articulate a clear and explicit theory of action, but at the least, USED should require an explicit theory of action as part of the NIA consortium proposal expectations. A theory of action outlines the intended components of the system, while clearly specifying the connections among these components. Most importantly, a theory of action must specify the hypothesized mechanisms or processes for bringing about intended goals. In the case of the NIA, the theory of action should describe how the particular clear goals will be achieved as a result the proposed assessment system(s). Further, USED should require proposals to clearly articulate how the educational system will get from “A to B” as a result the proposed system. In other words, what processes must be in place in order for the consortium to achieve its goals and what empirical evidence exists to support the proposed expectations? The theory of action must explicitly describe prioritized design choices, e.g., influencing and shaping teaching and learning or measuring existing knowledge, or making cross-state comparisons. The theory of action is a check on the logic of the underlying assumptions of the various proposals and should be a critical aspect of the NIA proposal scoring process. Again, a theory of action is not just a bunch of pretty shapes and arrows created with a piece of software. It must be an empirically and logically based argument that outlines how the specific proposed system will fulfil the stated goals and how it will do so.

Purposes and Uses

As I mentioned in my introduction, the plethora of design requirements in the RTT notice will stress any (even comprehensive) assessment system. A very important likely use of the assessment system is as part of the next generation accountability system. I understand that Congress will ultimately write the reauthorization of ESEA, but I also know that USED and the White House will influence the process. Therefore, USED should try to predict as well as it can the likely accountability uses before letting the NIA. Additionally, USED has put forth many purposes, uses, and design requirements in the Federal Register notice inviting comments. There

are simply too many competing priorities for any system to meet. I urge USED to prioritize its intended (or hoped for) purposes and uses as clearly as possible. If USED is unable or unwilling to undertake this prioritization, you should (must?) require any consortium proposals to state (as part of their theory of action) its prioritized purposes and uses. Clarity on purposes/uses will serve as an important touchstone during complicated design deliberations. This is yet another rationale for funding multiple consortia. For example, one consortium could have as its highest priority to build assessments to support value-added models for teacher accountability, while another proposal might focus on creating very innovative assessments designed to push teaching and learning. It would be hard, especially in the limited time frame, to do both of these well.

Overarching Goal and Prioritized Purposes and Uses

In an effort to practice what I am preaching, I will provide my vision for a future assessment system by first stating my **main** goal for the system.

- ALL students should have meaningful opportunities to develop deep understanding of important content and critical skills to allow for viable postsecondary choices (e.g., college/work ready) and for becoming contributing members of society.

I propose a system that is intended to support this overall goal, but first I specify my prioritized purposes and uses:

1. Providing students opportunities to develop robust knowledge and skills for use in novel and complex settings by measuring a limited number of big ideas at deeper levels of understanding.
 - Developing a system with a much more intentional integration of curriculum, instruction, and assessment because we cannot address these challenges with just an “assessment fix.”
2. Measuring student longitudinal growth as a foundation for valid accountability systems and as information for school improvement.

Notice that I am limiting myself to two main purposes, because I do not think a system can do more than two-three well. I am intentionally not focusing on cross-state comparisons, not because I think there is anything inherently wrong with cross state comparisons. Rather, I think my proposed purposes will help meet the overall goal better and focusing on cross state comparisons might distract the system from the main goals.

I provide these goals and purposes more as an example—although it matches well with my proposed system below—of the type of expectations USED should have for the NIA. Any consortium should be able to clearly articulate the main goals, purposes, and intended uses of its proposed system. If the proposer cannot do this, they will have trouble implementing any sort of innovative system.

Theoretically Based Design Principles

The NIA must require proposed designs to be based on theoretically sound design models. Much too often—almost always—current state assessments are designed based on fairly parochial practices in spite of significant recent advances. Bob Mislevy has said that modern psychometrics is [unfortunately] the application of 21st Century statistics to 19th Century psychology. We could be doing much better and this NIA provides the perfect opportunity for us to do so. Evidence-centered design (ECD, Mislevy, 1994, 1996) is one of the best examples

of such theoretically explicit design frameworks and is now being applied to the redesign of the Advanced Placement courses and exams. A landmark National Research Council publication, *Knowing What Students Know* (Pellegrino, Chudowsky, and Glaser, 2001) helped to clarify and expand on Mislevy's ECD. These are not the only theoretically-based design options, but there are not very many! If USED hopes that the \$350 million RTT assessment funds will transform educational assessment, it must require that consortium proposals adhere to a well-vetted theoretically-based assessment system design. The NIA must require proposers to demonstrate familiarity with the particular design framework and, more importantly, precisely articulate how they intend to put the design into practice.

A Vision for a Reformed Assessment System

I articulate a design for a specific comprehensive assessment system. I propose this vision because I think it can radically change teaching and learning in the United States. This proposal is designed to build a coherent system that bridges curriculum, multiple forms of assessment, and supports for instruction. I am not necessarily wedded to every detail put forth here (or omitted for the sake of brevity), but I am wedded to the main components linking curriculum, assessment, and instruction. Many other experts (e.g., Braun, Wise, Baker, Darling-Hammond, Shepard, Gong, Pellegrino, Abedi) described a similar vision of a system that linked curriculum and assessment although there were slight differences in the details. I have no doubt that these experts could easily come together to work out the differences in the specifics. Therefore, my main purpose here is to paint a clear picture of what a reformed system might look like and to argue that unless a comprehensive system is proposed, it is unlikely that a consortium would meet the USED's transformative goals for U.S. education.

I am proposing a conceptually coherent comprehensive assessment system that incorporates explicit connections to research-based curricular units and includes the following components:

- ✓ End-of-year summative assessments built on well-articulated content and performance standards
- ✓ Interim performance tasks embedded in mini curricular units
- ✓ Formative assessment supports/prompts
- ✓ Focused professional development
- ✓ Actionable reporting system to help reveal student and school strengths and weaknesses

Reporting System

Taking the last item first, I argue that reporting systems need to be considered as an integral part of the design process. Assessment reports—the only way that we really have of communicating about the assessment system to the public—are all too often seen as an “add-on.” The reports must be conceived as a system of reports, with reports designed for specific audiences depending on what information they need in order to make decisions appropriate to their respective roles in the system. Most importantly, the reports must be “actionable” in that they lead users to engage in appropriate inferences, decisions, and instructional/programmatic actions. These reporting structures must support the theory of action. See <http://www.schoolview.org/> for a terrific example of what's possible.

The NIA should require specific information about many aspects of reporting discussed here, even to the point of requiring proposals to submit at least mock-up report designs for different levels of users and to describe how these reports fit within the theory of action.

The Curricular Units

As discussed throughout these comments, decoupling assessment and curriculum leads to a very limited form of assessment and one that ensures that the “rich get richer.” Including curriculum or at least curricular units as part of the design helps to equalize opportunity-to-learn (not that it alone will make it equal) and allows for the development of richer assessment experiences. I recognize that the federal government or even state governments for that matter are wary of prescribing curriculum. However, if we are interested implementing a truly innovative assessment system, adopting an agnostic stance toward curriculum is selling the effort short. If it helps, I would be happy to refer to these curricular units as “opportunity-to-learn” units or “assessment supports.” But for now, I refer to these as curricular units.

Depending on grade level, I envision implementing approximately 2-6 of these units throughout the year, varied by grade level. I would phase-in the development and implementation of these units and associated assessments over time by either implementing one or two units per grade level and content area or focusing on a few key grades at first. I would also suggest implementing different types of units with some as short as a few days with others as long as a couple of weeks. Each unit should be focused on a “big idea” of the domain and should be used strategically within existing curricula (e.g., perhaps at the end of a longer unit of study). These units would be designed to instantiate key aspect of the common standards, but should also be designed to extend and deepen what at this time looks to be a very weak draft of the common standards.

These curricular units (or assessment supports) would serve as the basis for interim performance tasks and as a context for summative assessment. These units should be designed in an online environment to capitalize on potential for innovation and supplementary training materials for teachers and supports for students. In addition to training materials for teachers on curricular implementation, these units should includes training materials and supports for implementing formative assessment and progress monitoring strategies within each unit. These units should be deep and flexible enough to use each year with new/comparable contexts such as a different science experiment or grade-level text, while assessing same concepts (e.g., standards).

The units could and probably should differ in depth and scope depending on grade level. I would expect high school units to be designed to fit within specific courses, while elementary units would tend to be a bit more generic. Further, at elementary and perhaps even at middle school, units and associated performance tasks that integrate content from multiple subject areas could be designed as a way to address legitimate concerns about the narrowing of the curriculum. I know there is a great interest in having teachers involved in the assessment system. This is an area where teachers working alongside skills facilitators/curriculum experts could make great contributions toward the system by helping to design these units.

Summative Assessment

The summative assessments should serve as a culminating experience at the end of each course or year in school. These should be administered toward the end of the school year when teachers have had as much time as possible to ensure that students have learned the expected knowledge and skills. However, instead of measuring students' ability to recall trivial information from earlier in the year, these summative tests should be designed to determine how well students are able to synthesize and use key concepts taught throughout the year. To the extent possible, these assessments should be designed using computer-based approaches to allow for the use of innovative item types and logistical efficiencies. The summative assessment would serve as the foundation for growth measurement although the summative assessment would NOT be the only contributor to school accountability.

Some of the content and specific examples used on the summative assessment will be drawn from the curricular units to help move past some of the superficial aspect of current summative assessments. This will allow for a richer representation of knowledge and skills (i.e., plenty of open-ended tasks) than is currently the case and will serve as a signal to educators of what is valued. While these assessments should, depending on capacity, be designed for online administration, moving to an online environment should not be done to feed the obsession with instant results. We still do not have appropriate automated scoring routines for content-based complex performance assessments and to the extent that rich open-response tasks are included in the summative assessments—I'd argue that these types of tasks should be included—we are still many years away from instant results. Further, I argue that the drive for instant results for end of year summative assessments has been driven by misconstrued accountability demands. We can certainly have very rapid turnaround of results, but everything comes at a cost.

Interim performance tasks

These rich and engaging tasks are the foundation of this proposed system. There is simply not enough time and the context is not appropriate to administer these rich tasks at the end of the year. That is not to say that at least a one or two rich tasks couldn't be included in the summative assessment, but the main focus of these experiences should be in the course of the school year. These tasks will be contextualized within the curricular units. In fact, these extended tasks would be the culminating experience of each of the units.

These tasks should be scored locally and incorporated within local assessment and grading (graduation) systems. Again, this is another way to include teachers and other educators in the full assessment system. If these tasks are to be included in the full school accountability system—which I argue they should—the local scoring can be audited to verify the accuracy and consistency so that these scores can be included fairly in the state accountability system. While this is my vision, I think it is fine to leave such accountability uses up to each state. During the early 1990s, this type of auditing (and providing feedback to schools/teachers) of local scoring of writing portfolios was one of the most effective ways to ensure internalization of the writing scoring criteria.

As discussed earlier, the tasks should be carefully designed using ECD or other legitimate theoretically-based principles to reveal students' need for additional support. While we should strive to capitalize on technology to design and administer these tasks, I would not require this at

this time, but would ask bidders for a plan for moving to an online environment. Because of the memorable nature of these tasks, many should be released each year, although a good portion should be held in a bank to be used again in the future.

Formative assessment

The curricular units and associated materials should be designed to facilitate formative assessment probes and processes. However, even the best materials are rarely enough to ensure that significant percentages of teachers adopt meaningful formative assessment practices. Therefore, any proposal should clearly explain how the consortium will structure and support professional development to increase teachers' capacity for implementing and using formative assessment to improve instruction. This will likely be part of many bidders' theory of action, but without clearly specifying the mechanisms to achieve widespread adoption of formative assessment strategies, formative assessment will just be a nice phrase in the proposals. Finally, while I have indicated a willingness to use interim tasks as part of local and state accountability systems, I argue to a firewall between formative assessment and accountability otherwise the purposes and intentions of formative assessment will become too easily corrupted.

Opportunity, Access, and Equity

I argue that we have much more of an instruction (OTL) than an assessment problem. The best assessment system cannot make up for lack of OTL. The proposed curricular units are designed to help level the curriculum and instruction playing field by providing supports for teachers to help them ensure that all students can access and learn the required knowledge and skills. Additionally, formative and interim assessments are included in the system to help build educator capacity and to help "catch" students before they fall so too behind.

The tasks used throughout my proposed system should be designed with multiple and varied opportunities for students to validly participate in the assessment system. We need to capitalize on tremendous advances in innovative technological approaches for access and accommodations, such as those offered by Nimble Tools, to help promote access and opportunity. Finally, assessment guidelines need to focus first on fair access and less on narrow definitions of comparability.

A "New" Psychometrics

Related my point about comparability above, a system such as the one I'm proposing will require some serious re-examination of our current psychometric practices. We've traded a lot (of validity) in the past for student-level reliability, smooth scales, and overly strict notions of comparability. There is no question that some aspects of the system that I am proposed will create serious equating challenges, but I have confidence that we can address these especially if we design more valid accountability systems than the current approaches. The foundations for "new" approaches have been established (e.g., Linn, Baker, Dunbar, 1991, Mislevy 1994, Pellegrino, et al, 2001), but still need more attention to work in large-scale, efficient practice. The NIA should push for requirements and expectations beyond the current "safe" technical methods.

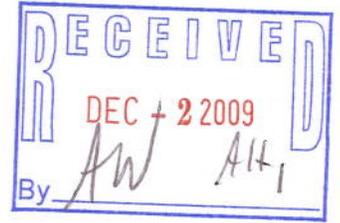
High Schools

The proposed assessment system should be situated in specific “indicator” or core courses up to some point (e.g., 10th grade), after which there should be more choice in the assessment (and accountability) system to allow for specialization and choice by students. The interim performance tasks can be used as part of a student accountability system like Wyoming’s or Rhode Island’s graduation systems. The bottom line, however, is that high school should not be seen as just an older group of elementary school students. If the goal is truly to create college/work ready students, then we need to focus on providing much more meaningful experiences in high school than has been the case. I argue, not surprisingly, that my proposed curricular units approach can enrich high school experiences for students all along the achievement continuum.

Some advice on the NIA

I do not envy the position of the USED in crafting this NIA and making awards. I offer the following suggestions—in no particular order—in hopes that they are helpful in your work.

- ✓ Be ambitious and innovative. Think 7-10 years out, as Randy Bennett noted, and craft this NIA to ensure that the awards put us on the path to the 7-10 year vision.
- ✓ Recognize that simply changing existing state assessment systems to measure new standards—without doing anything else differently!—will require at least three years from the date of the award. I cannot see how doing anything innovative and ambitious can take less than five years and more likely ten!
- ✓ Be crystal clear about your prioritized goals for the system.
- ✓ Do NOT specify the means unless you are absolutely certain of what you want. Let the bidders do much of the creative thinking to help you realize your crystal clear goals.
- ✓ Determine the absolutely essential pieces and then examine costs for additional components.
- ✓ Absolutely allow for multiple awards because as Laurie Wise noted, if you only make one award, you will be forced (politically) to be conservative.
- ✓ While at the Boston meeting, I argued that consortium should encompass the full K-12 system (or at least 3-12) for a given content area, I have been persuaded by Lorrie Shepard’s suggestion of allowing for narrower foci in order to increase the chances of successful innovation.
- ✓ Development is an ONGOING cost, not a one-time purchase!
- ✓ Recognize and embrace the differences between high schools and elementary schools.
- ✓ Reconsider the current practice of having every student tested on every item
 - Matrix sampling is still a viable approach
- ✓ I would definitely require bidders to include a well-specified theory of action that clearly spells out the goals and processes of the proposed system and to provide evidence that justifies their expectations.
- ✓ Further, I would require an essentially companion theory of action or work plan that describes how the consortium organization will support the theory of action of the curriculum/assessment system. This is not a piece to be taken lightly. The best ideas and plans will fall short without an appropriate organizational structure to support it.
- ✓ Similarly, the NIA and subsequent awards must recognize the critical operational and bureaucratic constraints include, but not limited to existing contracts, state laws, and procurement rules.



Race to the Top Assessment Program

Notice of Public Meetings and Request for Input
Assessing English Language Learners (ELLs)

**National Education Association (NEA), submitted by Luis-Gustavo Martinez,
Senior Policy Analyst, Education Policy and Practice Department**

Good morning representatives of the U.S. Department of Education.

I am here today representing the National Education Association (NEA), the nation's largest professional employee organization. Our 3.2 million members work at every level of education—from pre-school to university graduate programs. Our interest is to provide recommendations to the Secretary, who is particularly interested in innovative and effective approaches to assessment that will assist States in creating powerful and useful systems of assessment for English Language Learners.

The NEA recommends the Department to:

- 1) Ensure that the unique factors that impact the performance of ELLs and ELLs with learning disabilities are specifically addressed in the assessments that are used to measure the academic achievement of these students and reporting of the results.
 - When developing assessments, consider the specific characteristics of ELLs, in conjunction with standards. Assessments must be sensitive to various forms of diversity, including cultural, both within and across subgroups such as ELLs and ELLs with learning disabilities. It cannot be assumed that assessment or accommodations developed or adapted for one subgroup will be effective and valid for other subgroups. For example, the issues to be addressed in assessments and accommodations for ELLs and ELLs with learning disabilities are not the same.
 - Align and integrate standards and assessments that are specifically crafted for ELLs into the overall assessment system.
 - Incorporate available research, evidence and principles of fairness and equity for ELLs into assessment systems. (For example, use results from empirical research to indicate when ELLs may be tested in English on content-based assessments based on their level of English language proficiency.)
 - Provide the opportunities and resources necessary to ensure that ELLs have meaningful access to the content that is based on state standards.
 - Require multiple forms of evidence in the assessment of ELLs, including results of classroom-based assessments and performance of ELLs in their native language and/or in English, consistent with the language(s) in which they receive instruction or are best able to demonstrate their learning.
 - Understand the diversity within the ELL student population (such as linguistic and cultural differences; continuity of educational experiences inside and outside the U.S.) and act accordingly.

- 2) Require states to provide research-based recommendations for selecting and using appropriate accommodations for ELLs to ensure that these students have access to valid assessments of their content knowledge.
 - While the principles of universal design should be applied to the assessment system for ELLs with learning disabilities, base selection of assessments or accommodations on the specific needs of the students being tested.
 - Provide specific guidance for selection of assessments and/or accommodations for students with dual classifications (e.g., twice exceptional: ELLs with reading disabilities).
- 3) Require states to validate assessment systems for ELLs.
 - Include large enough numbers (95%) of ELL students in the validation process.
 - Control factors that negatively impact assessment outcomes for ELLs so that variables that are not the primary interest in assessments of achievement do not affect assessment results. (For example, a test in English is a test of English for ELLs; therefore, English language proficiency may affect students' ability to demonstrate their academic achievement in English.)
 - Require that states develop accountability systems that incorporate both growth and status measures. For example, emphasize growth when students are acquiring English language proficiency since language is a developmental process, and then shift the emphasis to a mix of status and growth when students have achieved the necessary proficiency (as determined through validation studies) to learn academic content taught entirely in English.
- 4) Support research to address major issues that complicate the design of appropriate assessment systems for ELLs. These include:
 - A universal definition for ELLs;
 - Appropriate identification of ELLs;
 - Psychometric properties of English language proficiency assessments;
 - Psychometric properties of both native language and English academic achievement assessments;
 - Psychometric properties of assessments for alternate assessments of academic achievement for ELLs, ELLs with disabilities, and non-ELLs with disabilities;
 - Accommodations and modifications for ELLs;
 - Criteria used for ELL student participation in testing, the effects of arbitrary criteria applied to subgroups (e.g., percentage of students who can be exempted, limits on the number of times students can take native language achievement tests, or specification of when ELL students have to be tested in English), and reasonable percentages of students for whom various alternatives, modifications or accommodations should be available;
 - Comparability of native language assessments, alternate assessments and regular assessments; and

- Language domains tested in Title I as compared with those tested in Title III for ELL students.
- 5) Provide incentives for states to work together to shape the conceptual design and construction of local and state assessments of academic achievement according to the characteristics of each specified subgroup. Federally fund research to address the most pressing technical issues related to assessments and accountability decisions for ELLs.
- Provide incentives and technical assistance for states to improve current assessments, apply universal design, and use multiple measures and growth models, as applicable to ELLs and ELLs with learning disabilities.
 - Examine the extent of applicability of the principles of universal design to the design of state assessments.
 - Establish criteria for ensuring that state assessments are relevant, equitable, valid, and of high quality for all ELLs and ELLs with learning disabilities.
 - Develop subject area assessments in languages other than English when students speaking a language form a significant proportion of the population.
 - Research the validity of common district assessments designed for ELLs and ELLs with learning disabilities as part of the state's comprehensive assessment system.

From: Mulattieri, Karen [kmulattieri@cicd99.edu]
Sent: Wednesday, December 02, 2009 12:49 AM
To: Race To The Top Assessment Input
Subject: Race to the top Assessment Program

Input Submission –December 1, 2009

ELL Trend Data –case study addresses the impact of high stakes English academic achievement testing on ELL students- actual progress over time

Race To The Top Comments-

1. Addresses the need for research in assessment practice to assist ELLs in demonstrating what they know and can do
2. Addresses effective programs for ELLs and need for quality assessments in languages other than English
3. Addresses the international benchmark of proficiency in more than one language as part of education
4. Addresses the need to make progress one of the key indicators in accountability for ELLs

Karen Mulattieri

Previous experience with statewide assessments:

Work with WIDA Consortium on English Language Proficiency standards and development of ACCESS language proficiency assessment

Member of the Illinois English Language Learner Assessment Advisory Committee for ten year

Presentation on ELLS and RI at Statewide Bilingual conferences 2007 and 2009

Karen Mulattieri

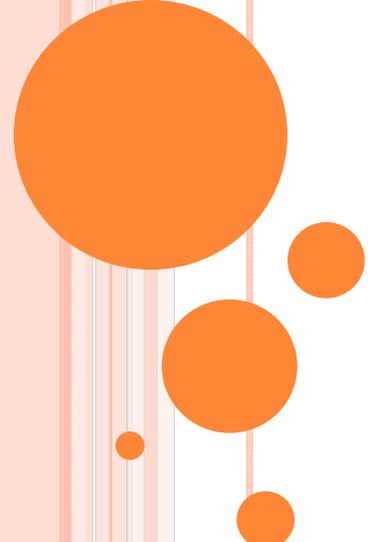
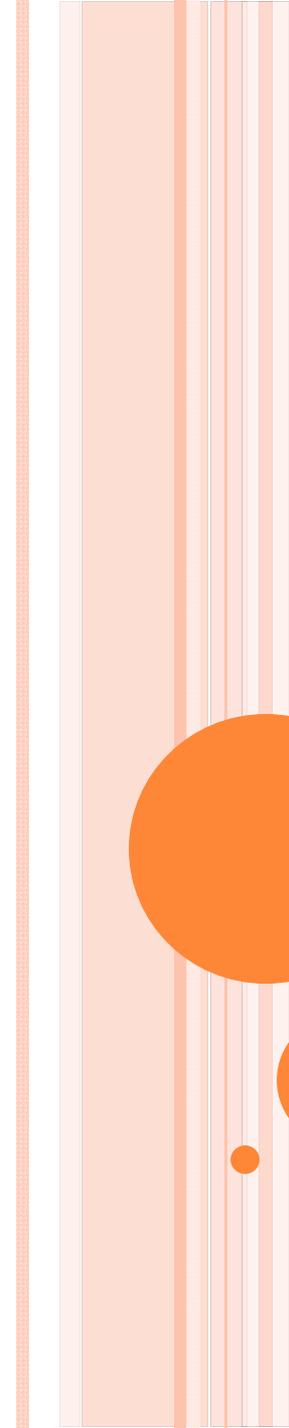
Assistant Superintendent for Student Services

Cicero Public Schools

5110 W. 24th Street

Cicero, IL 60804

708-863-4856



ELL DATA

Cicero Public Schools- Case Study

Karen Mulattieri, Assistant Superintendent

Dec. 2, 2009

KEY QUESTIONS

- How should we define a “successful school”?
- What quality indicators should be considered?
- How can we track progress?
- What if student groups take longer to eventually meet benchmarks?
- What value do we give to proficiency in two languages?



MORE KEY QUESTIONS

- Does meeting/exceeding on state assessments guarantee high school graduation?
- How do we measure bilingualism?
- How do we measure higher order thinking and collaboration?
- Which school experiences promote students' taking responsibility for their own learning?



SCHOOL DISTRICT DEMOGRAPHICS

- 94.6% Hispanic students
- Enrollment: 13,800
- 52% English language learners
- 84.7% low income students
- All schools are school wide Title I
- School Choice, SES, Title I Extended Day, and Summer School in place for the last three years



TESTING IN ILLINOIS

- All state assessments are administered in English
- ACCESS is the Language Proficiency Measure used across 21 states
- In 2008 ISAT became the State Assessment administered to all students including ELLS-new baseline was not created.



RESEARCH ON ELLS AND SECOND LANGUAGE ACQUISITION

- The primary language, developed in the context of social interaction, is fundamental to the thinking, learning, and identity of the individual.(Commins;1997)
- Students first and second languages interact with each other (Cummins 1979; Oller;1980)



RESEARCH ON ELLS AND SECOND LANGUAGE ACQUISITION

- Academic English takes 5-7 years to develop (Collier and Thomas, 2005)
- Meta-analysis show that existing evidence favors bilingual approaches over English-only immersion type programs (e.g., Slavin & Cheung, 2003)
- Eventually, able to close the gap between native English speakers' performance over time (Collier 2001, Ramirez et al, 1991)



STATE ENGLISH LANGUAGE PROFICIENCY DATA ON ELLS (ONLY 21.4 % HAVE RECEIVED 5-7 YEARS OF ESL)

Year	Percent Making Progress	Percent English Proficient
2006	78.9%	19.1%
2007	94.8%	32.4%
2008	92.7%	26.9%
2009	91.1%	25.7%



READING FIRST DATA 2009

K-3 STUDENTS ASSESSED IN ENGLISH LITERACY

Grade Level	DIBELS /Fall Students at Benchmark	DIBELS/Spring Students at Benchmark
Kindergarten	19%	31%
First Grade	30%	36%
Second Grade	29%	32%
Third Grade	31%	36%



STATE ASSESSMENT-READING ACHIEVEMENT

Grade Level Cohort	2008 Data- Meet/Exceed Standards	2009 Data- Meet/Exceed Standards
Grade 4	44	47
Grade 5	48	49
Grade 6	38	59
Grade 7	60	56
Grade 8	62	70



ACHIEVEMENT IN ENGLISH VS. ENGLISH PROFICIENCY

○ 2009:

- 25.7% were proficient in English – able to take an English achievement assessment
- 21.4% had 5-7 years of formal English instruction
- 44.1 % of ELLs were able to meet/exceed state standards in English Reading given the same assessment

Are we truly assessing what students know and can do?



ASSESSMENT IN THIS CONTEXT

- The District administers local benchmark assessments 3 times a year in both English and Spanish
- Students are making progress according to online assessments
- Teachers are using curriculum and instruction that is aligned to standards.
- English literacy is the area where students are not making AYP, this in turn affects mathematics.



UNIVERSITY PARTNERSHIPS

- Northern Illinois University-teachers earning masters' in reading
- 21st Century-Northeastern University
- University of Illinois at Chicago-joint grant proposal for grades 6-9 Mathematics



QUESTIONS?



Race To The Top

Assessment of English Language Learners

Input on Assessment of English Language Learners

Karen Mulattieri, Assistant Superintendent for Cicero Public Schools

Background:

Under NCLB, English language learners in grades K-12 are to be assessed annually to measure English language proficiency in listening, speaking, reading, and writing. The law also requires that ELLs be held accountable for academic achievement in the areas of reading, mathematics and science. This is the only group of students who are subject to two assessments annually.

The original wording of the law requires that the annual measure of academic achievement must be aligned to standards and designed to be valid measures of what ELL students know and can do academically.

Current situation in Illinois:

ELLs take the ACCESS measure of English language proficiency. In most cases cohorts of students demonstrate progress in English from year to year. In a statewide study the scores of students on both assessments were cross referenced to determine the point where ELLs could perform academically commensurate with their English Speaking peers. The ACCESS scores in necessary in at most grade levels, was determined to be 4.8 composite with a 4.2 in literacy. Statewide a small percentage of students meet this benchmark.

Academic Assessments:

In fall 2008, the U.S. Department of Education in conjunction with the Illinois State Board of Education determined that the academic measure to be used with ELLs in grades 3-8 would be the same assessment used with English speaking students (ISAT). There have been two administrations of ISAT to date; the result has been more schools failing to make AYP. Districts with large groups of ELLs are particularly judged in a manner that is not fair.

Best Practice for instruction for ELLs:

Research and many educational associations such as the International Reading Association, recognize the crucial role that the primary language plays in instruction of ELLs as they gain English proficiency. Students need to be able to understand complex academic concepts in order to progress. Use of the native language especially in literacy instruction is key in instruction in the core academic areas.

Dual language or two way immersion programs are proven to reduce the achievement between ELLs and their English speaking peers. Quality assessments in languages other than English can best inform instruction in these programs.

Best Practice in assessment-

Assessments of academic achievement should match instruction. In the case of Illinois, native language instruction is mandated. Bilingual and dual language programs rely upon native language instruction in core academic subjects.

Currently, all accountability measures in grades K-8 are administered in English with one accommodation allowing students to construct responses in Spanish.

Goals for Race to the Top:

One of the goals of race to the top is to track college readiness. All ELLs enter school speaking a language other than English. In the current accountability system, the focus has been on the acquisition of English. Fluency in a language other than English is an entrance requirement for many colleges and universities.

International benchmarks are also mentioned in the rules for the grant applications. Most countries around the world promote fluency in more than one language. In the U.S. we have a history of promoting fluency in English.

Recommendations

1. The USDE needs to consider fluency in more than one language as goal for all students.
2. In Race to the Top-valid ways of assessing ELLs academic content knowledge need to be researched and perfected. Online assessments with the use of graphics and performance tasks that reduce the language lead hold promise.
3. Learning progressions for English acquisition need to be formulated.
4. Formative assessments in the languages other than English, particularly in literacy need to be considered as part of a comprehensive assessment system for ELLs and students in dual language instruction.
5. Accountability systems for ELL need to make progress a key component of the equation until English proficiency is reached.

Most ELLs are also of low socioeconomic status and have a history of dropping out of secondary school. With the proper instruction, assessment and support, this trend can be reversed, which I believe is one of the major goals of the "Race to the Top" initiative.

From: Anna Nicotera [Anna@publiccharters.org]
Sent: Wednesday, December 02, 2009 3:36 PM
To: Race To The Top Assessment Input
Cc: Brooks Garber
Subject: Race to the Top Assessment Program

Anna Nicotera
National Alliance for Public Charter Schools Input on Race to the Top Assessment Program
General Assessment Input

In response to the third question in the Project Management section in the Rate to the Top Assessment Program notice, the National Alliance for Public Charter Schools recommends that state consortia be required to include representatives from the charter school authorizers and representatives from the states' charter schools in the development and implementation of the proposed assessment systems. Charter school authorizers retain the responsibility of approving new charter schools and determining charter school renewal and closure. Valid, reliable, and timely performance assessment outcomes are critical to making these decisions. As a result, charter school authorizers are important stakeholders who must have a seat at the table when the new systems of assessments are designed and implemented.

Anna Nicotera
Research and Evaluation Director
National Alliance for Public Charter Schools
(303) 333-4325 (w) / (303) 257-0884 (c)
anna@publiccharters.org
<http://www.publiccharters.org>



**Educational Testing Service
Response to Request for Input on the
Race to the Top Assessment Program**

December 2, 2009

Prepared for:

U.S. Department of Education
Office of Elementary and Secondary Education
400 Maryland Avenue, SW, Room 3E108
Washington, DC 20202

Executive Summary/Introduction

To ensure that your input is fully considered, we urge you to identify clearly the specific question, purpose, and characteristic that each of your suggestions addresses and to arrange your submission in the order of the questions listed later in this notice. Please also include a description of your involvement, if any, in statewide assessment practices.

Educational Testing Service (ETS) respectfully submits this response.

Description of Involvement in Statewide Assessment Practices

ETS has been involved in K-12 assessment for decades. At the federal level, we have held contracts since 1984 to develop, administer, and report the National Assessment of Educational Progress (NAEP). Under contract to the College Board we also develop national assessments that play important roles in K-12 education. These include the *Advanced Placement Program*® (AP®), the SAT®, and the *Preliminary SAT/National Merit Scholarship Qualifying Test* (PSAT/NMSQT®). ETS develops national-level assessments for other clients, most notably for the Educational Records Bureau and the Southern Regional Education Board. Moreover, at the state level, we are either developing or have developed statewide assessments for California, Florida, Georgia, Indiana, Maryland, Mississippi, New Jersey, Puerto Rico, Tennessee, Texas, Virginia, and Washington. In California, we also work with the California State University system to develop the Early Assessment Program (EAP), a college readiness supplement to end-of-course components of that state's assessment system.

ETS provides a range of services to states and other clients. These include psychometric research and statistical analysis, assessment development, program management, production and delivery, communications, and policy analysis. The types of assessments for which we contract with states include No Child Left Behind (NCLB) summative assessments, along with their alternate and modified versions; high school end-of-course assessments; high school exit examinations; and Title 3 English language proficiency tests.

In the pages that follow, we provide one set of possible answers to the questions raised in the U.S. Department of Education's (the Department's) Notice of Public Meetings and Request for Input on the Race to the Top assessment program. Before beginning a direct response to these questions, we offer a brief introduction.

Introduction

Advances in technology, coupled with innovative assessment task design and advanced psychometric and cognitive models, make it possible for us to obtain a richer, more intelligent, and more nuanced picture of what students know and can do than ever before. While the historic opportunity to change the direction of education is real, so too are the challenges inherent in any change in assessment paradigm. At the heart of these challenges is one point that is often missed: Different stakeholders will set diverse priorities for an assessment system. Some of these stakeholders value snapshots of what

students know and can do at fixed points in time, and they consider the use of these data for accountability purposes as the highest priority. Others value obtaining multiple points of data that can be used to evaluate schools and teachers systemically. For some, instructionally actionable data at the student level for the purpose of improved instruction is the main system goal, while others are more interested in data at higher systems levels for auditing or “return on investment” types of decisions. Most want formal assessments to be as short and inexpensive as possible, while others would trade some cost and time efficiency to have more authentic, complex, and reliable tasks. Some stakeholders require data that are unambiguously comparable across states, local education agencies (LEAs)/districts, schools, and children, while others would rather see some substantial state and local control over the content of assessments.

No single assessment, not even an integrated-assessment system, can optimally serve all possible purposes. Any assessment design, therefore, is a compromise. Tests that provide optimal instructional feedback may not be the best way to get an overall snapshot of what students have learned over the course of a school year. The need for formative information is not necessarily consistent with the need for data that can be used to evaluate teacher or school effectiveness. Tasks that model good instruction are not always consistent with desires for tests to be as short as possible and for scores to be returned immediately. The desire for comparability of data across jurisdictions conflicts with wishes to allow those jurisdictions — and their teachers and curriculum specialists — substantial and variable input into the form and content of assessments. The need for low operational cost may be at odds with many other goals of the system. Efficiency in the long term involves investments in technology and human capital in the short term.

Policymakers should consider the four principles following from this discussion:

- » First, we should think of systems of assessments rather than individual tests, as this approach is likely the only way to satisfy the various information needs identified by stakeholders.
- » Second, we are at a moment when new technologies and assessment methodologies provide us with an unprecedented opportunity to satisfy many perceived needs in a carefully structured integrated system.
- » Third, we must realize that, even in a complex system, we will need to choose among competing and conflicting priorities.
- » Fourth, we must stage the creation of the new assessment system to accommodate reality, because even if we know what we want to do and how we want to do it, the existing assessment infrastructure in the U.S. is a limiting factor in implementation.

This document represents an attempt to create a high-level framework for an assessment of common-core standards. We arrived at this framework in the following way: First, we considered the general requirements and desired characteristics of such an assessment system. Then we considered various factors and made judgments about competing priorities. Different decisions about priorities would certainly result in different assessment designs, and we endeavored to point out places where alternate decisions might have such impact. Ultimately, some areas require further research and more thought.

Finally, we defined the desirable system as “Generation 2” and recommended a “Generation 1” transition system to achieve many of the goals of the ideal system sooner than would be possible if we waited for all elements of Generation 2 to be feasible. Because of all these considerations, it is important for readers to understand that this document is only one of a broader set of possible answers, and is meant to inform the Department’s thinking rather than to propose a single path forward.

General Assessment Questions

- 1)** Propose an assessment system (that is, a series of one or more assessments) that you would recommend and that meets the general requirements and required characteristics described in the notice. Describe how this assessment system would address the tensions or tradeoffs in meeting all of the general requirements and required characteristics. Describe the strengths and limitations of your recommended system, including the extent to which it is able to validly meet each of the requirements described in the notice. Where possible, provide specific illustrative examples.

We believe the following to be key design elements of a forward-looking assessment system:

1. The educational system needs both accountability and instructionally actionable data, and no single test will be optimal to provide both. Therefore, we believe that the goals of this new effort will be best served by an integrated-assessment system that includes summative and formative or interim elements built to a common framework. If the American Recovery and Reinvestment Act of 2009 (ARRA) funds support only the development of the summative elements of the system, the Department should ensure that the system and system infrastructure are designed to work with formative and interim elements that are designed and developed by others.
2. The system must measure common standards and must allow for state-to-state comparability on the common standards. To accomplish this, the new summative measures should have a set of common components assessing the common standards, and produce scores and performance indicators that are comparable across states. However, the system should also allow states to augment this core with materials of their choosing to produce separate state-specific information.
3. The summative portions of this battery will need to include, at a minimum, end-of-year assessments for grades 3 through 8 in both mathematics and reading/language arts. At high school, the system may include either “end-of-course” or “end-of-domain” assessments. The elementary- and middle-school assessments should support growth modeling and across-grade comparability. The assessments should also support within-grade proficiency standards. While we believe that these end-of-year and end-of-course/domain assessments should be part of the system, we also believe we should consider using data collected over the course of the year as part of the summative system (see point 9 below).
4. Assessment designers will likely need to incorporate international benchmarking and facilitate comprehensive alignment efforts, although the methods for accomplishing these goals have not yet been determined.
5. The tests should be delivered on computer or other similar technology. Student mastery of emerging standards can likely not be measured based on paper assessments alone. Further, summative assessments should make use of adaptive administration, although adaptive models will need to make allowances for the full range of item types needed to measure emerging

constructs, including those that will be scored by humans. We envision that such a system will ultimately support the on-demand needs of a personalized education system. However, the technology to effectively administer computer-adaptive tests on a large scale in a narrow summative assessment window is not available yet in many states. Therefore, we may need to consider the possibility that while complete technology delivery is a goal for the Generation 2 assessment system, transition to these technologies may need to be staged over the period of Generation 1 implementation.

6. The development of assessment tasks will be based on an evidence-centered design (ECD) process that involves experts and stakeholders. To measure the intended constructs, the tests will likely need to use a range of tasks and stimulus materials, and will need to include more than traditional multiple-choice questions. Important decisions will need to be made regarding how constructed response questions are scored, though we picture a mixed model that uses technology and professional (e.g., teachers and other subject matter experts) scoring that is supported by assessment technology infrastructure. Such a system will also provide opportunities for professional development.
7. Compared to current summative tests, items and tasks should be created based on an improved understanding of learning and development, both to promote better interaction with formative elements of the system as well as to provide models consistent with good instruction.
8. Tests should be as accessible as possible to students with disabilities and English language learners, and designers should make use of technology to improve such accessibility.
9. Certain forward-looking ideas should be considered that may or may not be ready for operational implementation at the time of initial rollout of the new system. Perhaps most important among these considerations is that summative assessments may not be single-testing events but could augment end-of-year assessments with data collected over the course of the year. The use of interim elements as part of a summative system could also provide ways to experiment with the use of new item types and technologies.
10. We should have careful plans in place to validate assessment scores and claims made based on them, as well as a long-term research agenda to continuously improve the efficacy of the assessment system for its intended purposes.

2) For each assessment proposed in response to question 1), describe the—

- Optimal design, including--
 - Type (e.g., norm-referenced, criterion-referenced, adaptive, other);
 - Frequency, length, and timing of assessment administrations (including a consideration of the value of student, teacher, and administrative time);
 - Format, item-type specifications (including the pros and cons of using different types of items for different purposes), and mode of administration;
 - Whether and how the above answers might differ for different grade levels and content areas;
- Administration, scoring, and interpretation of any open-ended item types, including methods for ensuring consistency in teacher scoring;
- Approach to releasing assessment items during each assessment cycle in order to ensure public access to the assessment questions; and
- Technology and other resources needed to develop, administer, and score the assessments, and/or report results.

- Optimal Design

We believe the summative assessments should have two major components: a common-core assessment and an optional state-specific assessment. Our understanding is that states may augment the common-core standards with their own standards, as long as the common-core standards represent at least 85 percent of the universe of standards in the state at any grade where common-core standards exist. Thus, the common-core assessment system must provide data on the common standards that are strictly comparable across states and must allow states to measure state-specific content as needed.

Because there could be both common-core standards and state additions, the tests would likely have at least two major components. The first would be the test of common-core standards. This would be consistent across all participating states, LEAs, and schools. The common components of the test will be designed to yield state, LEA, school, and individual results on the common-core standards and will not include state-specific augmentation. The second component could be composed of state-specific content or augmentations. Such augmentations could focus solely on the unique state-specific standards that are in place or provide additional measures or coverage of common-core standards. These augmentations would be analyzed in tandem with common-core items to yield state-specific results.

Why do we believe that the common-standards components of the summative measure should not be customizable, and that state choices should be located in state-specific sections? Comparability of results on the common-core standards and test development efficiency will be high priorities of the system. Comparability across states and the economies of scale will be enhanced if there is a common assessment of the common standards. Other designs are possible if the ability of states to customize the common-core assessment is viewed as desirable, but these will likely threaten comparability of results and will lead to higher cost.

In system terms, the approach we recommend means adopting a single national delivery package and permitting states (or groups of states) to add components as needed, as opposed to “opening up” the common materials for each state. Note that we do not mean that the same exact test form is required, but rather the same assessment, which would be available in equivalent (or adaptive) forms consistent with test security.

This approach allows some states to decide they do not need state-specific content, without affecting the comparisons on the common components (which embedding items in the common core would risk).

This approach has other advantages: Even if a single consortium develops the common-core assessments, states would be free to work with whomever they wished for state-specific components. If developers of the common-core components of the system were to work toward some open and shared standards for test material, packaging, and delivery, all components could be delivered as a single test by any number of assessment-delivery systems. (We comment on this further in the response to bullet 4 under this section [Technology and other resources needed to develop, administer, and score the assessments, and/or report results].) Alternately, the developers of the common-core assessment could build some special components that could be used at state discretion.

Note that in any of these models, provisions will need to be made for pilot/field testing new content. For the common components, this could be accomplished either through a variable section or by embedding pilot/field test items within operational sections.

One open question is how big a system (in terms of assessment exercises) would be needed to maximize security. The answer will depend on the length of the test window, which in turn depends on the number of students who can be tested at any time. This answer also will be affected by the speed with which test developers can rotate content, or the number of different aggregations of content we can provide.

A second open question concerns the length of the individual tests. It is likely that tests at grades 3 and 4 will be limited to 50 minutes, while tests at grades 5 through 8 will take 60 to 120 minutes (for both common and state-specific components). High school tests could, conceivably, take between 2 and 3 hours. If extended tasks are used, assessment time may need to exceed these limits.

– Type (e.g., norm-referenced, criterion-referenced, adaptive, other)

The system must support both common and state-specific performance levels. A comprehensive system might work as follows: There could be a single-scale score and a set of achievement levels on the common test component. This would allow for comparisons among participating states and the placement of individual scores in the context of the common standards. Recall that this is possible because each state in a consortium is taking the same assessment on the same standards.

The common-core standards assessments will likely need to be internationally benchmarked. The easiest way to accomplish this is through judgmental processes: either through the use of the internationally benchmarked standards as key descriptors of goals in a level-setting process, or through

some assurance from an independent body that the standards themselves conform to international best practice and that the assessment is aligned with the standards. Alternately, the system could rely on statistical linkages to international studies such as Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS). Regardless, a key step involves meeting with stakeholders to determine the specific uses stakeholders wish to make of the international benchmarks.

Our proposed design assumes that the new assessments will have performance standards. Therefore, using appropriate methods and sources of information to set standards will be of key import. Standard setting is often not considered when designing an assessment, but the validity of claims made based on the assessment will be no stronger than the performance standards allow. Assessment designers should help ensure that crucial evidence is brought to bear regarding topics such as what successful students around the world know and can do in different grades, and what sorts of texts students should be prepared to encounter to succeed at the next grade. Overall, we should have a solid evidentiary basis for stating that students have reached a level that will allow them to succeed in future education.

The comments above relate to the scale and performance levels for the common-core components of the summative assessment. In addition to this, the assessment will need to have separate state-specific scales and levels for states that augment the common core with their own materials. In all likelihood, these would be based on state-by-state analyses of the conjoined sets of items (that is, common plus state-specific). In practical terms, states may find it challenging to explain major differences between their standards and national standards. But the system needs to support these types of data.

Use of Adaptive Testing

As mentioned above, we believe that the summative-assessment system should make use of adaptive administration in Generation 2. Whether or not all elements of the system can use adaptive administration in Generation 1 is yet to be determined. The answer will depend on the type of adaptive models we wish to use, and the availability of technology for universal computer administration.

A variety of adaptive testing approaches may be used when the assessment system reaches maturity (e.g., traditional item-level adaptive testing, multistage testing, variable or fixed-length testing). The appropriate adaptive testing solution will depend on the content and structure of the exams.

Some arguments in support of adaptive testing follow:

- » It allows for on-demand testing.
- » It allows for somewhat shorter testing times than linear testing, which helps from various perspectives, particularly if access to computers is an issue.
- » It allows us to measure the “higher” standards, while at the same time gaining some meaningful information about what lower performers know and can do.
- » Considered appropriately, it may allow us to identify standards on which students are struggling without unduly lengthening tests. Particularly in reading/language arts with a heavy emphasis on authentic reading, we believe variations in traditional computer-

adaptive testing approaches (e.g., section-based or passage-based adaptivity) can be implemented in an advantageous manner. Again, this will allow for far more personalization than traditional assessments.

- » It will allow us to get better return on the investment in open-ended/performance-based testing.

One possible challenge is the use of items that require human scoring in an adaptive system. There are in fact ways to use such items. In a multistage system, for example, routing decisions can be made based on a machine-scorable stage, with performance or open-ended exercises requiring human scoring administered during later stages.

While we believe the assessment should be adaptive, it is not certain the system could be adaptive in the first year of administration, even for the interim-assessment system. Large-scale piloting of items would be necessary before rollout. However, given issues associated with calibrating a pool under suboptimal motivational conditions, it is likely that the rollout year of the program would require assembling a large number of linear tests and assigning these randomly to candidates. The system could, however, use adaptive administration in subsequent years.

- Frequency, length, and timing of assessment administrations (including a consideration of the value of student, teacher, and administrative time)

In the previous sections, we have for the most part discussed the tests as if they were given at fixed points during some course of study (either the end of a school year or the end of high school). Furthermore, we believe that such end-of-year or end-of course tests should be part of any coherent system of assessments. However, this is not the same as arguing that they should be the *only components* of a summative system.

There are several ways in which one could consider other “assessment events” or data sources to be formalized parts of the summative-assessment system. In one family of approaches, there would be multiple assessments over the course of the year whose results would be aggregated into a summative score or scores. Such an approach could conceivably take one of two general forms. In the first, a larger assessment that would theoretically cover the entire year would be broken into component pieces covering different, and possibly non-overlapping, sets of content and skills. For example, a three-hour test might be broken into three one-hour tests that would be given over the course of the year. In this conception, the end-of-year assessment would essentially cover the last third of the year. A similar possibility is to build assessments around discrete instructional units (even if those were not equally spaced over the course of the year).

A variant on this approach is a system in which the end-of-year assessment did cover the entire year’s worth of content, but earlier standardized tests covered content from the first part of the school year in more depth. This is similar to the “midterm-final” approach used in many universities and high schools, in which scores from midterms and finals are averaged according to some preset weights and often combined with other information to derive a final grade.

There are obvious advantages to such approaches, as well as real challenges. On the positive side, one would get some early-warning data on students from the summative system itself; students might be able to retake modules they have failed over the course of the year. Because such systems would allow more aggregate data, they might give more stable results. On the other hand, such a system almost certainly involves making decisions about the ways content and skills are to be ordered (or at least combined) in the curriculum, and this may be beyond what is possible. While the aggregate data may be solid, the reliability of the periodic measures may be lower than one might like, which will be a problem if those data are used on their own for high-stakes purposes. Finally, in the second of these models, the system would need to be prepared to deal with a possible conundrum. If two LEAs got the same average scores on the end-of-year assessment, that phenomenon would normally be interpreted to mean that those two LEAs ended that school year “in the same place.” Rating one LEA higher because of performance on intermediate ratings might be problematic.

We describe an alternate model, used in some other countries, below. There would still be an end-of-year assessment, but accountability scores would also use data from standardized projects conducted over the period of the course of study (for example, research papers, laboratory reports, or book summaries). Scores from these projects would represent a fixed percentage of the final summative score.

This model would have clear advantages and disadvantages as well. By making these sorts of tasks part of a formal accountability system, this model encourages the use of tasks that are elements of good instruction and learning. In addition, this approach avoids the problem that usually keeps these sorts of tasks out of large-scale testing: They simply take too long to be included in a fixed-event assessment. These kinds of tasks might also provide a logical place to rely on teacher scoring and to enjoy the professional development benefits attendant upon it. Finally, centrally designed tasks and scoring guides may be able to mitigate certain comparability issues.

Our recommended transition to a new assessment system in two generations can allow for experimentation in these approaches without disrupting the utility of the accountability testing system. During Generation 1, the end-of-year or end-of-course assessment can be an “event” test, with the pressures of fast turnaround of results and the benefits of low cost, emphasizing or exclusively containing machine-scorable items. This event test can be supplemented with results from carefully controlled, but not necessarily identical, interim assessments that take place throughout the year, consisting of items in various formats; these assessments can be computer-adaptive. As we learn more and get technologies and operations in place to make the assessment system work more fluidly, we can advance the innovative item types and administration methodologies into the end-of-year assessment in Generation 2.

There are a number of issues that would need to be addressed in making such an interim-assessment system operational. It would need mechanisms to help ensure that students themselves completed the tasks. While steps might be taken to standardize task protocols and scoring rubrics, short of adoption of a common curriculum, some choice of tasks would need to be provided at the local level. Even with the best safeguards in the world, such choice, combined with local scoring, will almost certainly call into

question the strict comparability of results both over time and across jurisdictions. This is not a reason to reject such approaches, but rather represents the sorts of tradeoffs that must be considered and suggests the sort of research that is necessary. It may be possible to find interesting compromise positions: We might conceptualize an accountability system in which not all data elements are used for cross-jurisdiction comparisons, for example.

The use of assessments or projects conducted over the course of the year as part of a formal summative-assessment system is a concept deserving of thoughtful consideration. There are challenges to be met before such a system could be implemented, and the existence of such a system presupposes infrastructures for data maintenance and transfer that are currently beyond the scope of many states. Thus it is possible that these assessment features will begin as part of the state augmentations described above, until the time that they can be added to the accountability system. We believe that strong, forward-looking end-of-year assessments will be part of the system.

- Format, item-type specifications (including the pros and cons of using different types of items for different purposes), and mode of administration

Stating a firm position about item types is in many ways premature: Final internationally benchmarked standards do not exist at all grades. Decisions about the sorts and arrays of tasks that ought to be included on these assessments should be the result of a careful ECD process in which we gather expert groups, review research, and identify the sorts of behaviors that would convince us that students have reached the stated standards. Simply stated, we want to use the assessment task or item that most appropriately measures the construct desired.

However, we need working assumptions. Our task design should be guided by the general goal of measuring each construct as validly, effectively, and thoroughly as possible. This will certainly involve a range of exercise types that move well beyond traditional multiple-choice. These may include, though not be limited to, scenario-based tasks, long and short constructed responses, tasks that involve the exercise of technology skills, and simulations. This is particularly true given the general goals of providing college readiness information, eliciting more than content mastery information (i.e., problem solving and critical analysis), and exploiting the assessment medium (namely online technology).

During the design effort, other questions will emerge about the sorts of items and tasks that can be used. These will surround issues like the use of audiovisual stimuli (as called for in the CCSSO/NGA English-language arts standards), as well as interactive tasks involving spreadsheets and databases. One interesting matter that will need to be resolved early in the process concerns the inclusion of tasks that measure reading/language arts standards for speaking and listening (if these are in the final version of any set of standards). This is not uncommon in current state standards, but these skills are rarely if ever covered in assessments (which are normally limited to reading and writing). Decisions will need to be made about how to assess in these areas, as this has broad implications for test design and administration. One possible approach is to include listening and speaking in the individual score portions of high school tests (which can be longer), and only assess these skills at state discretion in tests

at earlier grades depending upon the goals of assessing listening and speaking or the outcome measures desired in these domains.

If we are to do something new and different, it is necessary that items and tests be developed with an awareness of how students learn. A test built around an understanding of available learning progressions is likely to be a better provider of information to formative components of the system. Items that model good learning and instruction should make “teaching to the test” less of a problem. Of course, this sort of thinking cannot mean that we fail to meet psychometric standards for quality, score comparability, and fairness, particularly given the high-stakes nature of the potential use for high school graduation, college readiness/college placement, and possibly college admissions. Finding the appropriate balance will be the key.

Our proposal to phase in the new system in two generations interacts with the question of item and task type. During Generation 1, the tasks and items that take longer to administer and score, or require computer administration, could be limited to the interim system, and the end-of-year assessment can rely only on the types of items and tasks that can be scored by computer, thus hastening the availability of results for accountability purposes. We have concerns that such limitations on the summative assessment would narrow curriculum and teaching, but these concerns are tempered by the fact that results from the interim assessment would also be part of the accountability system. In addition, we would be working toward a Generation 2 system in which these performance items/tasks would also be contained in the summative assessment event itself.

— Whether and how the above answers might differ for different grade levels and content areas

We assume that the summative-assessment system will include end-of-year reading/language arts and mathematics assessments at grades 3 through 8, all of which need to produce individual scores as well as aggregate scores and will need to work together to track student growth. These end-of-year assessments may not be the only components of the summative system. At high school, we believe two summative models are possible: either end-of-domain assessments in both reading/language arts and mathematics that cover the knowledge and skills needed to be ready for college and career training, or a series of end-of-course assessments. Each approach has advantages and disadvantages, depending on the priorities selected.

One should not assume that a single assessment model or design will make sense at all grades and in all subjects. For example, tests used at early grades will almost certainly be shorter than those used at the high school level. It is also likely the case that the types of exercises used may vary across grades and subjects, as may the mix of machine and human-scorable items. The amount of technology familiarity we can expect of test takers may also not be consistent across grades. It is even conceivable that the constructs covered at grades may vary: For example, a grade 4 reading/language arts test may focus on reading and writing skills, while a college readiness measure may also include measures of student abilities to listen to lectures. Certain underlying goals, like growth modeling, may be easier to achieve at elementary than at secondary levels. Finally, even if the “end state” is to have assessments that are

similar across ages and subjects, the transition plan may not be the same. We may be “more ready” to test grade 8 reading/language arts in a computer-based setting than we are to assess grade 4 mathematics.

We cannot determine the specific ways in which answers vary by age and grade until specific grade- and subject-specific standards are finalized.

- Administration, scoring, and interpretation of any open-ended item types, including methods for ensuring consistency in teacher scoring

To optimize the speed and cost-effectiveness of scoring a range of non-traditional items, we should be prepared to adopt a range of strategies. First, we may need to push the limits of what can be scored electronically: machine-scorable must not equal multiple-choice. Computerized-scoring systems are getting more effective all the time. Second, we can and should develop better ways to analyze data obtained from exercises such as simulations that go beyond simple student responses. Third, while some tasks can be machine-scorable, we must realize that emerging standards will likely necessitate the use of items that, given the current state of scoring technology, will require human scoring for some number of years. If this is true, we will have to find ways to balance the need for these items with other imperatives. We will also need to make effective use of technologies for distributing responses for scoring, and for monitoring and assuring the quality of such scoring. To summarize, we believe it is likely that the new assessment system will need to make use of three types of scoring: simple machine scoring using online testing, intelligent scoring using online technologies, and human scoring using online technologies.

Human scoring is, of course, in many ways a positive. It allows items that are not constrained by limits of the current electronic-scoring systems. The use of teachers in the scoring process would also represent a powerful professional development activity. Teacher scoring in a system that will also be used for teacher evaluation will necessitate careful safeguards. Therefore, any final design will need to find ways to use human-scorable items in ways that optimize the instructional and professional development impact of those items, without placing undue or unrealistic burdens on the system. We should also be prepared to make aggressive use of emerging computer constructed response scoring technologies, to make sure that teacher involvement is in fact professional development and not solely additional labor. We believe there are ways to involve teachers in scoring, without necessarily expecting them to conduct all the scoring (at least of the common-core standards components that require rapid score turnaround). The good news is that much progress has been made recently in using automation in human scoring in ways that improve quality and professional development potential.

Ensuring Consistency, Reliability, and Accuracy in Scoring

Given the high-stakes nature of these proposed assessments, helping ensure reliable and accurate human scoring is critical. We propose a multilevel, multifaceted approach, because it is most effective at establishing that only the raters who learn to use the scoring rubric accurately are allowed to begin scoring, and it verifies that raters stay on track throughout scoring. The following is a list of some of the procedures that are used to monitor and train raters to help ensure consistency in scoring – regardless of whom makes up the rater pool (i.e., teachers or professional raters).

- » Rater calibration — occurs prior to operational scoring and tests the rater’s abilities to appropriately apply the scoring rubric to specific items
- » Response randomization — occurs during operational scoring and randomizes responses distributed to raters
- » Double reads — occurs during operational scoring and requires two independent rater scores for items, if required by the design of the program
- » Response distribution rules — occurs during operational scoring and helps ensure that no systematic biases are introduced into scoring
- » Scoring leader backscoring and validity papers — occurs during operational scoring and helps ensure that raters are scoring according to the scoring rubric and rules
- » Trend scoring and equating — occurs during operational scoring when items are reused between administrations to determine a statistical comparison

The combination of these procedures creates a strong framework of checks and balances that protects scoring fidelity, while attempting to minimize the amount of additional scoring time and cost needed to establish a strong and defensible process.

During Generation 1 of the assessment system, types of items that either require human scoring, or use of as-yet-unproven technologies for computer scoring, should be limited to the formative system and/or the interim-assessment system, with results aggregated to the accountability system. As we gain confidence in their use for high-stakes purposes, we can graduate them to the end-of-year summative assessments for Generation 2.

As mentioned in the Introduction to this response, there are competing priorities that have a lot to do with the type of scoring used, especially on the end-of-year assessment. How important is rapid turnaround of results on the year-end summative assessment? Is it important enough to limit the end-of-year assessment to machine-scorable items, or can we take the time to do human scoring of some items, adding a few weeks to the processing required before getting score reports?

- Approach to releasing assessment items during each assessment cycle in order to ensure public access to the assessment questions

Although costly, the release of test questions is very beneficial for a number of reasons. Released items for the common assessment will enable students and educators to understand how the content standards will be measured and will give students practice with the various item formats used on the tests. This kind of information is especially valuable for open-ended questions, with their accompanying rubrics and sample responses. To the extent to which we succeed in making the exercises models of effective instruction, released exercises can provide a useful toolkit to teachers. In addition, released exercises tend to demystify assessments.

We have provided a more complete discussion of releasing test questions in response to question 3 under “Specific Technical Assessment Questions.”

- Technology and other resources needed to develop, administer, and score the assessments, and/or report results

One of the major questions facing the designers of a common-standards assessment is “How much technology, how soon?” Certainly, the current state of technology availability in many states and the current price structures of testing programs would argue that an assessment system should offer a paper-based test, or at least a program that could be administered on paper as well as online. In spite of this, we believe that, as soon as it is practical, the assessment of common standards should be computer-based (or other-technology-enabled) tests in which paper is used solely for certain special accommodations. We describe this ideal version of the Race to the Top assessment program as Generation 2 throughout this document, and believe that the transitional system (Generation 1) should consist of a steady march toward the eventual goal of having almost all of the system be computer-administered.

There are several reasons for recommending that the entire system be computer-delivered in Generation 2:

- » Emerging standards in both mathematics and reading/language arts define constructs that can only be measured through the use of technology. This is likely to be true in subjects such as science as well. Maintaining parallel paper and computer systems on which results are supposed to be interchangeable would effectively prevent measurement of such skills. This “assessment tail wagging the education dog” has been a large criticism of education reform efforts in the past, and we want to avoid this.
- » Technology allows for the use of a range of forward-looking exercise types, including item types that ask students to engage with digital content and formats, and brings to bear skills that wouldn’t (and couldn’t) be invoked on a paper test.
- » Testing some skills on paper may simply yield invalid results in the future.
- » Technology allows for flexible (adaptive) and on-demand testing, which should be a part of this design.

- » Technology allows for electronic scoring of some sorts of items, and thus for use of a broader range of items than does paper-based testing. Technology also facilitates the distribution of student responses to teachers, monitoring the quality of teacher scoring, and increased opportunities for professional development in terms of assessment development and scoring.
- » Rapid return of scores and seamless data/information interchange is facilitated by technological delivery.
- » If the summative assessment is delivered via a technology platform based on accepted interoperability standards, it could feed data to, and receive data from, the interim and formative segments of the system, thus creating an integrated, balanced system.
- » Technology will continue to improve, become easier to use, and become more common in the future such that our proposed system will be operationally feasible.
- » Technology allows for provision of a range of accommodations for students with disabilities and English language learners that might not otherwise exist.
- » Using technology administration as the single delivery paradigm simplifies issues with comparability.

This decision, of course, has major operational implications. Even with expanded technology access, we cannot rely solely on mass administrations, so scheduling becomes essential. Testing windows will need to be open long enough to accommodate test takers, and exercise pools will need to be large enough to protect test security. The final system must allow for tradeoffs between assessment purpose (like high-stakes graduation decisions) and the size of the testing window allowed. Finally, because it is likely that state-specific content will be developed by a number of different entities, we would need a set of data transfer and delivery protocols that could be used by all involved.

During Generation 1, we recommend that computer administration be used as much as possible for the interim assessments and formative assessments, at least. This would allow for the use of newer item types and scenarios that measure 21st century skills and Information and Communication Technology (ICT) literacy throughout the year. It would also allow for the build-up of capability and capacity in the scores for the eventual transition of the entire end-of-year assessment to technology delivery in Generation 2. During Generation 1, the summative assessment can be administered via technology in those states and LEAs that are ready for it, but might have to be administered, at least for two years or so, by paper in other states and LEAs. While this puts severe limitations on the types of items we can include in the summative assessment during Generation 1, and slows down turnaround time for some LEAs and states, it would allow for quicker implementation of the overall assessment and avoid the problems that would occur in forcing the system into total technology administration before the infrastructure and operational base is ready.

3) ARRA requires that States award at least 50 percent of their Race to the Top funds to LEAs. The section of the notice entitled Design of Assessment Systems – LEA-Level Activities, describes how LEAs might be required to use these funds. What activities at the LEA level would best advance the transition to and implementation of the consortium’s common, college and career ready standards and assessments?

The Race to the Top funds provides LEAs with an extraordinary opportunity to participate in the development and implementation of next generation assessments. In particular, resources directly available to LEAs provide the chance to build capacity in ways that improve teaching and learning. Additionally, the investment in and among LEAs will help prepare all students for college and careers.

The following LEA-level activities would, in our judgment, best advance state consortia common standards and assessments.

Development of Formative Components of Assessment Systems

As part of a balanced next generation assessment system, LEAs can help develop formative components designed to work with the summative components that are centrally developed. These components can include rich performance tasks that are closely aligned to classroom practice, reflect learning progressions, serve to reinforce learning, and identify gaps in knowledge and skills.

While it is also possible for LEAs to be involved in the development of the interim components of the system, this is a bit more challenging if those interim components will be used for accountability purposes via combination with the summative assessments. The interim assessments would have to have some degree of standardization and security for that plan to work.

Professional Development

One of the most enduring and useful expenditures of Race to the Top funds at the LEA level would be to promote professional development activities related to a next generation assessment system. Without teacher and school staff involvement in and understanding of what the changes are and why they are being implemented, the new assessment elements introduced by this initiative will at best be minimized; at worst they will be frustrating to staff and eventually ignored.

The following are some of the ways that school and LEA staff can learn about and better appreciate the changes to their assessment program:

Assessment Literacy: A next generation assessment system will bring new terminology and concepts that could confuse and intimidate those who are put in the position of explaining the new assessments to parents and the general public — terms such as common standards, 21st century skills, international benchmarking, and adaptive testing. Assessment literacy requires an understanding of types and purposes of assessments, how to glean information from summative assessments, and the use of student achievement data in teacher performance evaluations. Teachers need to be thoroughly aware of the changes that will result from the new assessment

system because they are the front line, talking daily with parents, neighbors, and others outside of education. Their understanding and support are critical to the implementation and success of a new assessment system.

Writing/Reviewing Test Questions and Scoring Open-ended Questions: Training teachers and staff to both write/review new test questions and score open-ended questions has been shown to be one of the most beneficial professional development experiences for those who provide and manage instruction. Likewise, involving teachers in item and test development activities provides important opportunities for training in the writing of clear and accurate items that are aligned to the standards, as well as effective assessment design. As previously noted, teachers can also help to develop formative assessments — the perfect arena for local staff to make a significant contribution to the state’s assessment system. Because part of the cost of using teachers to score assessments is hiring substitute teachers to replace them in the classroom, this is an appropriate use of the LEA flow-through funds under the Race to the Top assessment program.

Capacity Building

As states adopt common standards, considerable time and effort will be needed to align local curriculum, assessment, and instruction to the new standards. LEAs need resources, direction, and support to develop benchmark and formative-assessment materials and encourage the use of multiple measures that work together to achieve common goals.

Funding at the local level could help manage logistics, communication, and technical support to create sustainable programs that align curriculum and instruction at each grade level with the common and state-specific standards.

Capacity building involves purchasing instructional programs and resources, and doing the professional development that enables teaching and learning of the new standards to be most effective.

In addition, we should encourage LEAs and schools to collaborate with others and share resources and best practices, which includes a focus on capacity building at all levels. County and regional cooperatives can pool resources to purchase services that may otherwise not be available to individual LEAs.

4) If a goal is that teachers are involved in the scoring of constructed responses and performance tasks in order to measure effectively students’ mastery of higher-order content and skills and to build teacher expertise and understanding of performance expectations, how can such assessments be administered and scored in the most time-efficient and cost-effective ways?

Human scoring adds challenges for any assessment system. Human scoring means that it takes longer to release scores. It also raises assessment costs and psychometric challenges regarding data comparability, particularly over time. Of course, human-scorable items also allow for the measurement of skills beyond what is possible in a machine-scorable system. And finally, scoring itself can be a powerful professional development activity. This tension leads to a key question: How do we mitigate

the negatives and enhance the positives? How do we improve quality and control cost and time? The answer involves, at least partly, an effective use of online scoring technology.

Student written responses were traditionally scored in a face-to-face (F2F) setting, with raters assembled in a central location. This approach had some real advantages. There were few technology needs, as the actual student books were used in scoring. F2F scoring allows for personal interaction between raters and scoring leadership. However, there were two main limitations of this approach. First, real-time quality control and assurance tools in paper-based systems were limited. Second, F2F scoring required expenditure on travel, lodging, subsistence, facilities, and equipment.

Online scoring has helped address both of these weaknesses. The systems themselves provide real-time quality control and assurance tools. The systems also obviate the need for F2F training: One can create either online or virtual F2F environments and dramatically reduce costs for travel, lodging, subsistence, facilities, and equipment through the use of online distributed scoring. Distributed scoring allows participation in the scoring process by teachers with computers and internet connections from anywhere in the world.

Approaches to Involving Teachers in Constructed Response Scoring

We propose alternative ways to meet the Department's goal of involving teachers in the assessment process — ways that would promote buy-in and opportunities for professional development. Employing teachers to complete the constructed response scoring of field test or operational items is one approach to achieving the goal of involving teachers in the assessment development process. The degree to which teachers are involved in scoring, however, must be balanced with regard for the apparent conflict of interest. In an accountability context where the outcomes are consequential for teachers as well as for students, a rater pool of teachers might compromise public acceptance of the validity of the outcomes. This is not to argue that teachers should not be involved, but rather that the system must help ensure that this involvement does not threaten the validity of scores.

In addition to carrying out all operational scoring, there are other types of teacher involvement that can increase confidence in test scores from a stakeholder's perspective, and reliability and validity from a psychometric perspective. One approach would be to invite a diverse panel of teachers to evaluate the alignment between scoring criteria and the broader performance expectations described in the set of K-12 standards. During such evaluative exercises, teachers would map the scoring criteria for each content area to the performance expectation they judge to be the best fit. The degree to which these judgments align with the standards and with one another would provide information about how well the performance expectations are used by the scoring criteria. Another approach might entail teachers' input in pilot item and scoring guide review. When constructed response items are field tested, teachers would review the items and score responses using the accompanying guidelines. The merit of the items and scoring guidelines would be judged based on the quality of responses received and how well the guidelines can be applied to those responses. Teachers participating in this activity would interact with colleagues from other LEAs and states and gain valuable professional development experience in crafting performance assessments.

Another alternative to actual scoring is to get teachers involved in a range of quality control processes such as exemplar (sample) response selection. Teachers would collaborate with assessment developers to choose the responses used for training, certification, calibration, and validity monitoring of raters. This highly deliberative session gives participants the opportunity to discuss the item, the expected performance standard, and the scoring rubric in detail. This process results in samples of student responses established as clear examples that represent each attribute of the scoring rubric. One set of exemplars is used to train raters for operational scoring; these exemplars serve as reference points during the scoring process. The others are used in rater scoring quality monitoring.

5) Given the assessment design you proposed in response to question 1), what is your recommended approach to competency-based student testing versus grade-level-based student testing? Why? How would your design ensure high expectations for all students?

The distinction between grade-based testing and competency-based testing is neither clear nor absolute. For example, standards for grade 4 students define competencies we expect students at that level of schooling to have obtained. It is possible to build assessments of those competencies. The decision about whether to administer those assessments to all grade 4 students or to allow grade 3 students to take tests if their teachers view them as ready to do so does not necessarily change the nature of the assessments. That having been said, there are ways to address the question above that help illuminate choices test developers will need to make in designing the new system.

One other point of definition: The phrase “competency-based testing” has been used in two ways. The first is to describe an assessment system in which students take tests when they are ready to show mastery of the competencies covered in that test. So in the example above, a grade 3 student might take the grade 4 competency test if ready. For purposes of discussion, we will call that “testing-when-ready.” The other meaning of “competency-based testing” tends to refer to separate testing of distinct competencies or clusters of competencies (which in the current discussion resemble standards). We will call that “assessment of specific competencies.” Given the nature of the question above, we assume the former usage is intended, although we will say a few words about both.

The summative assessments of common-core standards can combine elements of competency-based and grade-level-based testing. As mentioned above, at least in the early years of the new assessment, we will need to administer end-of-grade or end-of-course assessments to allow for the collection of system-level accountability data. We also recommend that the summative system make use of adaptive administration. In this case, off-grade content may be selected for either high- or low-performing students. However, all students would be tested with on-grade content, and the only use of off-grade content would occur as a result of adaptation. Finally, this helps ensure high expectations for all students in that all children will be evaluated against the within-grade rigorous standards, even if off-grade content is administered.

While the system will begin as a grade-based assessment regime, it could easily evolve into one in which people test when they are ready. For example, the system could include an end-of-grade 5 mathematics test, which will measure student mastery of the appropriate standards. Once we achieve the eventual

goal of a completely computer-based adaptive system, there is no reason that this could not evolve into an on-demand system in which students test when they are ready. So in this example, if a student in grade 4 felt ready to show mastery of grade 5 content and skills, he or she could. Of course, this system loses some of the data advantages of having fixed snapshots in time (for example, this may make it harder to make school comparisons, or implement value-added models). In addition, this sort of flexibility in testing has implications for instructional management in schools. Policymakers will need to determine whether these sorts of challenges are worth it given the advantages of an on-demand system.

We also believe that this system can accommodate the other meaning of competency-based testing, that individual testing or test-based events focus on individual competencies or clusters of competencies. This is not to recommend such an approach, but simply to state that it is possible. We argue that the summative system may not be a single testing event, and may rather be a combination of testing events that occur over the course of the year. In a system where there are multiple tests given at fixed intervals, these tests can in some cases focus on specific competencies. This implies that we can agree on the clusters of competencies that should be combined in the intermediate tests. Alternately, it implies a library of competency tests that states and LEAs can select and use with some discretion, although such a system would certainly reduce the comparability of results.

The use of a testing approach in which the summative system uses data from multiple tests does not necessarily assume competency-based testing of this sort, of course. There are several reasons not to consider separate assessments of specific competencies. Such testing has been criticized for encouraging inappropriate disaggregation of skills that should be viewed and assessed as integrated. Additionally, such approaches are easier to implement in mathematics than in reading/language arts.

In summary, the high-level system design has elements of grade-based testing, but could evolve into one that includes competency-based testing, whichever way one defines that term.

6) Given the assessment design you proposed in response to question 1), how would you recommend that the assessments be designed, timed, and scored to provide the most useful information on teacher and principal effectiveness?

Student and school effectiveness cannot be gauged based simply on percentages of students who reach standards. Different schools face different levels of challenge, and different teachers add varied levels of value. A system that has the measurement of teacher and school effectiveness as a goal requires data on the amount and nature of student improvement over time. In other words, if we are to use student performance information as a source of data on teacher and school effectiveness, we must have data on student growth.

Given the overall interest in student growth metrics (and the use of such metrics in teacher evaluation), we believe the assessment system should support cross-grade comparability, and the assessment will need to be set up to allow for such comparisons. This work will, of course, be greatly facilitated if the content standards and expectations are coherent across grades. In addition to supporting growth

modeling, cross-grade comparability facilitates another element we view as desirable in the system: the ability of flexible administration engines to select out-of-grade content for either advanced or struggling students.

There are interesting questions that will need to be answered in this area. For example, while it is likely that some constituents will want to see assessments at grades 3 through 8 on a vertical scale (perhaps mistakenly thinking vertical scales are required for growth measures), it is not at all clear that high school assessments should (or need to be) placed on such a scale. Frankly, the notion of comparing performance in various high school subjects, such as chemistry and algebra II, is problematic in itself. In the past, states have not tended to require this, and high school content may not be as friendly to cross-grade comparability. But there is a real need for data on whether or not high school students are proceeding as necessary.

It is worth mentioning that there are several ways to produce measures of growth and cross-grade comparability. How the requirements of specific growth models affect the system will need to be studied, but such considerations are beyond the scope of this response.

A well-structured student assessment system can be one source of data to be used in evaluating teacher effectiveness. One thing policymakers will need to consider is how to use these data in conjunction with other relevant pieces of information.

Specific Technical Assessment Questions

1) What is the best technical approach for ensuring the vertical alignment of the entire assessment system across grades (e.g., grades 3 through 8 and high school)?

From a technical standpoint, a vertically aligned assessment system is best developed within the context of coherent, vertically articulated content standards and performance expectations. An effective system includes within- and cross-grade alignment of standards-based instruction that is informed by assessment results and supported by ongoing professional development. Vertical articulation across grades is evident in performance level descriptors and the cut scores established to differentiate proficiency levels at each grade.

The assessment system would support within-grade proficiency measures as well as cross-grade comparisons of individual student performance. The cross-grade comparisons are desired to measure growth and determine at each grade level tested whether a student is on track toward college or career readiness by the time of high school completion. Recommendations for the best technical approach to helping ensure vertical alignment of the assessment system, and for scaling and reporting to support cross-grade inferences, will depend on a number of factors, as they pertain to the adopted standards and performance levels.

The provision of coherent, vertically articulated content standards and performance expectations could permit the construction of cross-grade or vertical scales. A vertical scale entails the notion of learning or growth across time. Because the common assessment tests students across a grade range with articulated content, longitudinal types of inferences based on scale scores would be possible if a vertical scale were implemented.

It should be noted that developing a vertical scale is technically complex, and may not be feasible in all situations. In these situations, use of within-grade or horizontal scales does not preclude measuring student growth. Although horizontal scales support direct comparisons of scale scores between same grade cohorts only, additional statistical procedures may be used to track individual growth over time.

We describe considerations for vertical scaling and other options to estimate student growth across grades in the context of vertically aligned content standards and performance expectations below.

Using Cross-grade Scales (Vertical Scaling)

As stated above, one approach to facilitate measurement of student growth is to develop a vertical scale. In a vertical scale, for each content area (e.g., reading/language arts or mathematics) scale scores run continuously from the lowest grade tested to the highest grade tested, with substantial overlap of the scale scores produced by adjacent grades. On the vertical scale, “Proficient” might be a scale score of 350 in grade 3, 380 in grade 4, 400 in grade 5, and so forth. The difference in a student’s scale scores at adjacent grades is a measure of the amount of academic growth achieved by that student.

With a vertical scale, an ideal goal is to have scale scores that have the same meaning if they are obtained from different test levels (e.g., a 400 “means the same thing” or represents equivalent knowledge or achievement whether it comes from the grade 4 test or the grade 5 test). Also, differences between scores are ideally comparable for gauging amounts of academic growth. For example, a student who grows from a scale score of 350 to 370 (20 points) would be demonstrating the same amount of growth as a student who grows from 390 to 410 (20 points).

In addition, for ease of interpretation and use in measuring growth, vertically scaled item pools allow the use of adaptive administration engines to select out-of-grade content to provide additional diagnostic information for either advanced or struggling students. Note that under current interpretations of NCLB, the content of assessments used for accountability must be aligned to grade-level content standards; however, the diagnostic out-of-grade content could be administered as an augmentation to the grade-level test. In the context of interpreting individual academic growth for students, the construction of a vertical scale is critical for these types of inferences.

Listed below are some technical considerations in producing the vertical scale:

- » To produce a vertical scale, it is assumed that the tests at adjacent grades have substantial overlap or articulation in content, and that a single, major dimension (e.g., overall mathematics achievement) explains most performance differences. An additional consideration is that vertical scaling makes the implicit assumption that the same construct is being measured at the top and bottom of the score scale, and this assumption may be difficult to justify when the vertical scaling includes many grades. Caution is needed in comparing growth in different parts of a vertical scale, whether comparing growth of low-achieving and high-achieving students or students in substantially different grades.
- » Beyond the expert judgment involved in establishing vertical articulation of content, a vertical scaling requires special data collection and analysis. (In the data collection, students take test items that measure content from adjacent grades in addition to content from their own grade; their relative performance on the two sets of items determines the relative difficulty, or scaling, of the different sets of items.) This is of particular concern for the high school assessments under the current paradigm. For example, administration of comprehensive assessments in grades 9, 10, or 11 could perhaps support cross-grade scaling if based on content that builds sequentially from lower grades, such as might be the case in certain versions of integrated mathematics curricula. On the other hand, administration of end-of-course assessments under traditional curricula (such as algebra I, geometry, algebra II or biology, chemistry, and physics) might not articulate. Of course, the fact that an approach might not work in high school does not render it ineffective at earlier grades. The new system need not rely on a single approach.
- » A similar issue for consideration involves the state-specific content, and the degree to which it is a) aligned to the common-core standards and b) vertically aligned across grades within a state. It is likely that state-specific content may differ substantially across states; thus, it may be that a vertical scale is developed for the common-core component only, for the purpose of cross-state comparisons. In other words, the vertical scale based on the common core will

measure growth only on the common core. If a state wanted to include their state-specific augmentation in the vertical growth measurement, then a state-specific vertical scale would be needed.

Using Within-grade Scales (Horizontal Scaling)

As stated previously, vertical alignment of the assessment system is best realized in the context of vertically articulated content standards and performance expectations. Although appealing for reasons listed above, vertical scales are not the only way to measure student growth. When assessment systems have separate within-grade scales, whether by design or because vertical scales are not feasible, alternatives that do not require a vertical scale may be used to estimate student progress across grades. For example, regression-based techniques to estimate growth percentiles or growth trajectories may also be used to determine whether students are on track to achieve the desired level of performance at various designated time intervals.

In contrast to vertical scaling, use of separate within-grade scales is not predicated on the assumption that a single major dimension is measured across grades; thus, it allows more flexibility in content differentiation. This is an advantage when considering how to assess students at the high school level where the content domain is more varied and course-specific.

Selection of statistical methods for estimating student growth should take into account the transparency of the statistical method used and its interpretability for the public and educational decision makers.

Summary

Recommendations include coherent, vertical articulation of content and performance standards across grades, with instruction aligned to standards and assessment; design decisions that support test properties adequate for the intended use; and the use of vertical scales if feasible. It is expected that the assessment system would support within-grade (horizontal) proficiency measures as well as cross-grade comparisons of individual student performance. Student growth measures may be obtained with or without a vertical scale, and there are some limitations associated with both options.

While beyond the scope of this response, one cautionary note is in order. Regardless of the ultimate assessment design and scaling procedures, given the high stakes associated with test results, we recommend careful monitoring of continued alignment. As with any educational reform involving new content standards, there may be changes resulting from clarification of standards and/or implemented curriculum, with implications for parameter estimation, and the stability of scales and the validity of performance standards established in the early years. To this end, we recommend periodic studies to evaluate whether revisiting performance standards and/or resetting baseline scales may be warranted.

2) What would be the best technical approach for ensuring external validity of such an assessment system, particularly as it relates to postsecondary readiness and high-quality internationally benchmarked content standards?

International Benchmarking

The phrase “international benchmarking” has several possible meanings. It can be defined as a content activity aimed at determining what leading students around the world are being taught. International benchmarking can be a level-setting activity, in which we set performance expectations (and related cut scores on tests) at points where they suggest readiness to compete in a global economy. Finally, international benchmarking sometimes refers to studies that compare actual educational attainment in different countries. TIMSS, PIRLS, and Program for International Student Assessment (PISA) are examples of such studies. These definitions of “benchmarking” are not contradictory: Individual studies can undertake all three sets of activities.

What is the value of international benchmarks? They allow countries as well as local educators and educational researchers to better understand the relative strengths and weaknesses of their education systems and possibly to identify best practices and plan appropriately.

If we are to make international benchmark data integral to the new assessment of common standards, we must have an ongoing plan to maintain the validity of those standards. This plan must cover the three benchmarking definitions described above. The Department should consider periodic support for curriculum reviews in high-performing countries, to make sure that “common-core standards” are keeping pace with international advances. In addition, as expectations around the world change, it may become necessary to update cut scores on the common-standards assessments. Finally, the Department should promote some linkage between the common-core assessments and international surveys (possibly through studies linking NAEP and these surveys, since all states already participate in NAEP). This linkage would likely be conducted through special studies in which some students take both the international survey and NAEP, or in which randomly equivalent samples do the same.

Postsecondary Readiness

Another issue relates to the evaluation of postsecondary readiness and external validity of the common assessment. Both predictive and convergent/discriminant validity can be used to support the assessment system in the context of postsecondary success. Predictive validity is concerned with the use of test scores to predict a variable of interest. Not surprisingly, students with higher levels of measured academic skills are more likely to have higher grades and graduate from college than their less able peers. Likewise, if students with low common assessment scores have to take college remedial courses, then this is an indication that these students were not adequately prepared for college-level material. A number of studies have investigated the predictive validity of the SAT or ACT® on grade point average (GPA) in the first year of college (e.g., Kobrin, Patterson, Shaw, Mattern, & Barbuti, 2008) and for other types of college outcomes. In a similar fashion, the common assessment could be used to predict first year college GPA or other outcome variables such as college graduation status. Because SAT and ACT test scores are the second-most important factor in college admissions decisions after high school

grades (Hawkins & Lautz, 2005), another option is to use the common assessment to predict college entrance performance on the SAT or ACT examinations.

There are numerous indicators of postsecondary success, including postsecondary institution enrollment, persistence, remediation, and degree completion; employment; military enlistment; and earnings that also can be used to demonstrate convergent and discriminant validity. These types of studies would require a concerted data collection effort but not one that is unprecedented. As of 2008, 20 states were providing postsecondary feedback reports on high schools, including information about each high school graduate's participation in postsecondary education (Hanover Research Council, 2008). Unfortunately, state strategies for implementing the postsecondary feedback report — including how they collect data, produce the indicators, and report information — vary significantly. For example, while most states include enrollment and remediation data in their feedback reports, only a handful report persistence rates and first year college GPA, and even fewer provide data on degree completion or workforce participation. For the common-core assessments, there would need to be more coordination in the definition and organization of these types of information in order to support these types of validity inferences.

3) What is the proportion of assessment questions that you recommend releasing each testing cycle in order to ensure public access to the assessment while minimizing linking risk? What are the implications of this proportion for the costs of developing new assessment questions and for the costs and design of linking studies across time?

We believe that the release of exercises should take place in three stages: at the prototype stage, after initial pilot testing, and on an ongoing basis after operational testing. Because released items are an important part of a testing program, a number of prototype items should be disclosed as soon as possible after the assessments have been designed, even before any items have been pilot/field tested. The prototype items should be widely available on the Internet so that students and teachers can begin to understand the overall assessment design as well as the item formats.

After items have been pilot/field tested, a predetermined number of pilot/field tested items should also be released, possibly with one or more statistical characteristics (e.g., item difficulty) accompanying each item. Both of these releases should include sufficient numbers of items to demonstrate how key standards will be tested and to illustrate the complete set of item types that will be used.

Traditionally, the items released after pilot/field testing comprise either a representative set, perhaps the equivalent of 15 percent to 25 percent of the items an individual student may take, or the number of items that appear in a full-length testing session. We recommend the latter approach, given the importance of this assessment.

After the program is up and running operationally and pools have grown to a “steady-state size,” we believe that each year new content addition should be more or less balanced with the release of exercises from the operational pool. So for example, if we add 100 new grade 4 mathematics items to an adaptive pool, we would release roughly 100 for use by educators. We would further propose that the

released exercises be accessible in at least two ways. First, teachers who wish to use these exercises for classroom purposes should be able to retrieve them from the Web. Second, students and teachers should be able to access the items in a way that emulates an actual administration (in other words, that is adaptive and that produces a scaled score, if teachers are willing to score the open-ended items). Such uses would be low-stakes, of course.

This sort of an aggressive release plan has the advantage of giving teachers and students more complete information about the knowledge and skills to be assessed as well as the level and type of performance that are expected. In addition, the release of forms rather than single items avoids an appearance of secrecy regarding the tests. However, it can be argued that releasing one or more complete or multistage forms increases the likelihood of teachers focusing on test items rather than student learning. Also, the cost of releasing intact forms, even smaller multistage forms, is much greater than the cost of releasing representative sets.

The plan for releasing items must be fully crafted prior to beginning item development. It is important to develop a sufficient number of items, with the customary overages, to allow for the planned releases. A careful plan will permit the release of items without any risk to linking items.

High School Assessment Questions

Provide recommendations on the optimal approach to measuring each student's college and career readiness by the time of high school completion. In particular, consider:

- 1)** How would you demonstrate that high school students are on track to college and career readiness, and at what points throughout high school would you recommend measuring this? Discuss your recommendations on the use of end-of-course assessments versus comprehensive assessments of college and career readiness. (Note: If you recommend end-of-course assessments, please share your input on how to reconcile the fact that college and career ready standards might not include all of the topics typically covered in today's high school courses.)

It would be ideal to measure readiness in grades 10 and 11, so that students are encouraged to use their last two years of high school, and especially grade 12, to increase their mastery of the readiness standards.

To demonstrate that high school students are on track to college and career readiness, there are two available models:

- » States would administer comprehensive reading/language arts and mathematics assessments in grades 10 and 11, and these assessments would be designed to directly measure the college and career readiness standards. Cut scores would be established to indicate degrees of readiness, with sufficient time for students to increase their readiness.
- » States would administer end-of-course assessments (in such courses as English II and III, geometry, algebra II, biology, and chemistry) and supplement those assessments with sets of questions that specifically measure readiness standards. For the purposes of giving readiness scores, items from the end-of-course assessments would be combined with the readiness items to form a reliable full-length assessment. This model has been in use in California's Early Assessment Program (EAP) since 2005. The state administers a set of readiness items along with their grade 11 end-of-course assessments: English grade 11, algebra II, and high school summative mathematics. Under this model, a similar supplement could be developed for grade 10 and for science, which currently are not part of the California EAP.

Because the material typically covered in today's high school courses is usually broader than the requirements of the college and career readiness standards, California selects half to two-thirds of the items on the end-of-course tests to use for the EAP measure. The selected items meet specific statistical and content specifications.

Some educational leaders in the U.S. have recommended that we employ a system of extensive examinations in high school, similar to that used in other countries, like the United Kingdom. There are many advantages to this proposal, especially in that it drives the high school curriculum and high school teaching to prepare students for extensive, multifaceted assessments that are worth teaching to. The type of system we have proposed here is not incompatible with such an approach, especially if the system we propose is used for grades 3 through 8 and another system is used for high school. We assume that the proponents of the examination approach are submitting their comments and advice to this request for information separately, so we will not deal with this subject in detail here.

Questions on the Assessment of English Language Learners

- 1) Provide recommendations for the development and administration of assessments for each content area that are valid and reliable for English language learners. How would you recommend that the assessments take into account the variations in English language proficiency of students in a manner that enables them to demonstrate their knowledge and skills in core academic areas? Innovative assessment designs and uses of technology have the potential to be inclusive of more students. How would you propose we take this into account?

Importance of Assessing English Language Learners

English language learners comprise a large and rapidly growing portion of the K-12 population in the U.S. According to recent statistics from the National Clearinghouse for English Language Acquisition (2007):

- » Currently 1 in 9 K-12 students in the U.S. is an English language learner; by 2025, as many as 1 in 4 students may be.
- » English language learners are prominent in our largest states. For example, in California, more than 25 percent of the student population are English language learners; in Texas, more than 15 percent.
- » Many states that have not traditionally had large numbers of English language learners have experienced rapid large-scale influxes in recent years.

The import of these data for the next generation of common summative assessments is clear: In order for it to successfully serve the purposes for which it is designed, the new assessment system must provide fair and valid information about the skills and abilities of English language learners.

Factors Influencing the Assessment of English Language Learners

The challenges of designing and implementing an assessment system that is appropriate to both the general population and to English language learners involve factors related to language, to educational background, and to culture. These factors are briefly outlined below. For a more detailed discussion, see Pitoniak et al. (2009).

Language

- » *Different language backgrounds:* While approximately 80 percent of English language learners come from Spanish-speaking backgrounds, it has been estimated that more than 400 languages are spoken by English language learners nationally.
- » *Varying levels of proficiency in English:* English language learners can range from true beginners with minimal English skills to students with levels of fluency approaching that of native speakers. A particular challenge for assessment is posed by the fact that there is no predictable relationship between age (or grade) and level of English proficiency; one English language learner in grade 10 might be a true beginner just learning the English alphabet,

- while another might be highly fluent in English. Students at these two extremes — and at all levels in between — will interact very differently with content assessments.
- » *Differing profiles of English language proficiency:* English language learners may have varying degrees of relative proficiency in oral versus written English and in interpersonal versus academic English. Students who are able to converse fluently in English may not have the literacy skills required to negotiate a standardized test.
 - » *Varying levels of literacy in native language:* As English language learners vary in the degree to which they can read and write in their native language, it is important not to assume that they will be able to understand written test directions or other test content in their native language.

Educational Background

- » *Varying degrees of formal schooling in native language:* In addition to native language literacy levels, the degree of formal schooling in the native language also affects English language learners' content-area skills and knowledge.
- » *Varying degrees of formal schooling in English:* English language learners vary both in the number of years they have spent in English-medium schools and also in the type of instruction received there (e.g., bilingual, full English immersion, English as a second language).
- » *Varying degrees of experience with standardized testing:* Any testing format — including multiple-choice items, constructed response items, and computer-based administration models — will likely impact different cohorts of the English language learner population differently depending on their degree of previous exposure.

Culture

- » *Varying degrees of acculturation to the U.S. mainstream:* Students who are unfamiliar with American culture may be at a disadvantage relative to their peers because they may hold different assumptions about the testing situation or the educational environment in general, have different background knowledge and experiences, or possess different sets of cultural values and beliefs, and therefore respond differently to test directions and questions.

The Goal

This question calls for recommendations for the development and administration of assessments for each content area that are valid and reliable for English language learners, as well as for the broader population.

The primary goal of any effort to make content-area tests valid and reliable for English language learners is to reduce the level of construct-irrelevant variance stemming from English language learner status. That is, we should work to help ensure that the tests are accurate assessments of what English language learners know and can do in reading/language arts and mathematics. Of course, the elimination of construct-irrelevant variance assumes an understanding of “construct relevance.” This careful definition of the constructs of interest is a first important step. The relationship of the construct to English

language learner status is, of course, different for reading/language arts than for mathematics. For reading/language arts, proficiency in English is a fundamental and essential enabling skill. For mathematics, the situation is a bit more complex. If the construct of interest is *mathematical skill exclusive of language skill*, we should provide English language learners with as pure a measure of mathematical skill as possible. If the construct of interest is defined as *the ability to perform and communicate about mathematics in an English-medium classroom*, however, the task is more subtle: We should make sure that the English language load is low enough to minimize construct-irrelevant variance, even while recognizing that some degree of communicative competence in English is part of the construct.

Several additional goals can also be identified as elements that will support the general goal above:

- » Maintain the inclusion of English language learners in the nation’s assessment and accountability system and continue to use the progress of English language learners toward meeting standards as an essential criterion by which the success of schools, LEAs, and states is measured.
- » Provide assessments that will give meaningful information about all English language learners, including those whose performance levels are currently well below grade-level expectations. These assessments should provide accurate information both about English language learners’ current skill levels and about the relationship between those skill levels and grade-level expectations.
- » Minimize the need for lower-proficiency English language learners to take tests that are inappropriate for them, ones on which they cannot perform meaningfully and will have a frustrating and discouraging testing experience.
- » Minimize the need for English language learners to be double-tested by, for example, taking both an English language proficiency assessment with substantial reading and writing components, and also a separate reading/language arts assessment focused on reading and writing (unless there are meaningful differences in the information provided by the different assessments).

Recommendations

To help ensure that the assessment system enables English language learners to demonstrate their knowledge and skills in core academic areas, we recommend the following:

- » Set the same standards for English language learners as for all other students in terms of what constitutes “proficient” or “advanced” performance. Do not, in any way, lower the level of assessments or expectations for English language learners.
- » In the guiding documents for the assessment system (curriculum standards, framework documents, test specifications, item specifications, etc.) include construct definitions that provide clear operational definitions of the role of English language proficiency in the construct. For example, the September 21, 2009, draft of the *College and Career Readiness Standards for Mathematics* contain several standards implying that communicative skill is

part of the construct of interest (e.g., Core Practices 2, “Construct viable arguments;” Statistics Core Skills 3, “Interpret data displays and summaries critically; draw conclusions and develop recommendations.”). For these standards to be used in an effective assessment of English language learners’ skills, such statements should be elaborated to clearly define the level of linguistic skill called for.

- » Enlist the meaningful participation of teachers experienced in working with English language learners in the development of all aspects of the assessment system including the standards, the framework documents, the test specifications, and the item specifications. Make sure that such experts participate on item review committees (for content and for fairness); in item analysis; and in differential item functioning (DIF) analysis.
- » Establish a policy to help ensure the use of accessible language, providing guidance on how to minimize construct-irrelevant variance. The policy should specify that non-construct parts of the assessment (e.g., directions, many elements of the mathematics assessment, questions — though not passages — on the reading/language arts assessment) be phrased in language that is as clear and accessible as possible.
- » Include English language learners as a distinct subgroup in all item tryouts (including one-on-one tryouts of item types), small-scale pilot tests, and large-scale pilot tests.
- » Consider carefully the role of constructed response items requiring written responses in the test design. Because writing is the most challenging language skill for most English language learners, using tasks that require written responses to assess anything other than writing skill itself poses a considerable risk of introducing construct-irrelevant variance.
- » When constructed response items are included, consider the needs of English language learners in designing the items, developing the scoring rubrics, and selecting responses for training and for public release.
- » When scoring constructed response items, train raters to distinguish between construct-relevant and construct-irrelevant ways in which English language learner responses tend to differ from the responses of native speakers, and document those procedures.
- » In designing the assessment system, include formative assessments sensitive to the particular (and varied) developmental needs of English language learners.
- » Consider appropriate accommodations for English language learners (including English language learners with disabilities) throughout the test design and development process, including the design of the test delivery format. (See response to question 2, below, for a more thorough discussion of accommodations for English language learners.)
- » Consider carefully the use of pictures, graphics, and other nonverbal forms of communication in the test design. Such alternate modes of communication can help English language learners understand what is asked of them but can also create barriers to the effective assessment of students with visual impairments.
- » Provide valid information about the performance of students whose skills and abilities may be some distance below what is called for in the standards. (As noted below, adaptive testing may help in this regard.)

- » Use scales that allow one to track student progress over time. English language learners at lower levels of English language proficiency have a strong likelihood of performing below grade-level expectations, and it normally takes several years for English language learners to acquire enough English to successfully function in an English-medium classroom. As a result, English language learners need an assessment system that both accurately measures where they are and also tracks their performance over time, providing information about improved skills even while those skills remain below grade-level expectations. Consider a hypothetical example in which scores on a content assessment are reported on a scale of 0 to 100 and the “proficient” standard is set at 50 for grade 5, 60 for grade 6, and 70 for grade 7. A newcomer may arrive during the grade 5 year with little or no English language proficiency and score a 15 on the grade 5 summative assessment. That student then makes good progress over the following year and scores 35 on the grade 6 summative assessment. She continues to make good progress over the following year and scores 60 on the grade 7 summative assessment. An effective assessment system will show this student for what she is — a success story in the making — rather than simply labeling her as “below proficient” in grades 5, 6, and 7.
- » Consider the needs of English language learners in all research and validity efforts. Gather information on the performance of English language learners as a subgroup and consider these data in evaluating the performance and suitability of items and item types.

Finally, it should be recognized that any assessment system focused only on content assessments cannot meet all of the assessment needs of English language learners. To fully serve the needs of English language learners, an assessment of English language proficiency or an English language proficiency component to the general assessment system will be needed. Such a system should have a principled connection to the content-area assessments and should include a placement instrument that can be easily administered and locally scored; appropriate formative assessments; and a summative assessment of student progress in the acquisition of the academic English needed to learn in the content areas and succeed in English-medium classrooms. If thoughtfully designed, the system would collect and report valuable information about English language learner students’ skills and abilities while minimizing their need to be double-tested or to take tests not appropriate to their skill level.

Technology-enabled Assessments and English Language Learners

Innovative assessment designs and uses of technology have the potential to be inclusive of more students, including English language learners.

In considering the potential impact of technology-enabled assessments, it is worth noting that English language learners vary in their level of exposure to technology. Some English language learners may not have the technology skills that can be assumed in mainstream populations (e.g., the ability to use a computer mouse or basic skills in word processing). In addition, novel test formats (or ones that require linguistically complex directions) may disadvantage those English language learners who have limited experience with standardized testing. However, in the very likely circumstance that the standards call for working with digital and other types of technology (as do several standards cited in media

applications of the *College and Career Readiness Standards for Reading, Writing, and Speaking and Listening*), skills in this area will clearly be construct-relevant. The degree of familiarity with digital and other technological media should be clearly delineated in the guiding documents for the assessment system.

Three areas in which technology-enabled assessments have the potential to improve the assessment of English language learners are discussed briefly below.

- » Computer technology can make accommodations for English language learners considerably more efficient and user-friendly. For example, a computer-delivered test can be designed so that appropriate students have direct access to an online glossary (via “mouse over” or by clicking on hyperlinks). Also, computer-administration platforms can be modified to provide different lengths of time to students for whom this is an approved accommodation.
- » In presenting items calling for a written response, computer technology can present students with a range of planning tools, such as graphic organizers, to help them construct their responses. Such tools can be particularly helpful to English language learners, but it is important that the students have an adequate opportunity to become familiar with the tools before being tested.
- » An adaptive testing model can broaden the range of the scale covered in a single assessment, allowing accurate measurement of the skills of those English language learners who are currently some distance below grade-level expectations. Adaptive testing should be designed so that English language learners are not automatically shunted into a lower-level test, but have access to test items or sections that both (1) provide accurate information about their current skill level, as low as it might be, and (2) allow them to demonstrate proficient or advanced skill levels. In reading/language arts, for example, adaptive testing might allow a model in which the same passage could be used with different sets of test items: For students at or near the skill level specified by the standards, the test items could assess higher-level reading skills (e.g., challenging vocabulary, extended reasoning, subtle points of tone and voice) while students with lower skill levels could read the same passage but answer lower-level questions (e.g., literal comprehension and basic inferences).

2) In the context of reflecting student achievement, what are the relative merits of developing and administering content assessments in native languages? What are the technical, logistical, and financial requirements?

Content assessments administered in a language other than English (i.e., native language assessments) represent one type of testing accommodation that has been developed for use with English language learners. A testing accommodation is defined as “support provided to students for a given testing event either through modification of the test itself or through modification of the testing procedure to help students access the content in English and better demonstrate what they know” (Butler & Stevens, 1997, p. 5). The main purpose of providing students with a testing accommodation is to promote equity and validity in assessment. An interaction hypothesis has been proposed to justify the use of accommodations (Sireci, Li, & Scarpati, 2003). This hypothesis states that an accommodation will lead to

improved test scores for students who need the accommodation, but will not have an effect on the scores of students who do not need the accommodation.

Native language content assessments provide direct linguistic support for English language learners since the language used for the assessment has been altered (Rivera & Collum, 2008). Note that native language content assessments may not be valid for use in situations where English proficiency is integral to the construct being measured. To date, the number of research studies that have evaluated the effectiveness and validity of native language assessments is quite small. A recent meta-analysis by Kieffer, Lesaux, Rivera, & Francis (2009) reported that the only study to date that has rigorously investigated the use of Spanish versions of content assessments was conducted by Hofstetter in 2003. Her study used grade 8 NAEP mathematics items and involved two samples of students: one set of Hispanic students instructed in Spanish and a second set of Hispanic students instructed in English. A strong positive effect for the Spanish version of the test was found for the students instructed in Spanish, but a moderate negative effect was found for the students instructed in English. Based on this study, the use of native language assessments does not necessarily lead to improved performance on the part of English language learners. Most importantly, unless the language of instruction matches the language used for the assessment, there appears to be little to no gain in English language learner performance. It appears that unless students are familiar, through instruction, with the concepts and terms used in a given language, the use of an assessment in that language does not provide any discernible benefits.

In order to use native language assessments, there are a number of technical and policy considerations that must first be addressed:

- » *Comparability/validity*: Can an assessment that has been translated, or developed in parallel in a different language, be assumed to measure the same underlying construct or set of skills? Can we assume that scores from an assessment in two different languages have comparable meaning? A recent paper by Young (2009) specifies a conceptual framework for test validity research on content assessments taken by English language learners, and identifies eight separate indicators of test comparability, including reliability, internal test structure, DIF, and predictive validity. Similarly, Sireci (2009) has identified several quantitative approaches for evaluating test comparability, including evidence based on internal test structure and differential item functioning.
- » *Test translation/transadaptation*: The technical problem of translating assessments from English into other languages is a perplexing one that must be resolved. In many cases, it is necessary to go beyond directly translating a test from English into another language, by adapting the content of the assessment to account for the sociolinguistic and cultural differences between the two languages (a process known as transadaptation) (e.g., see Stansfield, 2003).
- » *Native languages for assessments*: States have provided assessments in a number of languages including Arabic, Cambodian, Haitian Creole, Portuguese, Russian, Spanish, and Vietnamese, but only Spanish is widely used (Sireci, 2009). The criteria for choosing appropriate native language(s) for assessments are unclear. If one criterion is that a certain

percentage of students must have a given language as their native language before the assessment is created, should that percentage be based on students at the state, LEA, or school level? Determining which assessments should be provided in another language is another consideration. Clearly, one would expect that assessments in reading/language arts should be administered in English, but should native language assessments be made available for all other subjects?

- » *Fairness considerations:* If content assessments are developed for some native languages, but not for others, questions concerning equity and fairness will naturally arise. In addition, testing accommodations may not be equally effective for all students, such that the use of native language assessments may benefit students with higher levels of native language literacy than other students. Furthermore, students may be proficient only in *speaking* their home language, not in reading it, and the language of instruction and the language of assessment must be aligned in order for native language testing to be useful (Abedi, Lord, Hofstetter, & Baker, 2000). Lastly, some states, such as Virginia, do not allow students to be tested in a language other than English. How would conflicts between the potential availability of native language assessments and state policies, such as Virginia's, be resolved?
- » *Financial considerations:* Developing, administering, and validating a native language assessment has been found to be as costly as producing a completely new assessment, even if the native language assessment was developed as a transadapted version of an existing assessment. Test translation is more expensive than translation of other types of documents due to the many additional steps and extensive reviews that must be built into the process (Stansfield, 2003).

In addition to native language assessments, states currently provide other testing accommodations to English language learners, including the use of dual language assessments, English dictionaries/glossaries, bilingual dictionaries/glossaries, native language instructions, response accommodations, scoring accommodations, and simplified language (sometimes referred to as linguistic modification of test items) (Rivera & Collum, 2008; Solano-Flores & Li, 2009; Young & King, 2008). The meta-analysis by Kieffer, Lesaux, Rivera, & Francis (2009) reported that the use of English dictionaries/glossaries was the only accommodation they investigated that showed a significant positive impact on the performance of English language learners. Studies of the use of bilingual dictionaries/glossaries and simplified language have found mixed results, with slightly positive to no impact on English language learners. However, further research on these and other testing accommodations for English language learners is clearly warranted before any definitive conclusions can be drawn about their effectiveness and the impact on the validity of test scores of English language learners. We suggest that consideration be given to these other accommodations for English language learners, in addition to the possible use of native language assessments.

Question on the Assessment of Students with Disabilities

1) Taking into account the diversity of students with disabilities who take the assessments, provide recommendations for the development and administration of assessments for each content area that are valid and reliable, and that enable students to demonstrate their knowledge and skills in core academic areas. Innovative assessment designs and uses of technology have the potential to be inclusive of more students. How would you propose we take this into account?

Researchers at ETS have analyzed the test data for several state assessments and have conducted experimentally designed research studies to examine the impact of testing accommodations on students with disabilities. Research results have shown that current state assessments are unreliable measures for a large portion of students with disabilities because they are too difficult relative to the students' current achievement levels. In some state assessments, the proportion of students without disabilities responding at chance level (or below) is less than 3 percent, but this percentage jumps to 10 percent to 20 percent for students with learning disabilities. In addition to inducing reliability issues for this population, these sorts of tests may have a negative impact on students' emotions and motivation, as well as the ability of the test to accurately measure student growth from year to year.

One possible solution to this mismatch between test difficulty and student achievement level is adaptive testing, which we have recommended as a component of the system for other reasons as well. There are a number of positive reasons for using adaptive testing models with students with disabilities. One of the most important reasons is that such tests provide a better match of the difficulty level of the test to the achievement level of the student. This is important because providing an assessment that is better matched to a student's achievement level will not only result in a more precise estimate of the student's skills, but it will also result in a less frustrating experience for the student. In addition, it may be possible for states to use an adaptive test design to objectively route some students with disabilities to a modified assessment.

One design includes a two-staged adaptive assessment which measures reading comprehension (using a read aloud accommodation) and reading fluency separately for students with reading-based learning disabilities who perform at (or below) chance level on a short routing test. This type of test design has the potential for allowing states to measure proficiency level, while also providing additional information to teachers (scores for two separate components of reading), providing students with test content that is closer to their current achievement level, and allowing a portion of students to use a read aloud accommodation.

Although adaptive testing models have the advantage of targeting the difficulty level of the assessment to the students' current achievement level, there are several disadvantages that are nontrivial for students with disabilities.

- » A potential disadvantage of adaptive testing may be the impact of divergent knowledge patterns in students with specific disability subtypes. For example, many learning disability classifications are defined by divergent cognitive profiles or lower achievement levels in specific academic knowledge areas or subskills. The implication is that students with learning disabilities defined by a deficit in mathematics fluency, for example, may perform poorly on relatively easy test questions that measure calculation but perform well on relatively difficult questions that measure estimation. The use of computer-adaptive tests in the presence of idiosyncratic knowledge patterns has been studied, and results show that scoring of adaptive tests is problematic when a test taker responds to questions in an unexpected way. Additional research would be required to determine the impact of this for students with disabilities.
- » Another disadvantage of implementing an adaptive test is that providing some testing accommodations can be problematic. This is particularly challenging in developing alternate format tests (such as Braille) for item-level adaptive tests because the selection of questions in an item-level adaptive test is based on the specific performance of the test taker on the previous questions. Therefore, it is impossible to assemble a test prior to administration. In addition, many computerized testing platforms do not provide magnification or prerecorded audio, and none of the existing platforms currently provide refreshable Braille. For these reasons, individuals who require Braille test forms do not currently participate in item-level adaptive tests. Instead, these test takers typically take an alternate paper-based linear form of the assessment.

Ultimately, adaptive testing, particularly multistage adaptive testing, holds promise for students with disabilities; however, it is not a panacea. Below are several recommendations for how the Race to the Top assessment program funds could be used to develop the infrastructure for delivering accessible assessments that target test questions to student achievement. If these research studies are conducted during Generation 1 of the program, the issues could well be resolved by the time we move to Generation 2 with large-scale, adaptive testing in the summative component of the system.

- » Specify that some portion of the Race to the Top assessment program funds be devoted toward the development of an open-source computer-based testing platform that is fully accessible to students with disabilities. This is no easy task and will require the collaboration of individuals with experience in assistive technologies (both developers and teachers), universal design of assessment, and the development of existing computer-based testing platforms. In addition, the Department has already invested funds to develop accessible computer-based testing platforms (NAEP Writing and NimbleTools®) so we encourage consultation with colleagues in the National Center for Special Education Research (NCSER), the National Center for Education Statistics (NCES), and the Office of Special Education Programs (OSEP) to build upon their progress.
- » Conduct studies to determine if adaptive tests (particularly item-level adaptive tests) accurately measure the achievement levels of students with disabilities.

- » Do not wait until the common assessment is developed to start planning for the development of alternate standards and alternate assessments, test forms, and test formats.
- » As the Department considers innovative test items and design features, consider the role of graphical material, animations, and other media in tests, and have a plan in place to maximize the adaptability of such materials and/or devise strategies to develop alternative item types to replace graphical types when needed.
- » Conduct research studies to document that the scores on all test forms and formats are comparable (or suggest ways to improve comparability). This is particularly true for any innovative technology-enabled assessments which may be proposed.
- » Develop test content in a format that allows testing vendors to easily render test content in alternate formats such as audio, Braille, and large print. This may involve providing text descriptions of graphics and providing audio descriptions, captions, and text transcripts for movie clips and animations.
- » Consider adaptive test designs which would allow the scores from the common assessments and alternate assessments based on modified achievement standards to be reported on the same scale.
- » Build upon developments in universal design and accessible assessments that have been funded by NCSE, OSEP, and the National Science Foundation (NSF), such as the National Accessible Reading Assessment Projects (NARAP). The NARAP Accessibility Principles for Reading Assessments, which includes supporting research evidence, can be found at <http://www.narap.info/publications/reports/NARAPprinciples.pdf>.

Questions on Technology and Innovation in Assessment

- 1) Propose how you would recommend that different innovative technologies be deployed to create better assessments, and why. Please include illustrative examples in areas such as novel item types, constructed response scoring solutions, uses of mobile computing devices, and so on.

Our response to this question is formulated as a set of recommendations.

Recommendation #1

Start with a long-term vision (5 to 10 years out) for a next generation assessment system and, only then, work backward to a set of steps to get there, including significant near-term ones.

Throughout this document, we have referred to “Generation 1” and “Generation 2” as the two major phases of the implementation of the Race to the Top assessment program. If we start with the definition of Generation 2, we can work backwards into what Generation 1 looks like, and conduct the tryouts and innovation labs for new components in the system during Generation 1, using the interim and formative components of the system.

The reasoning behind this recommendation is that it takes 2 to 3 years to create, review, pilot test, calibrate, and administer a new parallel form of a *paper-and-pencil multiple-choice* test. If 3 to 4 years is the end-state time frame for creating a technology-based, next generation assessment system, then the likelihood of achieving fundamental change is not going to be high.

Recommendation #2

In that long-term vision (and to the extent possible in the incremental steps), focus on such critical ideals as using technology to:

- 2a. *Measure important competencies that cannot be measured well in pencil-and-paper testing.*
- 2b. *Help teachers (and students) adjust instruction and learning.*
- 2c. *Model effective teaching and learning practice, so that the assessment becomes a worthwhile learning experience in and of itself.*
- 2d. *Make assessment fairer for all students, including those with disabilities and English language learners.*

We suggest focusing on ideals because the danger is that, if we do not, worthwhile near-term efficiency targets (e.g., improving score turnaround) may dominate to the detriment of more fundamental goals (e.g., measuring what’s important).

Each of the above subrecommendations, 2a through 2d, deserves further elaboration.

Recommendation #2a: *Use technology to measure important competencies that cannot be measured well in conventional form.* There are many ways in which we can use technology to

measure important competencies. Examples include having students use simulations of dynamic systems to interpret evidence, discover relationships, infer causes, and pose solutions; mathematically model problem situations with a spreadsheet; write on computer and read (nonlinearly) on the Internet; search for, and critically evaluate, information on the Internet; respond to reading or writing problems that require the integration of many text sources and of various document types (including nontext types like video and animation); fluently execute basic procedures (which can offer information that is formatively useful); carry out complex extended projects; and assemble digital portfolios of their work. None of these uses of technology should be done for its own sake but only if it is used to measure important competencies that could not otherwise be assessed.

Recommendation #2b: *Use technology to help teachers (and students) adjust instruction and learning.* When a student's summative test performance suggests the presence of either an overall proficiency deficit or of specific skill deficits, we should at the least provide "formative hypotheses" that point teachers toward students or skill areas of need, upon which teachers (and students) should follow up. As an alternative, we might route the student to a targeted diagnostic assessment.

Recommendation #2c: *Use technology to model effective teaching and learning practice.* We might use technology to model effective teaching and learning practice by building into test questions tools that practitioners use, and that students should be using routinely, in the course of their classroom work. Examples include making planning tools part of writing assessments, embedding into reading comprehension questions graphical organizers and tables for representing complex text (with appropriate alternatives for students with visual disabilities), and asking students to complete concept maps for representing physical or semantic relationships.

Recommendation #2d: *Use technology to make assessment fairer for all students, including those with disabilities and English language learners.* For example, embedding definitional links for difficult words into test questions (where vocabulary knowledge is not being tested) ought to lower irrelevant knowledge requirements for English language learners. A second instance is providing for students with print-related disabilities alternate representations of question components (e.g., translating stimulus text to speech, or describing orally the graphical components of an item). Finally, we can offer alternate questions measuring similar skills at similar difficulty levels, when a class of questions is important but not suitable for some students.

Recommendation #3

Understand the benefits and limitations of each technology before deploying.

All technologies have benefits and limitations. For example, automated scoring is operationally faster and cheaper than human scoring, and sometimes able to provide feedback on instructionally actionable performance components. But in many cases, automated scoring uses limited proxy measures, like

sentence and word length and sentence complexity, to *predict* a human score, and practicing the proxies may lead to higher machine scores but not necessarily to greater learning. A second example can be found in adaptive testing. Adaptive testing measures with precision throughout the skill range. However, in current implementations, it measures only a subset of what is important to test, potentially having the same (unwelcome) effects on instruction as current multiple-choice assessments are said to have. Therefore, it is important to recognize that there are tradeoffs associated with new technology that are best made by informed choice, rather than by accident.

Recommendation #4

Manage risk.

Most successful transitions from paper-and-pencil to computer delivery have put substantial time into planning and many have used a phased approach to implementation. Examples include Oregon and Virginia. Each state now delivers about 1.5 million summative tests annually on computer at the primary and secondary level, including for Adequate Yearly Progress (AYP) purposes. But it took those states the better part of a decade to achieve that level. The main point is that moving a large-scale testing program to computer is a very complex undertaking requiring, among other things, hardware and software availability and compatibility in all schools, extensive LEA training, and student familiarization. Getting the appropriate infrastructure and knowledge into place takes considerable time and effort. Our proposal is to use Generation 1 (up to four years) for tryout, experimentation, and building of infrastructure, and then launch Generation 2 in either the fifth or sixth year, with greater assurance of success. There is no need to repeat the entire decade of development it took the pioneering states, but there is also a danger in assuming that the other states can learn without their own trial and error on a small scale.

Recommendation #5

In the world of innovation, failure is a fact of life but one that can be put to beneficial use, so plan to fail but plan to fail early, often, small, and gracefully.

The value of this type of controlled failure is that it will make clear relatively quickly that an approach is unworkable or, in the best case, help successively approximate over time a practical assessment system with the least cost and harm to all concerned.

Recommendation #6

Fund multiple consortia so that significantly different assessment models (and uses of technology) can be explored and compared to one another, and consider giving preference to models that already have an existing theoretical base and that have been piloted.

The assessment industry knows a lot about how to create innovative technology-based assessments, including ones that ought to have positive effects on learning, so we should build on that existing knowledge. However, we know a lot less about how to create innovative technology-based assessments that are affordable, practical, technically defensible, accessible, and fair to all students, so there is great

value in funding multiple approaches. Therefore, we recommend that the interim and formative systems be based on more than one approach and be tried out in innovation labs in a variety of participating states. The summative system is where we need the greatest level of comparability and standardization, so there would not be a variety of approaches on that system, although some experimentation can take place in the state-specific component of the summative assessment.

2) We envision the need for a technology platform for assessment development, administration, scoring, and reporting that increases the quality and cost-effectiveness of the assessments. Describe your recommendations for the functionality such a platform could and should offer.

We have also formulated our response to this question as a set of recommendations.

Recommendation #1

Our first recommendation with respect to question 2 comes in two parts.

Recommendation #1a: *The platform should support the development, presentation, and scoring of assessments that represent as fully as possible not only the standards, but also the results of cognitive-scientific research because that research can help translate the standards to test specifications and to classroom practice.* The research suggests the need to measure higher-order thinking skills (e.g., conceptual understanding, problem solving, reasoning, critical thinking, strategic thinking), lower-level components (e.g., declarative knowledge, automaticity), and problem-solving processes (which have value for formative purposes). Additionally, the research suggests the need for assessments to model the habits of the mind that are characteristic of proficient performers in the domain. These needs require that the platform have the capability to collect timing data (to measure automaticity), collect keystroke and mouse-click data (to measure problem-solving processes), and integrate tools and performance criteria into test questions so that students learn to use those tools and internalize those criteria in their work.

Recommendation #1b: *The platform should support the development, presentation, and scoring of assessments that purposefully include dynamic stimuli (audio, video, animation), constructed responses of all types (written, spoken, digital representations of artifacts or of performances), simulations (e.g., of physical or social systems), information resources (e.g., Web sites, manuals), and scenario-based, extended exercises calling for the integration of multiple skills and knowledge components. The platform should also support the development, presentation, and scoring of traditional test questions.* This array of competencies and tasks is necessary because the types of tasks encountered, and the competencies required, in workplace and advanced academic settings *cannot* be effectively represented through traditional testing approaches alone. However, traditional approaches do have value for efficiently measuring some types of competencies and should also be included.

Recommendation #2

The platform should support frequent measurement, with the capability to aggregate information over time to form a summative judgment.

Frequent measurement could include multiple summative tests distributed across the school year; one or more standardized projects; electronic portfolios of student work; or combinations of these elements. We recommend that the platform support frequent measurement because we should be able to make more meaningful (and fairer) decisions about students, teachers, schools, and education systems if we combine evidence from multiple time points and from multiple sources.

Recommendation #3

The platform should minimize the influence of irrelevant factors on performance.

To minimize the impact of such factors, the platform should include student tutorials, practice tests, formative assessments, and instructional exercises that use the same interfaces, representations, and tools that are found on the summative assessments. The intention behind this inclusion is to give students multiple opportunities to become familiar with the mechanics and tools used in the test. In addition, the design of the platform should account for the needs of students with disabilities and English language learners in formulating these mechanics and tools. The end goal is to help ensure that test performance depends, to the maximum degree possible, only upon those aspects of student competency that are the intended targets of measurement.

Recommendation #4

The platform should support an advanced type of adaptive testing.

Traditional item-level adaptive tests require short tasks that are machine-scorable in real time. As a consequence, we should look toward new approaches to adaptive testing. One example is a traditional adaptive test section that routes students to an appropriate extended constructed response (ECR) section that, itself, might not be machine-scorable. A second example is a multistage adaptive test in which each stage consists of an extended, scenario-based task including both machine-scorable and ECR items, with routing from one stage to the next based only on the machine-scorable items. The rationale for this recommendation is that adaptive tests can provide more precise measurement than traditional tests for low- and high-performing students but, in their current form, adaptive tests omit the measurement of key competencies.

Recommendation #5

The platform should support online human scoring and automated scoring, as well as their combination.

Online scoring allows for geographically distributed human rating, with real-time monitoring of rater performance. We have discussed these benefits at length in response to question 4 under “General Assessment Questions.” At the same time, significant advances have been made in the automated scoring of many types of constructed responses including essays; short text responses; mathematics

equations, and numerical and graphical responses; and some types of spoken responses. Automated scoring can often provide, in addition, detailed feedback on student task performance. These approaches can potentially make scoring cheaper, faster, and better, especially when online human scoring and automated scoring are employed in combination.

Recommendation #6

The platform should make it easy to switch testing vendors or use multiple vendors.

The platform should represent test questions and automated scoring models in common formats (e.g., SCORM, QTI) so that questions and scoring models can be moved from one vendor's system to a subsequent vendor's system without undue time, cost, and effort. The rationale for this recommendation is that states should be able to make vendor selections without having to bear the cost of repeatedly converting test content and scoring.

3) How would you create this technology platform for summative assessments such that it could be easily adapted to support practitioners and professionals in the development, administration, and/or scoring of high-quality interim assessments?

While we present more recommendations for the platform's development here, we must point out that we do not recommend that a totally new platform be designed or built from scratch for administration of the summative assessment of common-core standards. This will be a high-stakes testing program in many respects, not only because of its accountability uses, but also because it represents the product of a tremendous amount of political will from the states and their stakeholders. It would be a tragedy for it to fail for reasons of technology and operational glitches. Version 1 of any software system usually has many bugs, and the kinds of systems we have described here as being necessary for the administration of the Race to the Top assessment program are very complex, with thousands of parts. Add to that the complexity of the technology infrastructure — computers, routers, servers, Internet connections — that must work together nearly flawlessly for a successful administration, and it becomes obvious that there are many places to fail in the rollout of the Race to the Top assessment program. We strongly recommend that the tests be administered using proven, existing systems, with desired enhancements being provided so that they are open-architecture applications that can be combined with existing platforms, provided those existing platforms are also QTI- and SCORM-compliant, open-architecture systems, as described in the answer to the previous question.

We treat this issue further in the answer to question 4, below.

Recommendation #1

The technology platform should allow for the possibility of making the interim-assessment part of the summative system, as well as for incorporating other sources of evidence like projects and portfolios.

The interim assessments, themselves, should be composed of an even greater mix and variety of innovative and traditional tasks as found on the end-of-year assessment, especially in Generation 1. Even in Generation 2, the need for faster score turnaround, greater comparability, and lower cost will

put more constraints on the summative component of the system than on the interim and formative components. In addition, the interim assessments should incorporate learning progressions, where such progressions are available. Finally, the interim assessments should be constructed such that they are learning experiences in and of themselves, not just tests.

The intention of this recommendation is to distribute the evidence used for summative judgments over additional sources, reducing the influence of a single, end-of-year assessment and employing the same model teachers routinely use to award course grades (i.e., take an average across quizzes, a midterm, final, and other sources). One should note that, in such models, the more interims (or other sources of evidence) there are, the less each counts individually. Interim assessment that is part of the summative system would more frequently model for teachers and students the competencies and tasks that are critical to proficient domain performance, and the learning progressions that are likely to lead there. Finally, such a model would give timely (but preliminary) formative feedback, pointing teachers to students and areas of need on which teachers (and students) should follow up.

Recommendation #2

The technology platform should have the capacity to offer a variety of teacher-optional, curriculum-embedded, formative-assessment materials linked to the standards and to the summative assessments (as embodiments of the standards).

In making this recommendation, it is important to note that interim assessment and formative assessment are distinctly different entities. The interim assessments primarily serve to help identify and document overall student status, whereas the formative measures are intended to provide more specific and targeted information for day-to-day classroom learning needs.

For the formative-assessment materials, the platform should have the capacity to include traditional items targeted at specific component skills; innovative tasks, projects, and portfolios targeted at skill integration, problem solving, reasoning, critical thinking, and conceptual understanding, among others; scoring rubrics to identify characteristics of good performance to teachers and students; exemplar student responses illustrating different score levels; pointers to additional, relevant instructional resources; learning progressions linking items, tasks, and instructional resources to the standards; and, finally, guidelines for teachers on a suggested process for using traditional items and innovative tasks for formative-assessment and instructional practice.

The rationale for this recommendation is that, on its own, interim assessment is insufficient for supporting classroom assessment needs. Teachers (and students) need curriculum-relevant items, integrated tasks, rubrics, interpretive materials, and instructional resources that they can use on a daily basis if they are to focus on, and make progress toward, achieving the standards.

Recommendation #3

The technology platform should have the capacity for teachers to add, modify, and share formative materials.

Teaching contexts and student populations vary, so the ability to customize is important. But many contexts and populations are similar enough that contributions by one teacher may be useful to other teachers, so mechanisms for sharing are critical.

Recommendation #4

The technology platform should have the capacity for teachers (and students) to formatively score constructed responses of all types.

The platform should be able to present rubrics, exemplar responses illustrating score levels, qualification sets (so that teachers and students know how well they are judging responses), and tools for annotating responses and recording scores. The intention behind this recommendation is that teachers and students can develop a shared understanding of what makes for good performance in a domain through scoring, particularly through identifying the features in responses that make those responses of higher or lower quality.

4) For the technology “platform” vision you have proposed, provide estimates of the associated development and ongoing maintenance costs, including your calculations and assumptions behind them.

As mentioned in the answer to the previous question, we do not recommend a new platform be developed for the administration of the Race to the Top assessment program, particularly for the summative assessment of common-core standards. Using existing platforms would not only save on the cost of developing a totally new system, which is not needed, but would also help ensure that the administration will work with less chance of failure. Of course, these systems may need to be updated to meet the needs of the new assessments.

The cost of developing the system would therefore be the cost of developing new enhancements or applications that would bring existing systems into compliance with what is needed to administer the new tests. These costs cannot be estimated until an inventory, or gap analysis, is done between the features and functionality of the desired system and existing test delivery systems. When that gap analysis is done, it will likely be different for each of the different existing systems. For example, Delivery System A from Company A might have gaps that would cost \$2 million to close, while Delivery System B from Company B might have gaps that would cost \$10 million to close. We would have to determine how much of that expense is reasonable to undertake as part of this development project, and how much should be left to the testing providers to close using their own sources of funding.

This issue interacts with the issue discussed in the section below “Project Management Questions,” and question 1 in particular. Our recommendation is for the Department to use the Race to the Top assessment program funds to design, build, research, and test the assessment system, but not to

operationally administer it consortium-wide in the states. The actual administration of the assessment in “live” census administrations should be done by the states using their normal procurement policies in a competitive marketplace. Because we are recommending an entirely computer-administered assessment in Generation 2, it means that states would procure services for the computer administration of the tests by qualified vendors of their choice. Of course, there would have to be some central repository and “keeper” of the assessment content and data, and some process for approving the vendors who are qualified to administer the Race to the Top assessment program. This is not an unusual need, and could be handled in a manner similar to how the GED® Tests are administered by providers authorized by the American Council on Education. In this model, the central repository contractor would be responsible for sending assessment components to authorized vendors in QTI-compliant “ready to use” format.

Assuming that there is a distributed delivery system to allow competition for administration and other services in the states, the Department would have to decide what percentage of the Race to the Top assessment program funds would go toward build-out of the system to facilitate the proper administration of the assessment with all of its attendant educational advantages. The modules developed by the consortium of states using the Race to the Top funds could be designated as open-source, open-architecture modules available to any providers of online test administration.

So, as far as ongoing maintenance costs are concerned, those would be the responsibility of the testing companies providing the services to the states, or the states themselves. Those dollars could come from the Elementary and Secondary Education Act (ESEA) funds, enhanced with states funds, that are normally used for administration of statewide tests required under the ESEA.

Project Management Questions

- 1) Provide estimates of the development, maintenance, and administration costs of the assessment system you propose, and your calculations and assumptions behind them.

The costs of a common-assessment system are very hard to predict in the abstract, as long as certain decisions remain unmade. We recognize that getting approximate estimates of cost are important to policy making and other decisions, but it would not be very helpful to say that the development costs could range from \$10 million to \$400 million, and the operational costs for administration and maintenance could range from \$10 per student to \$200 per student. However, the options we have heard considered and the range of possible decisions about the structure and size of the assessment could actually produce costs in that range. We would be pleased to provide more specific cost information once some decisions have been made.

What we have provided here is a description of how various characteristics of the program would affect cost.

First of all, it is important to separate **development costs** from **maintenance costs** from **administration costs**. While they are strongly related, there are some very important differences. For the purposes of this discussion, we assume that “development costs” include the design and initial creation of the testing system, up to but excluding its first “live” administration (when it is administered statewide in each participating state in the consortium, and the scores count for accountability). “Administration costs” include the live administration and the costs of delivering the assessments to the computers at which the students will be tested, as well as all the support services required, including constructed response scoring, customer service, and psychometric work and analysis to produce scores and summaries that take place each year the test is administered. “Maintenance costs” include maintaining and refreshing the item bank and keeping the software systems up to date.

Development Costs

Development costs depend on the type of assessment, number of grades and subjects, and number of forms to be developed. The number of students to whom the assessment will be delivered is not a major factor in developing paper-and-pencil tests, but it would influence the number of items needed in a computer-administered test if the ratio of students to computers is so high that it calls for an extended testing window. We must also consider the number of students needed for pilot/field testing, which depends on the pilot/field test design, assessment design, and type of administration is a factor. Other significant factors in development costs are associated with meetings, honoraria, and travel for teacher committees, administrators, and test developers. We have seen very large differences in the development costs for state assessments based on the state’s preference for how many meetings of in-state professionals are required. Can these meetings be conducted using technology, or do they have to be face-to-face? Are there meetings of special groups like bias committees and other stakeholders?

While the testing company can accomplish many of these activities highly effectively outside the state, frequent local meetings provide the great benefit of getting buy-in from in-state stakeholders.

Item tryouts can be very costly if they require a special testing occasion, which is likely during the initial development phase of the Race to the Top assessment. In order to get the best data, the students should be tested as close to the time of year as possible as when the assessment will be administered in a live program. That process could be quite intrusive in the Spring of 2011 and 2012, when states are still administering their existing assessment programs. Once the test is in operation, new items can be tried out by embedding them in the live assessments as they are administered online, but it is likely that the first phases of this new assessment would require a special administration and accompanying costs. Another factor to keep in mind is the difference between the new common standards and the standards previously in place in the states. If they are different enough, there will be an effect on student performance coming from the teachers' ability to teach the new standards and the availability of instructional materials and resources. This might cause the need for recalibration of the item bank over time.

A significant driver of development and maintenance costs is the size of the item bank. The larger the item bank, the more items there are that need to be developed, which increases cost. Factors that determine the size of the item bank include obvious ones, such as the number of standards that are measured and the number of items per standard that are needed for reliable scores (which in turn depends on the level of detail desired in score reports) — but also less obvious ones, such as the length of the test administration window and the item release strategy. The longer the testing window (because states test in different weeks or have to spread out testing over many days because of lack of access to computers), the more items are needed to protect security. The larger the portion of the test item bank that is released each year after testing, the more items needed to be developed to keep a live item pool for building the next year's pool.

Other factors that impact cost are the type of stimuli used and whether they are proprietary (i.e., requiring payment of fees for use). Many state assessments demand the use of published literature for reading comprehension passages, for example, and this demand involves a permissions process and payments of sometimes large fees for the use of that intellectual property.

Administration Costs

For administration costs, the number of students being assessed is a primary driver, as is the type of assessment (e.g., scorable by computer or requiring human scoring), and method of administration (computer-administered versus paper-and-pencil). Because we are proposing that the entire Race to the Top assessment be computer-delivered in Generation 2, except for special accommodations for certain students, we will not treat paper-and-pencil costs here, but would like to point out that they can be significant, especially for special forms like Braille and large print.

During Generation 1, we will likely have to allow for some states and even some LEAs to administer the test via paper and pencil. This introduces another whole set of costs to the system: composition,

printing, packaging, distribution, retrieval, document staging, scanning, and editing. The cost of a dual paper- and computer-delivered system is quite high, but might be unavoidable for at least a few years.

It costs more to have people read and score test results than to use a computer algorithm to score the tests. While automated scoring of constructed response items is possible and is widely used for certain item types, it is still not sufficiently developed to score all kinds of items and tasks likely to be required on the type of assessment we propose. We assume that at least part of the assessment would need to be scored by trained human raters, and that some of those raters be teachers. The size of the labor effort and the level of expertise needed by the raters is a significant factor in the cost. If the test is given online, much expense is saved in presenting the responses to the raters for scoring in an automated, or even distributed, system, but this is still a major cost.

Maintenance Costs

In many ways, the maintenance costs are driven by the ongoing decisions about item pools and assessment administration modes. The costs that drive those factors will obviously affect the ongoing maintenance costs of the system.

Estimating the cost for state-administered summative assessments is a complex task. We would recommend that the Department, in the models it promotes and funds, take full advantage of the competitive free-market system that exists today to spur innovation and drive down costs to states for test administration services.

Many think that a common assessment across, say, 30 states would greatly reduce the testing costs for those states. While this might be true for some, but not all, development costs, it is not true at all for administration costs. As a matter of fact, a common assessment might result in increased administration costs under certain conditions. If, for example, having a common assessment forced a narrower testing and scoring window across the participating states, this might force testing companies to increase their capacity for scanning documents in a system that relied on paper as well as computers during Generation 1, because more documents would be scanned in a shorter time. These increased costs for infrastructure would have to be passed on to states. If states wanted to keep their existing, spread-out testing windows, the current infrastructure is adequate, but that might require the development of more alternate forms to protect test security, which would increase development costs.

We are recommending that the entire program be computer-administered by Generation 2 and as soon as possible in Generation 1, except for some students with special needs. This creates different cost variables. Printing, shipping of paper, and scanning and paper handling are all but eliminated, which saves money, but there are additional infrastructural costs associated with the computer administration that can be significant. Again, we recommend that the states doing the common assessment create a structure that allows for multiple companies to provide test administration services for online administration, competing in the market to lower costs and improve quality and service. If this were the model, states can be sure to get the best service at the lowest cost, and also take advantage of innovations in delivery methods over time.

2) Describe the range of development and implementation timelines for your proposed assessment system, from the most aggressive to more conservative, and describe the actions that would be required to achieve each option.

A program as critical and complex as the Race to the Top assessment program must be designed and developed on a timeline that balances the need for an aggressive implementation plan and the necessity for a resulting assessment that is both valid and reliable. In addition, the timeline must also support input and involvement from a variety of stakeholders both from within the consortium of states and from the Department.

We are proposing that the system take place in two “Generations,” as follows:

Generation 1

The end-of-year assessment consists only of machine-scorable items, or very limited use of human-scorable items. The end-of-year assessment should be administered on computer where LEAs are ready to do so, or on paper otherwise. Note that this is a decision that needs to be made after careful deliberation. On the one hand, allowing for dual (paper-based and computer-based) administration would increase the participation rate in the new assessment system by states not ready to switch to computer-based testing, but it would also have negative consequences. The use of any paper-based testing in the end-of-year component in Generation 1 would have three limiting effects: (1) it would eliminate the possibility of using non-multiple-choice item types that could still be scored using automated scoring technology; (2) it would slow turnaround time of results because of the need for shipping, scanning, and scoring, and (3) it would cost extra money — both for the extra costs associated with producing, shipping, retrieving, and scanning a paper test and the extra costs of double production, the creation of extra forms, and the need to equate the two versions — that could be better spent on advancing other parts of the system. Because the first live administration of the Race to the Top assessment would not be until 2013 under the most aggressive timelines, our recommendations would be for states and the Department to make computer-based testing a condition for participation, and accept the consequences of some states deciding to wait it out. However, we are aware that many states feel strongly about having a paper-based option for Generation 1.

The interim-assessment system should be completely computer-administered, so that a variety of item types are possible. The interim system would also be computer-adaptive in some components that lend themselves best to that mode of administration. Data from the interim administrations is accumulated and added to the summative system for accountability purposes.

The formative-assessment system is computer-administered or paper-based, at the discretion and convenience of the teacher. The formative system is entirely for the benefit of the teacher and the learner, and data are not collected for any accountability purpose. Innovation labs are encouraged and funded throughout the participating states to try out new technologies and item types for eventual promotion into the interim system and then, perhaps, the end-of-year assessment for Generation 2.

Generation 2

The end-of-year assessment is computer-adaptive in all cases and consists of items that are scored by computer for the most part, including constructed response items. The interim system is the same as in Generation 1, and the formative system continues to be independent from restrictions, other than good assessment practice and alignment with the standards.

Timelines

We are assuming that the key part of the summative-assessment system would be administered in the Spring of the year, and that participating states would need some flexibility in the timing of administration that might result in a testing window from March through May. As pointed out elsewhere, the longer the testing window, the more items need to be developed to protect security, which increases cost. If the states could agree to a narrower testing window (e.g., April to May), we would have some relief in the development schedule in the more aggressive timelines. Also, if there would be an option of paper-based testing, all of these timelines would have to be pushed back by a year, because of an additional 4- to 6-month window required for preparation, printing, shipping, and other efforts. While the paper-based scenarios could be absorbed in the 2014, 2015, and 2016 timelines, those scenarios would be impossible in the 2013 timeline.

Also note that the 2013 timeline could not accommodate any interim assessments as part of the summative system in the first year of the administration of the Race to the Top assessment.

In **Figure 1**, we offer four potential timelines leading up to the first administration of the Spring assessment in 2013, 2014, 2015, or 2016. We strongly recommend that the 2013 timeline be considered unrealistic and that the earliest possible administration would be 2014 — and even that schedule is extremely aggressive. The reason we are presenting a timeline for a 2013 administration is to simply point out the steps that are necessary and show how much those steps have to be curtailed to achieve an administration date that soon.

Note that there are parallel tracks in each timeline for the development of content and the development of the technology systems to accommodate the new assessment. We have pointed out elsewhere that we recommend a system for administration that uses existing technologies for computer administration of the assessments, but there would still need to be some customization and additions to these systems. We believe that customization can be done on a parallel timeline with the content development. We should consider the content development path the critical path in this project, and it would dictate the amount of time that would be available for the technology work.

Below are descriptions of the steps we outlined in **Figure 1**:

- » **Formation of assessment design and technology design teams** — include meetings with representatives from all states within the consortium to evaluate designs and needs
- » **Assessment design** — include meetings with representatives from all states within the consortium

- » **Creation of test blueprints and test item specifications** — include review and input from stakeholder groups representing each state within the consortium
- » **Development of items** — include review and input from stakeholder groups representing each state within the consortium
- » **Assessment of technology infrastructure** — work with each state to determine technology gaps for online delivery of assessments
- » **Design and implementation of a plan to close technology gaps** — include a plan for schools across all states in the consortium and determine phase-in plan for online testing
- » **Design and customization of the online delivery platform** — work with the consortium to develop the online delivery system
- » **Provision of training to staff and students throughout the consortium states** — include training on the assessment design and content and on the online platform
- » **Integration of online testing platform** with other data systems within consortium states
- » **Small-scale pilot test of items and delivery platform** — administer items via the online system to small numbers of classrooms throughout the consortium and analyze
- » **Field testing of forms** — include creation, distribution and retrieval of documents for paper-and-pencil field-tests and/or distribution via the online platform for computer-delivered tests, scoring, and analysis of results
- » **Creation of operational assessments** — include review, reconciliation, and approval from stakeholder groups representing each state within the consortium

Figure 1: Potential Timelines for administration of the initial Race to the Top assessment in the Spring of 2013, 2014, 2015 or 2016

Timelines for implementation of RTTT Assessment

2013 Administration			2014 Administration			2015 Administration			2016 Administration	
Content	Delivery System		Content	Delivery System		Content	Delivery System		Content	Delivery System
Form assessment and technology design team		Oct-10	Form assessment and technology design team							
		Nov-10								
Assessment Design	Assessment of Technology Infrastructures	Dec-10	Assessment Design	Assessment of Technology Infrastructures						
		Jan-11								
Creation of Blueprints and Test Design Specs	Design and Implement Technology Gap Closing	Feb-11	Creation of Blueprints and Test Design Specs	Design and Implement Technology Gap Closing	Creation of Blueprints and Test Design Specs	Design and Implement Technology Gap Closing	Creation of Blueprints and Test Design Specs	Design and Implement Technology Gap Closing	Creation of Blueprints and Test Design Specs	Design and Implement Technology Gap Closing
		Mar-11								
Development of Items (including reviews)	Design and Customize the Online Delivery Platforms	Apr-11	Development of Items (including reviews)	Design and Customize the Online Delivery Platforms	Development of Items (including reviews)	Design and Customize the Online Delivery Platforms	Development of Items (including reviews)	Design and Customize the Online Delivery Platforms	Development of Items (including reviews)	Design and Customize the Online Delivery Platforms
		May-11								
Train Staff and Students	Integrate Online Testing Platform with Other Systems	Jun-11	Train Staff and Students	Integrate Online Testing Platform with Other Systems	Train Staff and Students	Integrate Online Testing Platform with Other Systems	Train Staff and Students	Integrate Online Testing Platform with Other Systems	Train Staff and Students	Integrate Online Testing Platform with Other Systems
		Jul-11								
Small Scale Pilot		Aug-11	Small Scale Pilot							
	Nov-11									
Field Test Forms		Dec-11	Field Test Forms							
	Jan-12									
Create Operational Assessments		Feb-12	Create Operational Assessments							
	Mar-12									
LIVE ADMINISTRATION		Apr-12	LIVE ADMINISTRATION		LIVE ADMINISTRATION		LIVE ADMINISTRATION		LIVE ADMINISTRATION	
	May-12									
		Jun-12								
		Jul-12								
		Aug-12								
		Sep-12								
		Oct-12								
		Nov-12								
		Dec-12								
		Jan-13								
		Feb-13								
		Mar-13								
		Apr-13								
		May-13								
		Jun-13								
		Jul-13								
		Aug-13								
		Sep-13								
		Oct-13								
		Nov-13								
		Dec-13								
		Jan-14								
		Feb-14								
		Mar-14								
		Apr-14								
		May-14								
		Jun-14								
		Jul-14								
		Aug-14								
		Sep-14								
		Oct-14								
		Nov-14								
		Dec-14								
		Jan-15								
		Feb-15								
		Mar-15								
		Apr-15								
		May-15								
		Jun-15								
		Jul-15								
		Aug-15								
		Sep-15								
		Oct-15								
		Nov-15								
		Dec-15								
		Jan-16								
		Feb-16								
		Mar-16								
		Apr-16								
		May-16								

3) How would you recommend organizing a consortium to achieve success in developing and implementing the proposed assessment system? What role(s) do you recommend for third parties (e.g., conveners, project managers, assessment developers/partners, intermediaries)? What would you recommend that a consortium demonstrate to show that it has the capacity to implement the proposed plan?

A successful consortium of states needs a detailed, well-defined organizing structure with a strong management plan. States should enter into an agreement or memorandum of understanding that explicitly lays out the consortium’s governance. We recommend that one state serve as the “lead state” to act as the fiscal agent to manage the grant money, and help ensure that the consortium meets all legal and regulatory requirements based on the states’ statutes and regulations. Only a state that has met all of the Race to the Top requirements should act as the lead state.

The consortium should establish a board with an internal organizing structure, clear voting rules, and defined operational procedures to manage their activities. The leadership should be collaborative and emphasize consensus building among the peer states. Regular timelines and meeting dates should be established for each phase of the effort with an annual evaluation of progress, with results shared among the state membership.

The board would select a project management team to work with the states and third party providers to manage the system components, such as grades 3 through 8 assessment development, high school assessment development, delivery and scoring/reporting, and outreach and professional development. The management team would also work with LEAs on a quality management and implementation plan, as well as a plan for how to engage the public and address concerns related to the new assessment system. Management team staff members should have a record of success in their area of expertise.

Third parties who support the consortium (technical consultants, assessment developers and psychometricians, operations contractors, and professional development support staff) need to have proven themselves in the marketplace as competent and successful. Each one should have a well-defined role by the consortium, and each should have the responsibility of sharing their activities with the entire third party team as well as the states so that all are aware of the totality of the activities.

As we mentioned previously, we believe the Race to the Top assessment program should be created in a way that encourages innovation, quality and service, and low cost from a competitive marketplace that is consistent with the procurement policies of the states. There is more than one way to accomplish this, but we would like to propose the following method:

1. The consortium is organized with a Lead State and Member States.
2. The Lead State, with the approval of the Member States, does a procurement to choose an Organizing Entity for the consortium. The Organizing Entity would be responsible for managing the business of the consortium, scheduling meetings, disseminating information, and other major duties.

3. The states in the consortium, with the help of the Organizing Entity, would write and submit the application for funding under the Race to the Top assessment program, following the Department's process.
4. Simultaneous with the submission of the application to the Department, the Lead State would initiate another procurement (e.g., Request for Proposal or RFP) for contractors needed to perform the work the consortium proposed in the Department application. The process would be managed by the Organizing Entity, which would be prohibited from bidding on any work in the RFP, so that the Member States had appropriate input into the choice of the contractors. We would recommend that contractors be permitted to bid on parts of the RFP, or the whole RFP, so that the consortium would eventually get a set of contractors who are the best at the work required, whether it is development, administration, psychometrics, technology, special forms, or other specialties. The award of contracts that results from this process would be contingent upon the approval of the application for Race to the Top assessment program funding by the Department. Note: It would make more sense if consortia bid with groups of contractors preselected, so that the Department could get a complete technical proposal with the application. However, we assume that many states' procurement laws would not allow this.
5. Upon award of Race to the Top assessment program funding by the Department to the consortium, contracts would be finalized with the successful contractors and the work would begin and be conducted according to the process described in the application and approved by the Department.
6. At the appropriate time in the development process, the consortium would do another competitive procurement, again under the auspices of the Lead State, to choose a Maintenance Contractor. The role of the Maintenance Contractor would be to keep the item bank and test bank secure, refresh the item bank and test bank, develop and maintain a process for authorizing Approved Assessment Providers, and interact with Approved Assessment Providers to help make sure they have what they need to deliver the assessments to the states in the consortium and to new states joining the consortium over time.
7. When the system goes "live," states would use their normal procurement processes to choose Assessment Providers. The states would add to their programs additional, state-specific elements, which would be provided by the Approved Assessment Providers. The Approved Assessment Providers would conduct all the activities for which the state needs a contractor, and would have access to the current year's forms of the Race to the Top assessment from the Maintenance Contractor.

We believe that a process like we have described here provides the best quality, consistency, and fidelity to the purpose of a common assessment, while permitting healthy and robust competition and innovation in the field.

References

- Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice*, 19(3), 16-26.
- Butler, F., & Stevens, R. (1997). *Accommodation strategies for English language learners on large-scale assessments: Student characteristics and other considerations*. CSE Tech Report No. 448. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.
- Hanover Research Council. (2008). *Measuring the participation rate of high school graduates in postsecondary education: A state-by-state analysis of data collection methodologies and college enrollment rates*. Washington, DC: Prepared for College Summit.
- Hawkins, D. A., & Lautz, J. (2005). *State of college admission*. Retrieved November 10, 2009, from the National Association for College Admission Counseling Web site: <http://www.nacacnet.org>
- Hofstetter, C. H. (2003). Contextual and mathematics accommodation test effects for English-language learners. *Applied Measurement in Education*, 16, 159-188.
- Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners: A meta-analysis on effectiveness and validity. *Review of Educational Research*, 79, 1168-1201.
- Kobrin, J., Patterson, B., Shaw, E., Mattern, K., & Barbuti, S. (2008). *Validity of the SAT® for predicting first-year college grade point average*. College Board Research Report No. 2008-5. New York, NY: College Board.
- National Clearinghouse for English Language Acquisition. (2007). *The growing numbers of limited English proficient students: 1995-96 – 2005-06*. Retrieved November 10, 2009 from: http://www.ncele.gwu.edu/stats/2_nation.htm
- Pitoniak, M., Young, J. W., Martiniello, M., King, T., Buteux, A., and Ginsburgh, M. (2009). *Guidelines for the assessment of English language learners*. Princeton, NJ: Educational Testing Service.
- Rivera, C., & Collum, E. (Eds.). (2008). *State assessment policy and practice for English language learners: A national perspective*. Mahwah, NJ: Erlbaum.
- Sireci, S. G. (2009). Validity issues and empirical research on translating educational achievement tests. In P. C. Winter (Ed.), *Evaluating the comparability of scores from achievement test variations*. Washington, DC: Council of Chief State School Officers. [Draft report dated June 17, 2009.]
- Sireci, S. G., Li, S., & Scarpati, S. (2003). *The effects of test accommodation on test performance: A review of the literature*. Center for Educational Assessment Research Report No. 485. Amherst, MA: University of Massachusetts.
- Solano-Flores, G., & Li, M. (2008). Examining the dependability of academic achievement measures for English language learners. *Assessment for Effective Intervention*, 33(3), 135-144.

- Stansfield, C. W. (2003). Test translation and adaptation in public education in the USA. *Language Testing, 20*, 189-207.
- Young, J. W. (2009). A framework for test validity research on content assessments taken by English language learners. *Educational Assessment, 14*(3-4), 1-17.
- Young, J. W., and King, T. C. (2008). *Testing accommodations for English language learners: A review of state and district policies*. College Board Research Report No. 2008-6 and ETS RR-08-48. New York, NY: College Entrance Examination Board.



MEMORANDUM

To: U.S. Department of Education
From: Alliance for Excellent Education
Date: December 2, 2009
Subject: Race to the Top Fund Assessment Program (Doc ID: RIN 1810-AB09)

The Alliance for Excellent Education (the Alliance) would like to thank the U.S. Department of Education (ED) for encouraging the state-led common standards process through the Race to the Top Assessment program. As you know, the Alliance has been working on the common standards process with the National Governors Association and the Council of Chief State School Officers, among other organizations. The Alliance appreciates the work ED has done to coordinate with and supplement the common standards process.

Thank you also for your efforts to improve the current assessment system and for reaching out to experts to determine how to best make these changes. The Alliance is committed to the goal of creating common core standards that provide evidence-based, internationally benchmarked statements of what all students need to know and be able to do in order to be college and career ready and we firmly believe that an aligned assessment system is a critical component of this goal.

Although many experts will surely weigh in on the technical aspects of an optimal assessment system, the Alliance's comments below focus on the needs of high schools in such a system. Also discussed is the need to implement these assessments well so that they can be used as teaching and learning tools.

Special Needs of High Schools in an Assessment System

End-of-Course versus Comprehensive Assessments

In the federal register notice, ED asks for recommendations on the use of end-of-course assessments versus comprehensive assessments of college and career readiness. The Alliance would like to stress that these two types of assessments are not necessarily mutually exclusive and can serve complementary purposes in some assessment systems. For instance, end-of-course exams could be used to assess a broader and deeper range of standards in a particular subject, while a comprehensive assessment could be used to benchmark whether a student is on track for college and career readiness upon graduation.

Non-Tested Grades and Subjects

ED did not specify in its notice whether it is planning to revise assessment requirements for non-tested grades and subjects. If these assessment requirements will be changed, the Alliance urges ED to take the following into consideration:

1. The current system of testing only once during high school makes it difficult to measure growth at the high school level.

2. The distinct nature of many high school courses (e.g, students typically take biology or algebra rather than tenth-grade science or ninth-grade math) makes it difficult to measure growth using only one assessment per year.

Because of these challenges of measuring student growth at the high school level, experts have suggested increasing the number of assessments at the high school level or using alternative measures of achievement. Some suggestions for such measures include performance on end-of-course assessments, rates at which students are on track to graduate from high school, interim assessment performance, and enrollment and performance in Advanced Placement courses. The Alliance would like to stress that any alternative measures used should be rigorous and comparable across classrooms. If alternative measures are being used for accountability, these measures should not provide incentives for teachers or other school and district personnel to lower expectations in order to improve passing scores.

Formative and Performance-Based Assessments

At the ED forum on November 13, a comment was made that formative and performance-based assessments should be included in a new assessment regime. The Alliance would like to encourage ED to carefully consider the appropriate federal role for these types of assessments. The primary federal government role in the formative assessment process is to provide incentives for implementation, such as supporting training for current and future teachers in how to align these assessments to common standards.

Performance-based assessments can be a useful tool for providing a more comprehensive understanding of students' knowledge and skills, particularly twenty-first-century skills like critical thinking and analysis. Similar to formative assessments, the appropriate federal role for performance assessments is supporting professional development, pre-service training, etc. The Alliance feels that it is important to ensure that assessments used for accountability purposes are uniform, objective and measurable.

Implementation Issues Related to Assessment Adoption

For as long as the Alliance has participated in the common standards process, it has emphasized that attention must be paid to the human capacity required to implement these standards well. It is critical that teachers and other school personnel be trained in how to access, analyze, and interpret assessment data and use it to improve instruction. It is also imperative that the data be made available in a transparent, timely, and user-friendly format. This type of training should be provided in teacher education and training programs, induction programs, and professional development. As demonstrated through evidence from systems in other countries, the involvement of teachers in the development and scoring of assessments provides them with a clearer sense of the standards and the instructional practices that can support them.

Other General Notes About Creating an Assessment System

Common Assessments or Comparable Assessments

Part of the goal of the common standards process has always been to be able to better compare results across states. If ED is leaning towards clusters of states developing multiple assessments, as discussed at the Boston hearing, it is critical that the tests and cut scores are comparable.

Role of the National Assessment of Educational Progress (NAEP)

The Alliance recommends that ED consider the role that NAEP could play in a future assessment system based on common standards, with particular attention to the historical information that NAEP provides.

From: Petska, Stephanie DPI [Stephanie.Petska@dpi.wi.gov]
Sent: Wednesday, December 02, 2009 8:47 AM
To: Race To The Top Assessment Input
Cc: Russell, Lynette K DPI; Petska, Stephanie DPI; Berndt, Sandra DPI; Kubinski, Eva M. DPI
Subject: Race to the Top Assessment Program
Attachments: Recommendations regarding Race To The Top Assessment Requirements and Students with Disabilities.docx

Submitter: Dr. Stephanie Petska, Director of Special Education, Wisconsin Department of Public Instruction

Title of Document: Input on Assessment of Students with Disabilities

Topic Addressed: Input on Assessment of Students with Disabilities

Text of Email:

One of the important and potentially ground-breaking opportunities posed by the Race to the Top Assessment competition is to ensure that Students with Disabilities have an equal chance to show what they have learned. As the Wisconsin Director of Special Education, I have seen the benefits when consideration of the needs of students with disabilities is part of the development of high-stakes assessments from the start to the end. Wisconsin is fortunate to have a collaborative relationship between our state's Special Education Team and our Office of Education Accountability. Without that relationship, the needs of students with disabilities can be overlooked or not addressed. Attached you will find several points that I believe are essential to ensuring the development of an equitable and effective assessment process, where all students are represented.

Name of Submitter: Dr. Stephanie Petska, Director of Special Education,
Wisconsin Department of Public Instruction
stephanie.petska@dpi.wi.gov

Title of Document: Input on Assessment of Students with Disabilities

Topic Addressed: Input on Assessment of Students with Disabilities

Introduction to Submission:

One of the important and potentially ground-breaking opportunities posed by the Race to the Top Assessment competition is to ensure that Students with Disabilities have an equal chance to show what they have learned. As the Wisconsin Director of Special Education, I have seen the benefits when consideration of the needs of students with disabilities is part of the development of high-stakes assessments from the start to the end. Wisconsin is fortunate to have a collaborative relationship between our state's Special Education Team and our Office of Education Accountability. Without that relationship, the needs of students with disabilities can be overlooked or not addressed. Below are suggestions that I believe are essential to ensuring the development of an equitable and effective assessment process, where all students are represented.

Require Universal Design

- Ensure that states/consortia develop an assessment accessible to all students.
 - Would remove the need to develop an alternate assessment aligned with modified achievement standards (also known as the 2% Test)
- Require accessible item development.
 - Specify allowable accommodations if needed.
- Work from clear, grade-level content standards.
- Consider the use of technology to enhance student ability to demonstrate what they know.
- Need to continue promoting the new generation of assessments that promote high expectations.

Require Considering Students with Disabilities (at ALL ability levels) in the Design of the Assessment

- Do not let the assessment of students with disabilities be an afterthought as so often in the past.

Require Concurrent Development of Alternate Assessment Aligned with Alternate Achievement Standards (AA-AAS or the 1% Test)

- There still continues to be a need for a test for students with significant cognitive disabilities.
- Alternate Assessment Aligned with Alternate Achievement Standards (AA-AAS) needs to be developed at the same time as the general assessment, and needs to require high expectations for students with significant cognitive disabilities.

- Alternate Achievement Standards need to be developed concurrently with the Common Core Standards and should be the basis for the AA-AAS.
- Consider the use of technology to enhance student ability to demonstrate what they have learned.

Prohibit Out of Level Testing

- Students who are tested at grade level are more likely to be taught grade level content, and have a better chance of being successful.
- Avoid lowered expectations for students with disabilities.

December 2, 2009

The Honorable Arne Duncan
Secretary
US Department of Education
440 Maryland Avenue SW
Washington, DC 20202



*Excellence and Equity
in Public Education
through School Board
Leadership*

Re: *National School Boards Association Response to Notice of Public Meetings and Request for Input to Gather Technical Expertise Pertaining to a Possible Race to the Top Program; published in the Federal Register on October 23, 2009*

The National School Boards Association (NSBA) representing over 95,000 local school board members through our state school boards associations across the nation is pleased to offer our comments regarding the proposed announcement on competitive federal grants to support a consortia of states regarding jointly developed common assessments. This program, if established, would provide approximately \$350 million, with at least 50 percent of the awards to the states to be used to provide subgrants to local educational agencies (LEAs).

Our comments do not recommend specific systems of assessments, but rather reflect our general thoughts regarding the development of assessments by a consortia of states, the appropriate role of the federal government, and the Department's framework in supporting LEA-level activities that are designed by the state consortium to support the development and implementation of its assessment system.

While NSBA remains strongly opposed to any efforts to develop, encourage, or require a single national assessment for accountability purposes, we do applaud the Department's proposed actions to establish competitive federal grants specifically designed to encourage consortia of states to develop assessments, and your acknowledgment that such efforts are appropriate for the states not the federal government. We concur with your position that the appropriate role for the federal government is to increase incentives to states and LEAs to create constructive remedies, and provide technical support to the states in developing those standards.

We also believe that in establishing incentives for consortia of states, such competitive grants should also be available to individual states. As proposed, a single state with a broad range of academic challenges, a highly diverse student population, and a demonstrated commitment to a high quality system of assessments meeting all requirements for the grants *except for its participation in a consortium* would be ineligible. Further, the size of the consortia should not matter so that those with fewer participating states and those with larger numbers of participating states could compete equally based on the quality of their proposals addressing both alignment with state standards and the rigor of the curricula.

The proposed requirement for states to award at least 50 percent of the awards to LEAs (including public charter schools identified as LEAs under state law) is very encouraging. As you are aware, research continues to indicate that there is strong consensus among states to ensure rigorous standards, strong curricula aligned with those standards, and valid and reliable systems of

Office of Advocacy

- *C.H. "Sonny" Savoie
President*
- *Anne L. Bryant
Executive Director*
- *Michael A. Resnick
Associate
Executive Director*

The Honorable Arne Duncan

December 2, 2009

Page 2

assessments that fairly and accurately reflect the performance of students, schools and school districts.

NSBA supports the proposed general framework regarding summative assessments that measure individual student achievement and individual student growth. We also support your position that assessments need not be limited to a single end-of-year assessment but could include multiple summative components administered at different points during the school year. Additionally, NSBA generally supports the proposed required and desired characteristics, recognizing that a system of assessments could be effective without necessarily possessing *all* the characteristics since the system needs to be tailored to the needs of the consortia of the states.

Further, we support the need for active involvement of the LEA in the design of assessment systems, and encourage the Secretary to clarify that the types of activities identified in the Notice are representative, but not necessarily required by every participating LEA. We also recommend that LEA activities should be framed around specific criteria and outcomes rather than administrative processes. As an example, the Notice proposes among the LEA activities, the development of a rollout plan for implementation of the standards and assessments together with all of their supporting components. While this may be appropriate for some LEAs, the role of LEAs could vary widely depending on the specific academic challenges being addressed, the capacity of the LEA and the extent to which LEA funds are invested.

Finally, as the Department develops final criteria and requirements to implement this new competitive grants program to support assessments based on state-developed standards, we urge the Secretary to ensure that such requirements compliment the provisions in the reauthorized *Elementary and Secondary Education Act* (ESEA), and that requirements in this competitive grant program would not become a condition for receipt of grants in programs that are designed for other purposes.

We very much appreciate the opportunity to comment. Questions concerning our comments may be directed to Reginald M. Felton, director of federal relations at 703-838-6782; or by e-mail at rfelton@nsba.org.

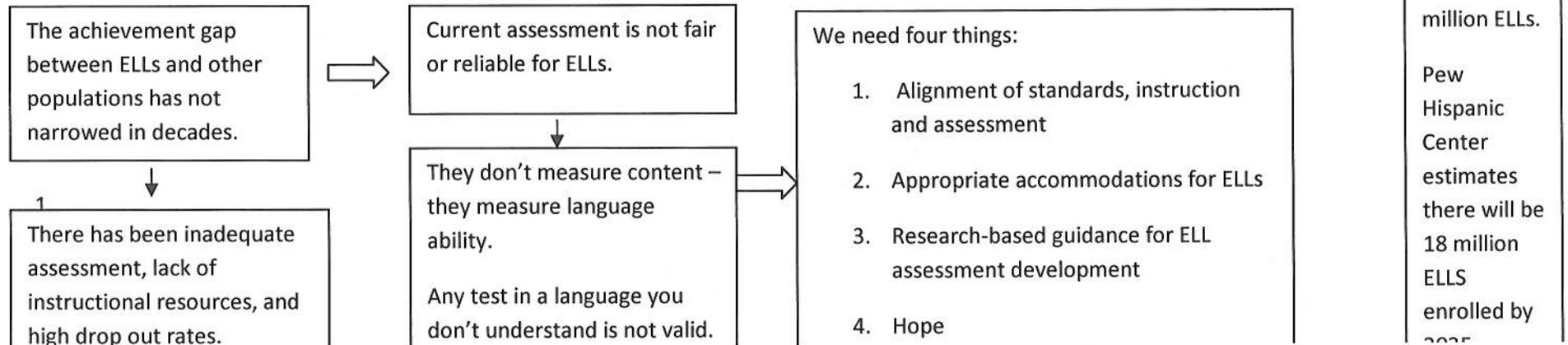
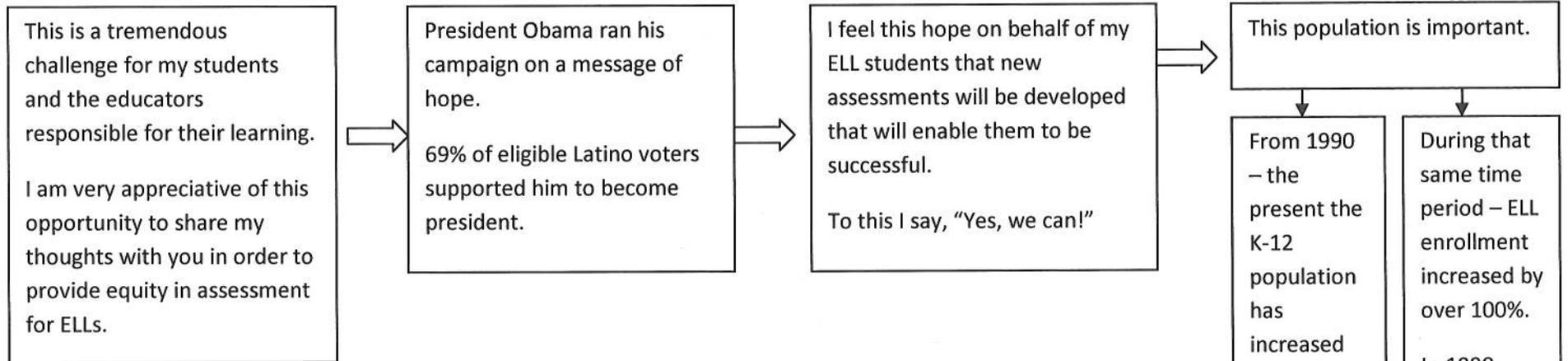
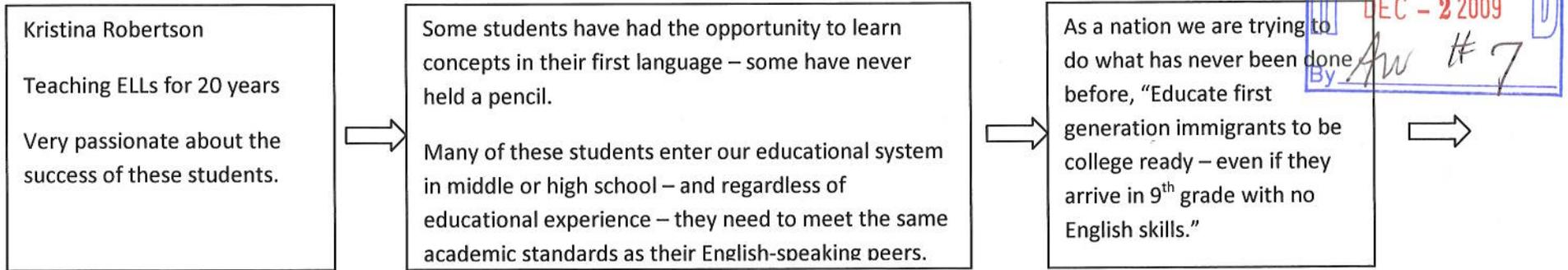
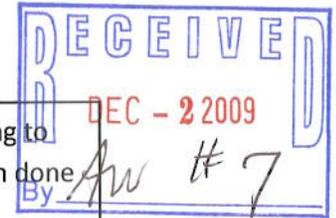
Sincerely,



Michael A. Resnick
Associate Executive Director

MAR: rf/kc

G:Adv/Regulations/2009/RTTTAssessments



We need to measure English language proficiency *and* content assessment.

We need to promote quality instruction and reliable data for ELLs.



Alignment is especially important for students with interrupted education. They are way below grade level and educators need accurate information to inform instruction and accelerate student learning.



We need:

- English Language proficiency standards aligned to English Language proficiency *and* content assessments.
- Implementation of uniform, standardized tests of English Language proficiency such as the WIDA consortium.
- Evaluate current English language proficiency assessments and increase training and resources.

If language level is not accounted for in high stakes testing then results are not valid and it will lead to negative consequences in addition to preventing ELLs ifrom meeting their educational goals.



We need to change:

- Test procedures
- Test materials
- Testing environments



Current testing accommodations were developed for Special Education students.



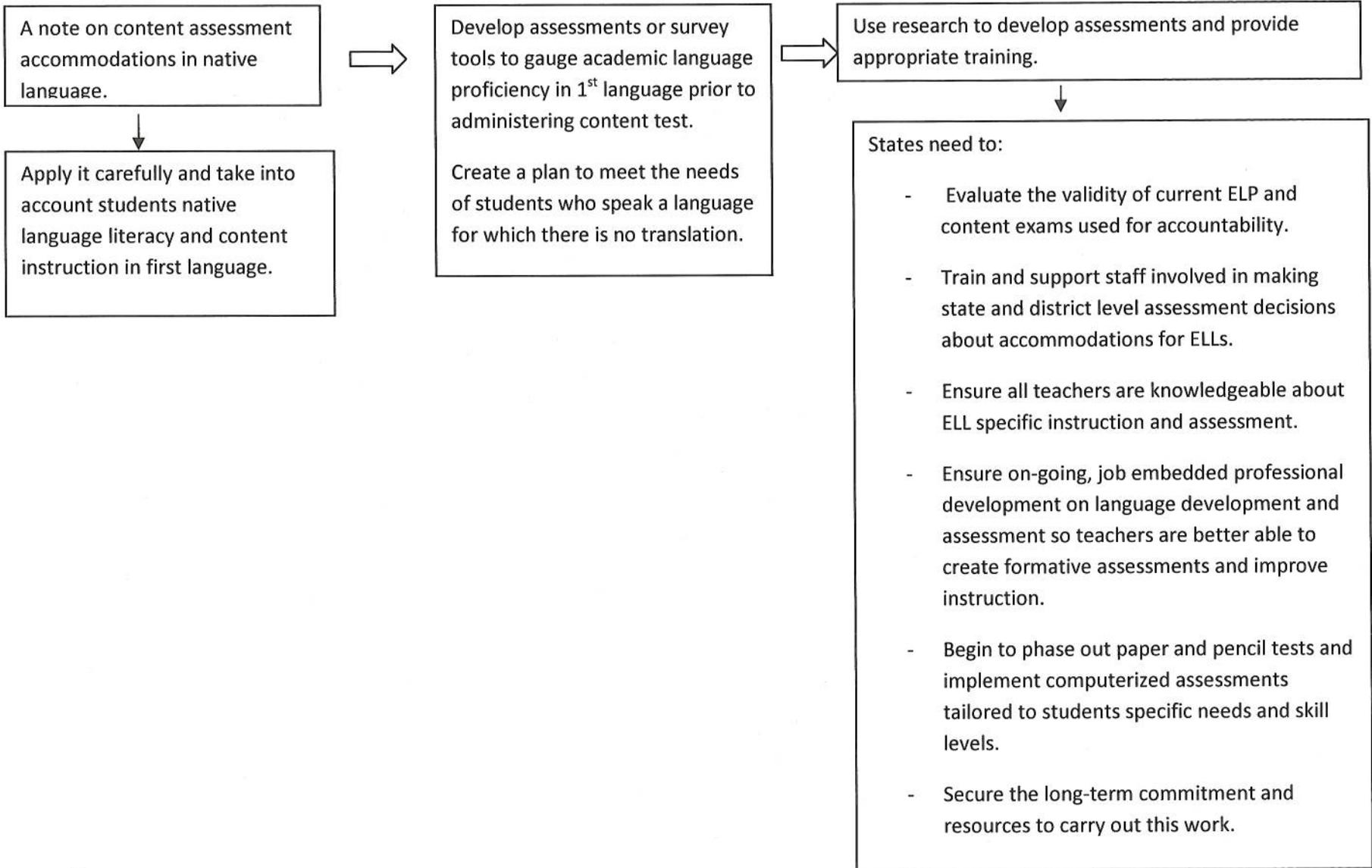
We need to find the best accommodations for ELLs. There are many promising research-based accommodations for ELLs.

For example:

- On screen pop-up English language dictionaries and glossaries
- Side-by-side dual translation – Spanish/English
- Translated Spanish assessments – if students are knowledgeable in content area.
- “plain English” version of tests – modify syntax, grammar, and vocabulary to avoid ambiguity.



ELLs continue to take and re-take tests only to “fail” over and over because of language barriers. They internalize this shame of failure and become discouraged. I have seen my students spend six hours in a single day trying to perform well on a content test that was too difficult for them. They didn’t want to give up – they wanted to be successful. It was heartbreaking and unfair.



We need to rely on the guidance of researchers who are highly qualified in language development and quality assessments.



Researchers to assist in this work include:

- Jamal Abedi
- Diane August
- David Frances
- Margo Gottlieb
- Charlene Rivera



Let's give hope to my students.

I came here with great hope and a quest for educational equity for my ELL students.

I leave this meeting confident in the expertise and dedication of the assessment developers and the US Department of Education.

I know when it comes to developing quality, valid assessments for ELLs –

I say, "Yes, We can."

Thank you.

ASSESSMENT OF ELLS:

**ASSESSMENT AND
INSTRUCTION DESIGN FOR
OUR CLIENTS.**

EJ RODRIGUEZ, DENVER PUBLIC SCHOOLS
RACE TO THE TOP

WHO IS AN ELL: RECENT IMMIGRANTS

- Mostly from Spanish speaking nations like Mexico, but other languages are also represented (Vietnamese, Arabic, Russian, etc.).
- These students might have a good level of formal education when they come from large urban centers in their countries, but we don't have a way to assess the extend of their formal education.
- In DPS 52% of students are growing up of Spanish Speaking families

EMERGING BILINGUALS

These are students born in the US and have been exposed to both the family language and English during the years 0 and 5.

Literacy skills on both languages vary.

STUDENTS OF MIGRANT FAMILIES.

These students often experience interrupted formal education and other factors that limit content and literacy development as measured by school expectations.

CURRENT STATE OF ELL ASSESSMENT

- We don't have a way to know what students know in L1, even when they had formal schooling before arriving in the US.
- The current assessments systems only want to measure what students know in English and content area assessments become language tests for ELLs.
- This affects the accuracy and reliability of the assessment.

WHAT SHOULD AN ASSESSMENT DO?

- We need assessments that respond to the following critical questions:
 - What do students know of the content?
 - What students can express related to that content in the English language?

CHARACTERISTICS

- Continue to focus on State, National and/or international standards.
- Content assessment: Differentiated to the language proficiency of the student.
 - Include native language and English in a developmentally appropriate manner.
 - Uses scaffolds for both languages, and are less dependant on the written word (Oral presentation of the prompts, graphic design, pictures, simplified language, drawings, technology based)

TECHNOLOGY CAN ASSIST LITERACY DEVELOPMENT

- Literacy development tools with language management tools:
 - These programs develop and use the strengths in the first language and gradually transfer them into the second language.
 - The degree of interaction with both languages is managed by the teacher using student progress information.

SOME EXAMPLES

- LITERACY
 - Imagine learning
 - ELLis by Pearson
- **MATH**
- Java tools for Interactive Demonstrations

THE FUTURE OF ASSESSMENT

- We envision assessments that:
 - can be administered quickly so teachers can use the assessment information for instructional purposes through out the school year.
 - Scoring and reporting that is more automated and produced by trained personnel on issues of second language acquisition.

Race to the Top Assessment Written Comment

Submitted by:

Michael Russell

Associate Professor, Educational Research, Measurement & Evaluation, Boston College

Director, Technology and Assessment Study Collaborative, Boston College

President, Nimble Assessment Systems, Inc.

December 2, 2009

Topics Addressed pertain to:

1. General Assessment Input
2. Assessment of English Language Learners
3. Assessment of Students with Disabilities
4. Technology and Innovation

The Race to the Top Assessment (RttTA) program holds promise to stimulate improvements in the quality and utility of test data provided by state assessment programs. As with all assessment, it is essential to clearly define the purpose(s) for assessing students and to then develop a program that supports enhancements aligned with that/those purpose(s). In the Notice of Public Meetings regarding the RttTA, several potential purposes of assessment are noted, including: a) holding schools, teachers, and potentially states accountable for student achievement; b) measuring student attainment of grade level standards; c) measuring student growth over time; d) providing information to help schools and teachers improve student learning; e) providing information to help parents and communities evaluate the quality of schools; f) comparing performance across students (normative); and g) comparing students to pre-defined performance levels (criterion). In addition, the Public Announcement identifies several aims for enhancing assessment programs, including: a) increasing the efficiency with which data is collected and returned to stakeholders; b) increasing teacher involvement in analyzing student work; c) developing innovative items; d) supporting the adoption of computer-based technological solutions; e) increasing the number of times test data is collected during the year; f) employing performance-based tasks to measure higher-order skills; g) adopting adaptive tests; and h) meeting the needs of students with disabilities and special needs (including English Language Learners).

In many cases, specific purposes of assessment and potential aims of the program are complimentary. As an example, adopting technological solutions holds potential to: a) increase the speed with which data is collected and returned to stakeholders; b) enhance a state's ability to collect multiple measures of learning during the school year to develop stronger growth models; c) tailor the presentation of information provided to parents, teachers, school leaders, community members, state leaders, and the federal leaders; and

d) improve accessibility for students with disabilities and special needs. In many cases, however, the potential purposes and aims are in conflict. As an example, the adoption of normative measures, performance-based tests, increased efficiency, measuring the achievement of grade level standards, and estimating student growth would likely work against each other, at least as they are commonly understood by the field. For example, performance-based tasks are typically designed to measure a sub-set of skills and knowledge included in grade level standards and require considerable time to administer and score, and are generally NOT designed to provide normative information.

The written comments that follow are designed to highlight some of the tensions that exist in the Public Announcement, have been raised in public comments, and which should be addressed when designing the program specifications. In many cases, lessons from prior efforts to develop and enhance state assessment programs are noted to support specification recommendations.

Tension 1: Primary Goal of the RttTA Program – Decrease Financial Burden on State Assessment Programs OR Improve Assessment Practices of State Programs?
Both the Public Announcement and public comments have identified two competing aims for the RttTA. On the one hand, there is a desire to improve the efficiency of state assessment programs and to decrease the financial burden for these programs by developing tests that can be shared across states. Clearly, given the high costs required to develop a sound assessment system coupled with the financial shortfalls in state budgets, decreasing the cost of testing is an important goal. On the other hand, there is a desire to improve assessment practices by developing and implementing innovative, technology-based solutions. While these two aims may be compatible in the long-run, in the short-term they are not.

Meeting the first goal is best accomplished by creating tests comprised of multiple-choice items with discrete correct responses, a technology that is well established in the field of testing. The primary challenge in accomplishing this first goal is reaching consensus across multiple states on the specifications for each test. Once consensus is reached, traditional methods of test construction, piloting, and implementation are well suited for developing a common multiple-choice test that can be employed across states with relatively little cost.

Meeting the second goal requires considerable investment in developing, experimenting with, and refining new methods and models of assessment. In many cases, these methods and models will build on shortcomings of prior efforts to enhance assessment. As an example, several states and testing programs experimented with performance-based tasks during the 1990s. While many of these tasks provide more authentic measures of student skill and knowledge, there were several technical shortcomings including challenges with reliably and efficiently scoring responses, and the validity of inferences about achievement across a domain. While technology and new approaches to psychometric modeling hold potential to overcome, or at least reduce, these challenges, considerable time and effort is required to fully develop and refine these approaches. Similarly, the development of innovative item types and approaches to providing valid measures of

student skills and knowledge will require considerable time and effort. While both lines of improving assessment practices can ultimately be incorporated into tests employed at low cost across testing programs, this payoff will take substantially longer to occur than simply employing traditional multiple-choice and short-answer items to develop a common assessment that effectively mirrors the tests currently developed by each individual state.

Tension 2: Form and Role of Consortium – Externally Directed OR Internally Directed
The RttTA announcement and public comments made to date emphasize that funds will be provided to a consortia of states working together to develop a common assessment. Consortia can take several forms. As one example, a group of states can contract an independent organization to play the lead role in directing and overseeing the development of a common assessment or enhancements to a common assessment. That independent organization can then contract with one or more test vendors or institutions to conduct the technical work required to develop one or more assessments or to develop enhancements for an assessment program. As an example, the development of the Algebra II test adopted this strategy. Several states with a common need (i.e., an Algebra II test) formed a consortium and contracted an independent organization (Achieve) to oversee and direct the consortium's activities. Achieve then managed and directed the process of working with states to create and reach consensus on test specifications. Achieve then contracted with a test developer (Pearson) to develop an Algebra II test. The consortium states then adopted (or had the option to adopt) the test as part of their state assessment programs. This approach was efficient and cost effective, and minimized the active involvement of states in the day-to-day work required to create a new assessment instrument. Despite the desire of several member states for the Algebra II test to be available in an accessible, computer-based mode (with multiple accommodations delivered via computer), this consortium model resulted in little innovation or advancement in the technology of testing.

An alternate consortium model takes the form of multiple states with a common interest coming together with one state playing the lead role, and each member state providing input on all activities, and support for specific sub-sets of activities. When needed, the lead state contracts with external organizations with expertise in specific areas of need. The New England Common Assessment Program (NECAP) provides a good example of this. The NECAP was the product on an Enhanced Assessment Grant awarded to Rhode Island. The grant allowed four states to work together to develop common standards and ultimately a common assessment. A subsequent Enhanced Assessment Grant awarded to New Hampshire allowed the NECAP states to partner with several additional states to enhance their assessment programs by exploring the development and adoption of universally designed computer-based test delivery methods that increased accessibility to test items for students with disabilities, students with special needs, and students developing English proficiency. As part of this first grant, the consortium tapped expertise from the testing industry and from academia to develop a common assessment and to begin to explore issues of accessibility. As part of the second grant, the consortium again tapped expertise to develop and refine cutting edge computer-based technologies and expertise that guided the consortium's work. While the NECAP consortium model

required considerably more active involvement by each member state and required more time to develop a tangible product, it resulted in a common assessment that was adopted across multiple states, and led to the development and adoption of an innovative, technological solution that can be easily adopted by other states and/or consortiums. In contrast to the Algebra II consortium, which also resulted in a common assessment instrument, the active participation of the NECAP states in directing the consortium activities coupled with involving multiple organizations, each with specific areas of expertise, enabled the development of sound common assessment instruments and innovative solutions that advanced the technology of testing.

Tension 3: Efficiency and Accessibility.

The RttTA program aims to provide funding for consortiums of states to develop and implement common assessment instruments in a relatively short period of time. Traditionally, test development has focused on creating an instrument that functions well for the majority of students. After a test is developed, accommodated versions are then created to allow students with disabilities and special needs to participate in the testing program. This process is effective for efficiently creating and piloting test items, and then assembling them into tests that function well for many students. However, the process of retrofitting a test to meet the needs of specific sub-sets of students with disabilities and special needs is often costly and problematic from a psychometric perspective, particularly when a retrofit violates the construct that is being measured by a given test or test item.

An alternate approach requires test developers to focus on accessibility throughout the test development process. As part of this process, construct definition must be considered in light of the various accessibility needs of any student who is expected to perform the test instrument. As part of the item development process, all elements of an item are examined for accessibility by any and all students, and alternate representations of some content are created to ensure that the content is accessible as possible for students with a given need without violating the measured construct. In addition, piloting of items and tests purposefully samples students with specific needs to examine the functioning of the items and test for all potential test-takers. Once development is complete, the delivery method then allows the appropriate version of an item or test to be presented to a student based on his/her individual need.

Clearly, the second approach is more demanding of the test development process, and may decrease the efficiency of developing a test. However, if a common assessment is to be of the highest quality and function appropriately for all students, this approach preferable to retrofitting for specific needs after the main assessment instrument is developed.

Additional Considerations:

In addition to the tensions described above, two additional issues must be addressed when developing the full RttTA specifications. These include building on existing solutions and

adopting common item banking and accessibility standards. Below, each of these are discussed briefly.

Building on Existing Solutions: Over the past five years, federal funding has supported the development of several innovations that are proving effective for improving the quality of state assessments. As one example, several projects have documented the value of applying principles of Evidenced Centered Design (ECD) when developing state assessments. While the methodology and tools for applying ECD may be in need of additional refinement, the underlying principles and procedures for applying ECD are well established and have been applied in multiple contexts. Similarly, several federal programs have provided funding for the development of universally designed computer-based test delivery which allows the testing environment and/or test content to be matched to the needs of each individual student. This approach has been demonstrated to dramatically reduce the need of test accommodations while also improving the validity of test-score-based inferences for students with disabilities, special needs, and/or who are developing English-language proficiency. Given the recent success of ECD and universally designed computer-based test delivery, it is sensible for the RttTA program to strongly encourage consortium to adopt demonstrated methods and solutions.

Common Item Banking and Accessibility Standards: The RttTA program holds potential for a consortia of states to develop one or more assessment instruments that a non-member state later adopts. When an instrument is computer-based, the test content can be either integrated and dependent on a specific delivery system or the content can be designed to be transferable across test delivery platforms. Similarly, as noted above, when developing a given assessment instrument, it is essential to consider accessibility throughout the development process. If the aim of the development process is to make test content that is portable across states and their test delivery systems, and which is equally accessible across settings, it is essential to establish and adopt common item banking and accessibility standards during the test development process. To this end, the Global Learning Consortium developed the Question Test Interoperability (QTI) and Access for All (AfA) specifications. These specifications provide the elements for creating common item banking and accessibility standards. More recently, several states have begun working with the Global Learning Consortium and other experts in the testing industry to develop an Accessible Portable Item Profile that builds on the QTI and AfA specifications to create a standardized method for structuring and tagging test items. The majority of this work will be completed prior to the awarding of the an RttTA funds, and should provide the foundation for common item banking and accessibility standards that assure that test content can be adopted by states that are not members of a given consortium and that the content contains all accessibility features required to meet the needs of all students.

Recommendations:

1. To stimulate the development of common assessment instruments, methods, and programs that have a high probability of adoption across states while also stimulating advancements to the technology of testing and assessment, an attractive option is to structure consortiums such that there are two categories of

membership. Implementing states would focus on developing, piloting, and being the initial implementers of a common assessment instrument designed to provide an annual measure of student learning. Development states would focus on a specific area of assessment that could be enhanced through the development of innovative methods that can be integrated into the common assessment. Such innovations may focus on developing interactive, accessible computer-based items that measure higher-order skills (e.g., simulated labs, manipulations, etc.), enhancing methods for reporting data to specific stakeholders, developing performance-based measures and/or teacher-based scoring systems, etc. As an example, this strategy was adopted for NH's Enhanced Assessment Grant, and has resulted in the implementation of universally designed computer-based test delivery for special populations participating in the NECAP operational test, accompanied by a set of studies conducted in partner states designed to explore and develop additional methods for increasing accessibility. Adopting (or allowing for) a tiered consortium membership structure holds promise to produce common operational assessment instruments in a timely and efficient manner while also supporting the development of innovative solutions that can be gradually folded into operational assessment procedures.

2. To avoid the shortcomings of the past, it is essential that accessibility be strongly emphasized and that consortia be required to address accessibility during each stage of their development, piloting, and implementation processes. To the extent possible, consortia should be encouraged to adopt proven methods designed to clarify the measured constructs, the evidence required to make inferences about achievement of those constructs, construct irrelevant factors that may decrease validity, task specifications, and acceptable alternate representations of test content and student evidence. To this end, consortia should be encouraged to embrace principles of Evidence Centered Design and Universally Designed Assessment/Test Delivery.
3. To maximize expertise and innovation, consortia should be strongly encouraged to separate all components of their development and implementation into discrete elements and to work with organizations with expertise in each component. As Clayton Christensen documents, it is difficult for any one organization to develop expertise across all areas of an industry, and it is even more difficult for an established organization to lead innovation. To maximize the benefits of the RttTA program, consortia should be encouraged to identify key components of their initiative and to work with a collection of organizations that specialize and/or are the leading-edge innovators for a given component. Again, the work of the NECAP states provides evidence that this approach is highly effective for developing a common assessment program employed by multiple-states, while also being proactive in exploring, researching, piloting, and adopting innovative methods that enhance that program.

Involvement in State Assessment Practices:

Over the past 15 years, I have engaged in a variety of research and development activities specific to state assessment practices. Among these activities are:

- a) Large-scale study of the effect of state testing programs on school and instructional practices
- b) Several studies that compare the effect of computer-based versus paper-based administration of writing tests on student performance and rater accuracy
- c) Several studies funded through Enhanced Assessment Grants that examine computer-based solutions to test accommodations and alternate assessments
- d) The development and implementation of a universally designed test delivery system for the NECAP state Science test
- e) Co-Author of a book title, *The Paradoxes of High-Stakes Testing*, that examines issues related to state testing programs



Universal Design for Learning and the Race to the Top Assessment Program

Input Requested in Federal Register: October 23, 2009 (Volume 74, Number 204)

More than 35 national disability and education groups representing higher education, general and special education interests are working together as the **National Universal Design for Learning (UDL) Task Force** to promote the use of UDL in today's diverse classrooms. For more information on the Task Force see www.udl4allstudents.com.

A key goal of the National UDL Task Force is to influence the Administration and Congress to reflect UDL principles as they relate to the four elements of the curriculum (goals, instructional materials, teaching methods and assessments) in all education policy and legislation. The Higher Education Opportunity Act included provisions on UDL and the House and Senate versions of the LEARN Act also incorporate UDL.

UDL is a scientifically valid framework for guiding educational practice that ensures accessibility in instruction and assessment. Therefore, it addresses Required Assessment Design Characteristic #2 for the Race to the Top Assessment Program, that assessments be accessible to the broadest possible range of students, including students with disabilities and English language learners. Also, students who would not be able to demonstrate knowledge and skills consistent with the goal of being college and career ready by the time of high school completion, on a traditional assessment, may be able to do so if the assessment is designed using the principles of UDL.

Assessments based on UDL would be designed from the beginning to provide 1) multiple means of recognition of assessment directions and stimuli, 2) multiple means of interaction and expression within assessment tasks, and 3) multiple means of engagement during the assessment.

We strongly recommend that the Race to the Top Assessment Program reflect the principles of UDL to ensure that the learning of all students is validly assessed. For more information on UDL see www.cast.org and www.udlcenter.org.

The National UDL Task Force appreciates your consideration of the following input for the Race to the Top Assessment Programs for ESEA reauthorization. Please contact Ricki Sabia at rsabia@ndss.org if the Task Force members can be of assistance or if you have any questions.

Ricki Sabia, Chair, National UDL Task Force, National Down Syndrome Society

Reggie Felton, Co-chair, Policy Committee, National School Boards Association

Nancy Reder, Co-chair, Policy Committee, National Association of State Directors of Special Education

Myrna Mandlawitz, Co-chair, Communications Committee, Learning Disabilities Association of America and School Social Work Association of America

Input on the Race to the Top Assessment Program

Topic: Assessment of English Language Learners

***Question #1:** Provide recommendations for the development and administration of assessments for each content area that are valid and reliable for English language learners. How would you recommend that the assessments take into account the variations in English language proficiency of students in a manner that enables them to demonstrate their knowledge and skills in core academic areas? Innovative assessment designs and uses of technology have the potential to be inclusive of more students. How would you propose we take this into account?*

Topic: Assessment of Students with Disabilities

***Question #1:** Taking into account the diversity of students with disabilities who take the assessments, provide recommendations for the development and administration of assessments for each content area that are valid and reliable, and that enable students to demonstrate their knowledge and skills in core academic areas. Innovative assessment designs and uses of technology have the potential to be inclusive of more students. How would you propose we take this into account?*

Our response to both these questions is the same. The application of UDL principles to assessments can result in a more authentic and accurate measure of the achievement of all students without having to make costly retrofits or rely on certain accommodations that may violate test validity.

Assessments that are developed using the principles of universal design for learning avoid construct irrelevant barriers to students showing what they know. For example, if the area being assessed is mathematics, science, etc., the ability to read is not the target skill, it is “construct irrelevant.” However, for students with disabilities, English language learners and others, these construct-irrelevant demands generate artificially low achievement scores. (Dolan, Rose, Burling, Harms, & Way, 2007; Rose, Hall, Murray, 2008). In addition, complex vocabulary that is not part of the material being tested may pose a significant barrier for certain students, resulting in poor test performance regardless of their proficiency with the subject or skill area being tested. (Clarkson, 1983; Helwig, Rozek-Tedesco, & Tindal, 2002; Helwig, Rozek-Tedesco, Tindal, & Heath, 1999).

An assessment can be developed using the principles of UDL to address these and other accessibility issues without affecting the validity of the construct being assessed. A UDL assessment would level the playing field and allow students to accurately demonstrate content knowledge and skill acquisition without the barriers traditional assessments often pose for students with disabilities, English language and others. This is possible to do without technology. However, digital media technology makes it easier to build-in supports to reduce these barriers (e.g. text to speech features, vocabulary supports, graphic organizers etc).



Race to the Top Assessment Program Comments

Input Requested in Federal Register: October 23, 2009 (Volume 74, Number 204)

The National Down Syndrome Society (NDSS) and the National Down Syndrome Congress (NDSC) are nonprofit organizations with more than 200 affiliates nationwide representing the more than 350,000 Americans who have this genetic condition. We appreciate the opportunity to provide input on the Race to the Top Assessment Program. Our comments are directed at the Assessment System requirements, characteristics and questions italicized below.

Although the input being requested is primarily about assessment design, it is important to note that the ability to design valid and accurate assessments is limited by the accountability rules and the perception they create about student ability. The many misconceptions about students who take the alternate assessment on alternate academic achievement standards and the interpretation of the regulations governing these assessments has led to serious unintended consequences regarding assessment eligibility, implementation of the assessment process and instruction..

Attached are our recommendations for amending the regulations governing alternate assessments on alternate academic achievement standards. The issues raised in these recommendations should help inform assessment design decisions. Also attached is a document addressing the myths regarding the alternate assessments.

Framework- Design of Assessment Systems –General Requirements

The Department is particularly interested in supporting the development of summative assessments that measure

—Individual student growth (that is, the change in student achievement data for an individual student between two or more points in time.

Measuring individual student growth is usually a required characteristic of assessment in order to apply a growth model for accountability purposes. In June 2009, a Growth Model Task Force, which was brought together by the National Center for Learning Disabilities, issued a report with considerations for including students with disabilities in growth models (see attachment).

One of principles asserted in this report is that all students should be included in any assessment and accountability system and be valued in the same way. We have moved past exclusion from assessments, for the most part, but face new challenges if we want to be sure that all students are fully included in growth model approaches to accountability. For students with disabilities, this means including those who participate in any alternate assessments as well as those taking the regular assessment with or without accommodations. We urge the Department of Education to review the report in its entirety.

Framework- Design of Assessment-Required Characteristics

With respect to the design of the assessment system, the Department would likely require that the assessments, at a minimum, meet the following characteristics:

- (1) Reflect and support good instructional practice by eliciting complex responses and demonstrations of knowledge and skills consistent with the goal of being college and career ready by the time of high school completion;*
- (2) Be accessible to the broadest possible range of students, with appropriate accommodations for students with disabilities and English language learners;*

College and Career Readiness

It is essential to recognize that the skills needed to be college and career ready depend very much on the college program and the career path of the students. However, at a minimum every student, including students with the most significant cognitive disabilities must be assessed on material aligned to the State content standard for the grade in which the student is enrolled. Since assessments drive instruction, it is critically important that assessments for all students drive access to the general education curriculum and inclusion in the general education classroom.

Many students with Down syndrome now attend college and University programs for students with intellectual disabilities (see www.thinkcollege.net for a data base of existing programs). As you know, the Higher Education Act of 2008 authorized the development and expansion of high-quality, inclusive model comprehensive transition and post-secondary programs for students with intellectual disabilities and the establishment of a coordinating center for technical assistance, evaluation, and development of recommendations for model accreditation standards. The Act also allows these students to be eligible for Work Study Jobs, Federal Supplemental Educational Opportunity Grants, and Pell Grants.

Accessible Assessments-UDL

It will be a challenge to design an assessment that elicits complex responses and demonstrations of knowledge and skills consistent with the goal of being college and career ready by the time of high school completion while ensuring that the assessment is accessible to the broadest range of learners, including students with disabilities and English language learners. Therefore the assessments on these skills and the complexity of the response required must be designed from the beginning with all learners in mind.

UDL is a scientifically valid framework for guiding educational practice that ensures accessibility in instruction and assessment (see www.cast.org and www.udlcenter.org). Students who would not be able to demonstrate knowledge and skills consistent with the goal of being college and career ready by the time of high school completion, on a traditional assessment, may be able to do so if the assessment is designed using the principles of UDL.

Assessments based on UDL would be designed from the beginning to provide 1) multiple means of recognition of assessment directions and stimuli, 2) multiple means of interaction and expression within assessment tasks, and 3) multiple means of engagement during the assessment.

We strongly recommend that the Race to the Top Assessment Program reflect the principles of UDL to ensure that the learning of all students is validly assessed. Additional information on UDL and assessments is provided below.

Assessment Questions- Assessment of English Language Learners

***Question #1:** Provide recommendations for the development and administration of assessments for each content area that are valid and reliable for English language learners. How would you recommend that the assessments take into account the variations in English language proficiency of students in a manner that enables them to demonstrate their knowledge and skills in core academic areas? Innovative assessment designs and uses of technology have the potential to be inclusive of more students. How would you propose we take this into account?*

Assessment Questions- Assessment of Students with Disabilities

***Question #1:** Taking into account the diversity of students with disabilities who take the assessments, provide recommendations for the development and administration of assessments for each content area that are valid and reliable, and that enable students to demonstrate their knowledge and skills in core academic areas. Innovative assessment designs and uses of technology have the potential to be inclusive of more students. How would you propose we take this into account?*

UDL

One aspect of our response to both these questions is the same. The application of UDL principles to assessments can result in a more authentic and accurate measure of the achievement of all students without having to make costly retrofits or rely on certain accommodations that may violate test validity.

Assessments that are developed using the principles of universal design for learning avoid construct irrelevant barriers to students showing what they know. For example, if the area being assessed is mathematics, science, etc. the ability to read is not the target skill, it is “construct irrelevant.” However, for students with disabilities, English language learners and others, these construct-irrelevant demands generate artificially low achievement scores. (Dolan, Rose, Burling, Harms, & Way, 2007; Rose, Hall, Murray, 2008). In addition, complex vocabulary that is not part of the material being tested may pose a significant barrier for certain students, resulting in poor test performance regardless of their proficiency with the subject or skill area being tested. (Clarkson, 1983; Helwig, Rozek-Tedesco, & Tindal, 2002; Helwig, Rozek-Tedesco, Tindal, & Heath, 1999).

An assessment can be developed using the principles of UDL to address these and other accessibility issues without affecting the validity of the construct being assessed. A UDL assessment would level the playing field and allow students to accurately demonstrate content knowledge and skill acquisition without the barriers traditional assessments often pose for students with disabilities, English language and others. This is possible to do without technology. However, digital media technology makes it easier to build-in supports to reduce these barriers (e.g. text to speech features, vocabulary supports, graphic organizers etc).

Many of the assessment design changes allowed for assessments on modified academic achievement standards are based on UDL principles. Numerous studies show that many of the students who score in the lowest 2% are students do not have IEPs. They would also benefit from UDL. Therefore, we urge the Department to use the research generated by the 2% rule with respect to UDL assessment features, to develop better assessments on grade-level academic achievements standards, rather than fund assessments on modified academic achievement standards.

Adaptive Testing

In the context of the digital assessments we feel it is necessary to provide a word of caution about the current enthusiasm for computer adaptive testing. Although, UDL

supports customization in assessments, the score must be an accurate representation of the student's knowledge.

The use of computer adaptive tests in the presence of idiosyncratic knowledge patterns has been studied and results show that scoring of adaptive tests is problematic when a test taker responds to questions in an unexpected way. Results also indicate that a fairly large number of students might have test results that are influenced by idiosyncratic patterns of knowledge. (Kingsbury, G.G. & Houser, R.L. (2007). ICAT: An adaptive testing procedure to allow the identification of idiosyncratic knowledge patterns. In D. J. Weiss (Ed.). *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing*. Retrieved 12/02/09 from www.psych.umn.edu/psylabs/CATCentral/). For example, the Associate Director of the NDSS Policy Center was told that her son with Down syndrome was reading on a first grade instructional level, based on an adaptive test, the same year he scored just 40 points below a proficient cut score of 385 on the eighth grade state assessment on the grade-level academic achievement standard. Additional research would be required to determine the impact of this problem on all students, including students with disabilities, and to determine how to address it.

In his testimony at the Race to the Top Assessment Program hearing in Boston, Skip Stahl of CAST conveyed the following cautionary message regarding adaptive testing and grade level alignment:

The proponents of computer adaptive testing often point to the “automatic” difficulty adjustments of that approach as enhancing student engagement by decreasing the challenge presented to them. This is the same rationale used to support the simplification of the curriculum for struggling students, identical to the “out of level” testing that results in moving students with disabilities further away from the mainstream curriculum. Universal Design for Learning seeks to maintain high achievement standards for all students through the use of customized scaffolds and supports that reinforce the importance of maintaining grade-level expectations for all learners.

In addition, a Pearson paper from 2006 acknowledges the challenges of providing certain accommodations on a computerized adaptive test. See page 11-12 at http://www.pearsonedmeasurement.com/downloads/research/RR_05_03.pdf.

The concerns we have described as well as the lack of a research base to support the use of adaptive testing for students with disabilities (or any students with idiosyncratic knowledge patterns) should be considered before the Department approves the use of adaptive testing.

Assessments on Alternate Academic Achievement Standards

As you know, current federal regulations allow states to develop and administer alternate assessments based on alternate achievement standards for students with the most significant cognitive disabilities. While this policy has been in place for some time, the consistency and availability of these assessments varies widely between states. A recent study by the National Center for Special Education Research, within the Institute Of Education Sciences, found that many states approach the alternate assessments on alternate academic achievement standards differently (Cameto, R., Knokey, A.-M., Nagle, K., Sanford, C., Blackorby, J., Sinclair, B., and Riley, D. 2009). Some states use a portfolio or body of evidence to constitute the entire assessment. Others use techniques such as a rating scale/checklist, performance task/events, or multiple choice/constructed response assessments. The inconsistent approach to these assessments across states creates varying standards and expectations.

We urge the Department to call for an analysis of the various types of alternate assessments on alternate academic achievement standards to see which are challenging, aligned to grade level content, can fit with a growth model and can be implemented without placing students in special education classes to collect evidence.

We also know from a new 7-state survey conducted by the National Alternate Assessment Center that 75 percent of the students participating in state alternate assessments on alternate academic achievement standards are reading sight words and using a calculator to do basic math operations. This finding suggests that many students assigned to this assessment may, in fact, be capable of participating in more rigorous assessments.

Common misperceptions and research-based recommendations for Alternate assessment based on alternate achievement standards

This discussion tool refers to alternate assessments based on alternate achievement standards (AA-AAS), assessments intended for students with significant cognitive disabilities. The acronym AA-AAS is used throughout.

Common misperception 1 – Many students eligible for AA-AAS function more like infants or toddlers than their actual age, so it makes no sense for schools to be held accountable for their academic performance.

Why is this misperception common?

People sometimes assume that students in this group all have very severe disabilities that limit what can be taught to them. This misperception also is rooted in the faulty assumption that all students must progress through infant and preschool skill development before any other academic instruction can occur.

Research Response: We have understood for many decades that waiting until these students are “ready” by mastering all earlier skills means they “never” will be given access to the skills and knowledge we now know they can learn. In the 1980s, we realized that these students were able to master many functional skills appropriate for their age regardless of whether they had mastered all lower skills. This caused a shift in thinking that resulted in a powerful, age-appropriate functional curriculum for these students.

In recent years, we have demonstrated that these students can learn grade-appropriate academic skills in addition to functional skills. Learner characteristics data from many states show us that MOST students who participate in AA-AAS have basic literacy and numeracy skills. These students are able to learn targeted grade-appropriate academics to an alternate achievement level, even when they have not mastered all earlier academic content.

Recommendation: Build accountability systems to ensure that all students who are eligible for the AA-AAS have access to and learn academic content expected for their same-age typical peers, to an appropriate but challenging alternate achievement expectation.

Common misperception #2 – Many students who participate in AA-AAS have life-threatening medical conditions or are not able to communicate.

Why is this misperception common?

People sometimes assume that AA-AAS students are a small homogeneous group of students with multiple problems that go well beyond what schools can actually handle; they assume that many of these students cannot speak, hear, or communicate in any way.

Research Response – The students who may be eligible for AA-AAS are generally less than 1% of the total student population or about 9% of all students with disabilities. Most of these students who are eligible for AA-AAS (90%) have effective communication skills with or without assistive devices. Only 10% of AA-AAS students have very severe and complex disabilities. These students (0.1% of total population of students) *can* communicate, but only if they are given opportunities to learn, including the use of assistive devices. The field of severe disabilities has worked from the “least dangerous assumption” for decades. We teach assuming that all students can build effective communication strategies.

Recommendation: For the very small (0.1% of total population of students) group of students who have the most severe and complex disabilities, educators should persistently and systematically seek successful methods to permit these students to first learn and then show what they know on an AA-AAS using multiple and varied communications strategies.

Common misperception #3 – Students in the AA-AAS can learn only rote academic skills, so AA-AAS should reflect only these skills.

Why is this misperception common?

People sometimes assume that the curriculum for students with severe disabilities has been based on math skills of time and money and reading skills limited to sight words because that is all they can learn. Thus, AA-AAS should focus on these same limited skills.

Research Response: It is true that research through the 1990s reflects a very narrow curriculum for students with severe disabilities, with instructional approaches relying on direct and repeated instruction that resulted in learning. Researchers now are finding strong evidence of academic skills and knowledge development among students who participate in AA-AAS, including abstract concepts and transfer of learning. Setting higher expectations for these students results in higher student performance on a range of grade-level content that can be demonstrated in large-scale assessments.

Recommendation: *Build AA-AAS approaches based on a curriculum framework that allows these students to demonstrate a range of grade-level content.*

Common misperception #4. – The AA-AAS has eliminated the teaching of important functional skills.

Why is this misperception common?

People sometimes assume that the addition of academics to the assessment and accountability systems for students with severe disabilities means that there is limited time for teaching functional skills like self-care, community participation, and safety. They believe that there is not enough time in the day to do both academics and functional skills.

Research Response: Many teachers have found that blended instruction in academic and functional skills yields better results for both. The “line” between academics and functional instruction begins to blur as teachers and parents discover how truly useful and satisfying increased literacy and numeracy skills are for these students, for quality of life and enjoyment, for integration into the community, school, or adult life, and for future employment.

Recommendation: *Provide training and support to teachers so that they can effectively merge academic and functional instruction where appropriate and so that they understand the vital importance of academic skills and knowledge to full participation in family, school, and community life.*

Common misperception #5 – An AA-AAS has to cover all of the same content that is on the general assessment for typical peers.

Why is this misperception common?

People sometimes assume that federal law requires the same content on all tests. At the same time they believe that the grade-level curriculum is too challenging and covers too much for these students to learn in a year, or ever.

Research Response: Federal regulations permit states to define the appropriate depth, breadth, and complexity of content coverage for the AA-AAS. States must show that these content priorities truly “raise the bar” of historically low expectations, and are clearly linked to the content that typical students in the same grade should know and be able to do. This is a shift for teachers who do not have experience with this content. Stakeholder and advisor understanding can ensure that AA-AAS are linked to the student’s grade (or grade band) but are reduced in scope and complexity from the general education assessment.

Recommendation: Provide training to teachers, and to other key assessment system stakeholders and advisors on what research and best practices documentation show these students are able to know and do when given the opportunity.

Common misperception #6 – Most AA-AAS are entirely individualized and differ for each student.

Why is this misperception common?

People sometimes assume that teachers make so many adjustments to the assessment for each student that there is no way to compare results from one school to another.

Research Response: A good AA-AAS allows a defined amount of flexibility in administration of the items and tasks because students with the most significant cognitive disabilities vary in how they take in and respond to information and requests. A good AA-AAS incorporates training, oversight, and structures to balance flexibility with standardization of procedures. Ongoing monitoring is conducted to ensure the assessments are administered, scored, and reported as intended.

Recommendation: All AA-AAS scores should indicate whether the student is proficient in an academic domain through procedures that allow flexibility and at the same time control for possible sources of error.

Common misperception #7 – An AA-AAS measures teacher performance instead of student performance.

Why is this misperception common?

People sometimes assume that teachers who are able to put together good-looking portfolios or examples, or who can choose student examples that make them look good, will have students who score higher than the students of teachers who may teach well but who do not spend time creating good-looking portfolios or examples of what their students do.

Research Response: A good AA-AAS requires test administrators who are familiar to the student because of the way they take in and respond to information and requests. That means that in most cases, teachers interact with the student to capture accurate evidence of what the student knows and can do. This teacher role requires high-quality scoring procedures that focus on scoring of independent student performance and control for administrator behaviors.

Recommendation: Train teachers on systematic data gathering procedures, provide oversight and monitoring to ensure they implement the procedures as intended, and design scoring processes to exclude evidence that reflects teacher behaviors instead of independent student performance.

Common misperception #8 – It would make more sense if teachers simply reported on the achievement of their own students rather than use an AA-AAS.

Why is this misperception common?

People sometimes assume that students with the most significant cognitive disabilities have IEPs that define what they should be learning. If that is so, then gathering data that already are used for the IEP is the best measure of the students' achievement.

Research Response: A good IEP will identify the services, supports, and specialized instruction needed so that the student can learn both academic and functional skills and knowledge. Data gathered on the specific goals and objectives in the IEP are important for individual accountability among IEP team members for these short and long-term goals and objectives, in all areas where the student has them. Some of these goals and objectives will specify the services and supports the student needs to *access* the general curriculum, but student progress based on the IEP does not provide accountability for student *achievement* of proficiency in the general curriculum. In contrast, AA-AAS are designed to provide data for system accountability to ensure that all students are provided access to and are achieving to proficiency in the general curriculum. The leverage of system accountability as well as individual accountability can yield far more opportunities for most students.

Recommendation: Design AA-AAS so that there are good data on the effectiveness of schools in providing access to the general curriculum as a complement to the individual accountability of the IEP.

Common misperception #9 – Some AA-AAS formats (i.e., portfolio, checklist, performance assessment) are better than others.

Why is this misperception common?

People sometimes assume that one method is better than another, with “better” meaning more technically adequate; the specific method that is considered better or worse is based on assumptions about methods based on preconceptions about testing design.

Research Response: Research on the technical quality of AA-AAS has shown that the format of the test is a poor predictor of technical quality. The nature of a “portfolio” or “checklist” or “performance assessment” can vary enormously, and a number of states now use hybrid models that combine elements of these approaches. Any of these formats by name alone can be of poor or high quality.

A good AA-AAS is built on a set of beliefs about how students with severe disabilities learn and demonstrate the academic content. Questions that need to be addressed include: What kinds of observations of their learning will give us evidence of what these students know and do in the academic content? What should we “see” when these students have been given appropriate access to the same grade-appropriate, interesting content as their typical peers? The answers to these questions help answer the question of what is the “best” format for the AA-AAS.

Recommendation: Select the format of the AA-AAS based on beliefs about academic teaching and learning for AA-AAS students.

Common misperception #10 – No AA-AAS can be a technically adequate measure of student achievement for accountability purposes.

Why is this misperception common?

People sometimes assume that the AA-AAS breaks all the rules of good design of large-scale assessments as judged by high quality psychometric evidence that has been used by measurement experts for a century.

Research Response: The challenges of designing AA-AAS are very new; prior to the 1990s, no large-scale assessment program included students with significant cognitive disabilities, and very few measurement experts had experience designing assessments for these students. Fortunately, there has been a great deal of work done since the 1990s on issues that have emerged in developing psychometrically sound AA-AAS. AA-AAS can be designed to produce valid and reliable information about student outcomes

Recommendation: State assessment offices should address three components of the assessment design as they develop and implement the AA-AAS: (a) description of the student population and a theory of learning for these students, (b) structure of the observations from the assessment, and (c) interpretation of the results. The technical defense of an AA-AAS starts and ends with these three components.



Issues for Updated Regulations on Alternate Academic Achievement Standards

Updated regulations on Alternate Academic Achievement Standards are necessary in light of the regulations on the modified academic achievement standard. The requested clarifications, below, are supported by the following resources: the Notice of Final Title I Regulations (December 9, 2003), the Guidance on Alternate Academic Achievement Standards (August 2005) and A Decision Framework for IEP Teams Related to Methods for Individual Student Participation in State Accountability Assessments (released in the toolkit that accompanied the modified academic achievement standard regulations). The updated regulations should clarify:

- That broad stakeholder input is required in the standards setting process to define alternate academic achievement standards.
- That the alternate academic achievement standard must be aligned with the State's academic content standards for the grade in which the student is enrolled (or, in the case of students in un-graded classrooms, the grade level commensurate to the student's age).
- That alternate academic achievement standards must provide (rather than merely promote) access to the curriculum. The provision ensuring that students with the most significant cognitive disabilities are, to the maximum extent possible, included in the general education curriculum, should be amended to delete the underscored limitation. If these changes are not made the regulations will undermine the requirement in IDEA that all students with disabilities must be enabled to be involved in and make progress in the general education curriculum.
- That guidelines for the IEP team on eligibility should include the following criteria:
 - (i)The student's disability has precluded the student from proficiency as measured against the grade-level or modified academic achievement standards, even with accommodations, as demonstrated by such objective evidence as--
 - (A)The State's assessments and

(B) Other assessment data that can validly document academic achievement.

(ii) The student's progress in response to high-quality instruction, including special education and related services designed to address the student's individual needs, is such that the student is not likely to achieve proficiency as measured against grade-level or modified academic achievement standards within the year covered by the student's individualized education program (IEP).

(iii) The determination of the student's progress must be based on multiple measurements, over a period of time, that are valid for the subjects being assessed.

(iv) The student is receiving instruction, by highly qualified teachers as defined in IDEA and NCLB, in the grade-level curriculum for the subjects in which the student is being assessed.

- That a student's eligibility for alternate academic achievement standards must be determined separately for each of the subjects for which assessments are administered (and the assessment must be designed to be administered separately in each subject)
- That the decision to assess a student based on alternate academic achievement standards must be reviewed annually by the student's IEP team, for each subject area in which the student is assessed based on these standards, to ensure that those standards remain appropriate.
- That students taking assessments on alternate academic achievement standards are not precluded from the opportunity to work towards a diploma.
- That IEP teams, including parents, receive training on the implementation of the eligibility guidelines.
- That appropriate guidelines should be developed to help IEP teams draft and implement IEPs for these students, including a requirement that IEP goals for the subjects assessed are aligned to the academic content standards for the grade in which the student is enrolled.

Race to the Top Assessment Program Assessment of English Language Learners

Edynn Sato, Ph.D.
Director, Research and English Language Learner Assessment
Assessment and Standards Development Services, WestEd
and
Director, Special Populations
Assessment and Accountability Comprehensive Center at WestEd
Phone: 415-615-3226
Email: esato@wested.org

Thank you for this opportunity to provide input on this very important topic, the assessment of English language learners (ELLs). Over the past several years, states and districts have been developing or refining their assessment, accountability, and support systems for ELL students. This has been no easy feat—assessment issues alone, involving ELL students, are complex and the implementation of new systems is challenging. Even so, promising practices are emerging, as is research that can help inform practice and policy.

We are at a critical juncture to systematically evaluate “lessons learned” from recent implementation efforts and findings from rigorous research in order to inform the federal, state, and local discussions related to the Race to the Top Assessment Program and implications for its design, implementation, and potential role in accountability, given upcoming reauthorization of the Elementary and Secondary Education Act.

There is a substantial achievement gap between our ELL student population, one of our nation’s fastest growing subgroups, and their English-speaking peers. ELL students also have a significantly greater school drop-out rate. A critical goal is to improve systems of assessment and support, as well as articulate policies that make possible the successful implementation of such system improvements, to help boost the achievement and graduation rates of our ELL students.

The Department has posed two questions of interest related to the assessment of ELL students (Source: <http://www.ed.gov/programs/racetothetop-assessment/executive-summary.pdf>):

1. Provide recommendations for the development and administration of assessments for each content area that are valid and reliable for English language learners. How would you recommend that the assessments take into account the variations in English language proficiency of students in a manner that enables them to demonstrate their knowledge and skills in core academic areas? Innovative assessment designs and uses of technology have the potential to be inclusive of more students. How would you propose we take this into account?
2. In the context of reflecting student achievement, what are the relative merits of developing and administering content assessments in native languages? What are the technical, logistical, and financial requirements?

In order to address these questions, I highlight some relevant information and resources, organized according to (1) upfront considerations, (2) considerations for development and implementation, and (3) evaluation of consequences. The general framework for the information I highlight is available to states and districts in the *Framework for High-Quality English Language Proficiency Standards and Assessments* (Framework) (Assessment and Accountability Comprehensive Center, 2009) for which I served as lead author for the Assessment and Accountability Comprehensive Center (AACC), collaborating with experts from multiple disciplines from across the nation. I will also discuss the Framework later in my statement.

1. Upfront Considerations

a. Defining “Proficiency”: Ensuring Appropriate Alignment of Expectations Across English Language and Academic and Career Readiness Content

Proficiency expectations (e.g., achievement standards, performance level descriptors [PLDs]) for English language proficiency and for the academic and career readiness content areas should be “aligned” in terms of the level and range of language skills and knowledge for “proficiency” in English that supports students’ engagement with and ultimate achievement of academic and career readiness content, and provides students the level and range of English language skills and knowledge necessary to be college and career ready (see **Attachment A**, pp. 7-9, for an example comparison of English language proficiency expectations and proficiency expectations in Reading and Mathematics for a state). This issue will only grow more difficult to address as states begin to implement the expected rigor of the Common Core Standards. Given the time that I am allotted today, I would welcome the opportunity to follow up and elaborate further for the Department the process for defining proficiency in a manner that would support more effective assessment of ELL students.

b. Characterizing Language Needed for Achievement in School: Language Demands and Language Progressions

For ELL students, fluency in conversational English is not sufficient to reduce the achievement gap; they also need to develop proficiency in language that effectively facilitates their access to and achievement of academic content (e.g., English language arts, mathematics, science, career-related content). Language that facilitates ELL student access to and achievement of content can be distinguished from language necessary in other contexts (e.g., social) in terms of its lexical, grammatical, and discourse features (Bailey, 2007; Cummins, 1980; Hutchinson & Waters, 1987; Dudley-Evans & St. John, 1998; Snow, Met, & Genesee, 1992). Failure to appropriately characterize language needed for achievement in school and subsequently provide targeted related language supports to our ELL students places us at risk of perpetuating the existing achievement gap, and our ELL students at risk of being excluded or marginalized from participation in educated society and inhibited or prevented from productively contributing to it (Delpit, 1998; Rumberger & Scarcella, 2000; Scarcella, 2003).

What I will refer to as *language for achievement* is needed by all students for long-term academic success and opportunities for professional growth. I provide a language

taxonomy that has been developed and applied to both English language proficiency and content standards in more than a dozen states. The taxonomy (see **Attachment B**, pp. 10-12) specifies elements of language in a manner that supports the generation of instructional and assessment tasks that educators can systematically use to facilitate the development of their students' English language needed for achievement—that is, students' language for achievement. This taxonomy also lends itself to establishing patterns of student development of English language proficiency (*language progressions*) and levels of attainment necessary to support achievement in the preK-20 context. Patterns of language development and the nature of language proficiency needed to successfully support achievement in school are useful for guiding decisions that better target instruction to students' strengths and weaknesses, as well as for informing assessment design and the scoring and interpretation of assessment results (Pellegrino, 2006).

The upfront considerations that I have highlighted here—defining “proficiency” and characterizing language for achievement in school—can help to guide the design and development of assessments in each content area such that they can be valid and reliable for ELL students, as well as help to purposefully account for the variations in English language proficiency of students in a manner that enables them to demonstrate their knowledge and skills related to core academic content and career readiness. I would welcome the opportunity to engage in a more detailed discussion with the Department about this work and its implications for the design and development of Race to the Top Assessments that are appropriate for ELL students.

2. Considerations for Development and Implementation

a. Linguistic Modification: Facilitating Access to Content for ELL Students at the “Intermediate” and “Advanced” Levels of English Language Proficiency

Both theory and research suggest that students, especially ELL students, could be constrained in showing what they know and can do if the test items used to assess their achievement are measuring factors other than their targeted content related knowledge and skills (Abedi, Hofstetter, and Lord 2004; Butler and Stevens 2001). For example, the complexity of the language (language load) associated with a test item might interfere with students' ability to demonstrate their understanding of the concepts being assessed (Rivera and Stansfield 2001). This interference has been found to be most pronounced for students with limited English proficiency, such as ELL students and non-ELL students who fail to achieve proficiency on state English language arts assessments (Abedi 2001). Research has shown that test items can be linguistically modified to reduce the complexity of the language used, without altering the grade-level construct being assessed; we at Regional Educational Laboratory-West (REL-W) have recently successfully completed such a study (Abedi and Lord 2001; Abedi, Courtney, and Leon 2003; Sato, Rabinowitz, Gallagher, & Huang, in press). Linguistic modification can support the development of more valid and reliable measures of what ELL students know and can do and more appropriate, meaningful comparisons of test scores from ELL and non-ELL students. (See **Attachment C**, pp. 13-17, for research-supported linguistic modification guidelines and strategies.)

b. Computer-based Tests: Providing Access and Balancing Flexibility and Standardization for All ELL Students

There has been a notable increase in general access to and use of computers by students (Goldberg, Russell, & Cook, 2003), and research suggests that computer-based tests (CBTs) can include features and functions that hold promise for use with special student populations (e.g., ELL students, student with disabilities) (Hart & Poggio, 2006; Poggio, Glasnapp, Yang, & Poggio, 2005; Thompson, Thurlow, & Moore, 2003). CBTs can identify a student's current level of language development and academic achievement (i.e., strengths and weaknesses) so that subsequent instruction can be adapted to help the student achieve intended learning and language objectives. CBTs also can include purposeful flexibility of administration (e.g., presentation and response supports such as graphics, audio, glossary of selected key terminology) so that, to the degree students need customized supports to be able to most accurately demonstrate what they know and can do, students can receive needed supports within the parameters allowable for "standardized" assessments for accountability purposes. That is, supports can be purposefully selected and made available in a manner that maintains a desired level of reliability and validity of the measure. Since language development generally is cumulative and dynamic (Riches & Genesee, 2006), and achievement is affected by students' language competencies, CBTs, with their typically immediate feedback on student performance, can help educators better understand and address in a timely manner the changing needs of ELL students as they develop English language proficiency and subsequently their academic achievement. (Note that I am referring to computer *based* tests, not computer *adaptive* tests [CAT]. CBTs are a broader category of testing that will allow the types of accommodations discussed above; whereas CAT, in its typical form, will not.)

As with the upfront considerations highlighted, I would welcome the opportunity to engage in a more detailed discussion with the Department about linguistic modification and computer based tests and their implications for the development and implementation of Race to the Top Assessments that are appropriate for ELL students. In conjunction with the definition of proficiency and characterization of language for achievement discussed previously (see Upfront Considerations), the Race to the Top Assessment Program has the potential to be designed, developed, and implemented to be appropriate and accessible to the full range of ELL students in our nation and support their achievement of rigorous content standards (e.g., Common Core Standards). At WestEd, we have been involved in the development of content and language proficiency assessments for numerous states. Our test development practices are informed by our research on special student populations and the technical requirements of assessment for these students, and our practices are sensitive to students' access needs. I would welcome additional time with the Department to share "lessons learned" and "trade-offs" (e.g., technical, logistical, financial) related to the valid and reliable assessment of ELL students.

3. Evaluation of Consequences

The Framework for High-Quality English Language Proficiency Standards and Assessments: A Tool for Guiding the Evaluation of Intended and Unintended Outcomes for ELL Students

As states and their districts have been developing or refining their assessment, accountability, and support systems for ELL students, the Assessment and Accountability Comprehensive Center (AACC), taking the lead role in partnership with regional comprehensive centers and other technical assistance providers, has helped states evaluate the effectiveness of their assessment systems vis-à-vis support for their ELL students. Although the *Framework for High-Quality English Language Proficiency Standards and Assessments* (Framework) focuses on English language proficiency (ELP) standards and assessments, it lends itself to a broader evaluation of assessments for ELL students and ways in which various types of assessments for ELL students (ELP and content) can be effectively coordinated (a downloadable PDF of the Framework is available at: http://www.aacompcenter.org/cs/aacc/print/htdocs/aacc/resources_sp.htm)

Key questions relevant to the evaluation of consequences include:

- Does the state have systems and structures for monitoring and improving the quality of its assessments, including a plan for ongoing procedures to maintain and improve alignment over time between the state's ELP assessments and content assessments?
- Is the state implementing ongoing quality control reviews to ensure that the system remains fully aligned over time (ELP and content)?
- Does the state rely on multiple sources of data/information (e.g., internal and external monitoring, qualitative data/analyses, quantitative data/analyses) for evaluating the quality and effectiveness of its assessment system (ELP and content)?
- Does the state have a process for using the information gained through its series of studies related to validity, reliability, fairness/accessibility, and alignment/linkage to eliminate gaps and address weaknesses, and does the state have a plan for regular quality review?
- Does the state have a process and schedule for monitoring the implementation of its assessments (ELP and content) and related consequences?
- Does the state help to ensure valid inferences and interpretation of assessment results? Do the state ELP and content assessments produce intended consequences, and have unintended consequences been considered and proactively and appropriately addressed?
- Does the state have a process for examining any accommodations used in terms of their appropriateness vis-à-vis ELL students, the degree of effectiveness of specific accommodations for different groups of EL students, their impact on the assessed constructs, and the inferences based on student performance on accommodated assessments?
- Does the state have a plan in place to support teacher in-service and professional development to ensure proper administration and interpretation of assessments and their results, including how to use such results to guide instruction?

I would welcome the opportunity to discuss in greater detail the implementation implications related to the questions I have just listed as well as the applicability of the Framework more generally to examining implementation challenges, “trade-offs”, and viable strategies for supporting the valid and reliable assessment of ELL students.

Thank you again for this opportunity to provide input. Given my allotted time, I hope that I have been able to highlight some information and resources available to inform the Race to the Top Assessment Program, particularly as it can serve the needs of our ELL students and effectively support their achievement.

A list of work cited is available upon request.

ATTACHMENT A: Example Comparison of a State’s English Language Proficiency Expectations and Proficiency Expectations in Reading and Mathematics (Grades 5 and 8)

Guiding Questions: How is “proficiency” defined? To what degree and how do/should the definitions of and expectations for proficiency be consistent and “align” across English language proficiency and the content areas? Is there sufficient information/detail to determine appropriate and adequate “alignment”?

English Language Proficiency	MEAP Reading	MEAP Reading
<p>Level 4 (Intermediate): Transitional Intermediate At this level students’ language skills are adequate for most day-to-day communication needs. Occasional structural and lexical errors occur. Students may have difficulty using and understanding idioms, figures of speech and words with multiple meanings. They communicate in English in new or unfamiliar settings, but have <u>occasional difficulty with complex structures and abstract academic concepts</u>. Students at this level may read a wide range of texts with considerable fluency and are able to <u>locate and identify the specific facts</u> within the texts. However, they <u>may not understand texts in which the concepts are presented in a de-contextualized manner, the sentence structure is complex, or the vocabulary is abstract</u>. They can read independently, but may have <u>occasional comprehension problems</u>. They produce written text independently for personal and academic purposes. <u>Structures, vocabulary and overall organization approximate the writing of native speakers of English</u>. However, errors may persist in one or more of these domains (listening, speaking, reading, and writing). (TESOL, 1999, p. 21)</p>	<p>Grade 5: “Met” Retells, summarizes, and builds inferences from text. Identifies and explains relationships among characters and themes within and across texts. Effectively addresses specific cross-text task, making connections and revealing understanding despite possible minor misconceptions. Demonstrates knowledge of different genres, including purposes, text elements, and features. Identifies author’s purpose and use of text elements and features to convey meaning. Uses syntax, semantic, and structural cues to determine meaning of some unknown words and phrases and multiple meaning words.</p>	<p>Grade 8: “Met” Builds inferences, summarizes, and applies knowledge from text. Connects relationships, themes, perspectives and universal truths within and across texts. Effectively addresses specific cross-text task, revealing overall understanding despite possible minor misconceptions. Demonstrates knowledge of different genres, including purposes, text elements, and features. Identifies how authors use text elements and features to enhance meaning and to make content accessible to readers. Determines meaning of some unfamiliar words and phrases and multiple meaning words encountered in context.</p>
<p>Level 5 (Proficient): Monitored (Advanced Proficiency) Students at this advanced level have demonstrated English proficiency as determined by state assessment instruments (English Language Proficiency Test - ELPT). They are expected to be able to <u>participate fully with their peers in grade level content area classes</u>. The academic performance of these students is monitored for two years as required by federal</p>	<p>Grade 5: “Exceeded Standards” Demonstrates insightful and accurate understanding by building inferences and making thorough connections within and across texts. Accurately explains relationships among texts, characters and themes within and across texts. Responds to specific cross-text task thoroughly and</p>	<p>Grade 8: “Exceeded Standards” Demonstrates insightful and accurate understanding by synthesizing and applying knowledge from text. Accurately, insightfully, and thoroughly synthesizes and applies knowledge gained from themes, perspectives, or universal truths within</p>

<p>law.</p>	<p>effectively without misconceptions. Shows understanding of different genres, including purposes, text elements, and features. Analyzes and evaluates author’s purpose and use of text elements and features to enhance meaning. Uses context clues to determine meaning of unfamiliar words and phrases and multiple meaning words in context.</p>	<p>and across texts. Responds to specific cross-text task thoroughly and insightfully without misconceptions. Analyzes purposes, text elements, and features of different genres. Analyzes and evaluates how authors use text elements and features to enhance meaning and to make content accessible to readers. Integrates multiple strategies to determine meaning of unfamiliar words and phrases and multiple meaning words in context.</p>
-------------	---	--

English Language Proficiency	MEAP Math	MEAP Math
<p>Level 4 (Intermediate): Transitional Intermediate At this level students’ language skills are adequate for most day-to-day communication needs. Occasional structural and lexical errors occur. Students may have difficulty using and understanding idioms, figures of speech and words with multiple meanings. They communicate in English in new or unfamiliar settings, but have <u>occasional difficulty with complex structures and abstract academic concepts</u>. Students at this level may read a wide range of texts with considerable fluency and are able to <u>locate and identify the specific facts</u> within the texts. However, they <u>may not understand texts in which the concepts are presented in a de-contextualized manner, the sentence structure is complex, or the vocabulary is abstract</u>. They can read independently, but may have <u>occasional comprehension problems</u>. They produce written text independently for personal and academic purposes. <u>Structures, vocabulary and overall organization approximate the writing of native speakers of English</u>. However, errors may persist in one or more of these domains (listening, speaking, reading, and writing). (TESOL, 1999, p. 21)</p>	<p>Grade 5: “Met” Perform basic (e.g., addition, subtraction) and complex (e.g., multiplication, and division) operations with whole numbers; perform basic operations with simple decimals and fractions. Show sufficient understanding of relationships between place-value and decimals, that decimals and fractions are parts of a whole, and that simple fractions and decimals are interchangeable. Apply to basic, routine real-world problems. Identify the fundamental characteristics and properties (e.g., symmetry) of two-dimensional and three-dimensional geometric shapes; recognize basic transformations of geometric shapes. Demonstrate understanding in reading, constructing, and interpreting simple and complex tables and bar graphs. Demonstrate appropriate application of basic measurement concepts (e.g., use of measurement tools such as rulers and thermometers, simple conversions) as applied temperature, perimeter and area of squares and rectangles in solving routine problems. Solve problems including the application of appropriate</p>	<p>Grade 8: “Met” Operate fluently on negative rational numbers, including addition, subtraction, multiplication, and division and can estimate all four. Understand and apply order of operations. Solve simple proportion problems and proportion equations in the form of $a c b d =$ Compute and estimate square roots. Calculate rates of change. Make connections between graphs, tables, and equations of a linear relationship, including directly proportional relationships. Solve applied problems involving linear and directly proportional relationships. Simplify algebraic expressions using addition and subtraction. Apply basic properties of real numbers. Solve problems involving similar figures and scale drawings. Understand relationships between similar figures, including angle, sides, area, and scale factor. Informally show that two triangles are similar</p>

	<p>operations, measurement, or estimation in each of the mathematics content strands; written solutions are organized and presented with both supporting information and explanations of how they were achieved.</p>	<p>using AAA and SSS. Interpret data using circle graphs, stem and leaf plots, histograms, and box-and-whisker plots. Represent data using histograms, stem-and-leaf plots, and box-and-whisker plots. Calculate relative and cumulative frequencies. Calculate mean, median, quartiles, interquartile range.</p>
<p>Level 5 (Proficient): Monitored (Advanced Proficiency) Students at this advanced level have demonstrated English proficiency as determined by state assessment instruments (English Language Proficiency Test - ELPT). They are expected to be able to <u>participate fully with their peers in grade level content area classes</u>. The academic performance of these students is monitored for two years as required by federal law.</p>	<p>Grade 5: “Exceeded Standards” Perform basic and complex operations with whole numbers and with simple decimals and fractions. Show substantial understanding of relationships between place-value and decimals, that decimals and fractions are parts of a whole, and that fractions and decimals are interchangeable. Apply to complex, non-routine real-world problems. Identify and apply the fundamental characteristics and properties (e.g., symmetry) of two-dimensional and three-dimensional geometric shapes; recognize complex transformations of geometric shapes. Demonstrate understanding in reading, constructing, and interpreting simple and complex tables and bar graphs; able to draw conclusions and/or make predictions from data. Demonstrate appropriate application of measurement concepts (e.g., use of measurement tools, simple conversions) as applied to temperature, perimeter and area of squares and rectangles in solving multistep, non-routine problems. Solve problems including the application of appropriate operations, measurement, or estimation in each of the mathematics content strands; written solutions go beyond the obvious and are organized and presented with both supporting information and thorough explanations of how they were achieved.</p>	<p>Grade 8: “Exceeded Standards” Formulate algorithms. Apply operations in more complex situations. Convert ratio quantities between different systems of units. Given a proportional situation, formulate multiple strategies to solve. Compute and estimate cube roots. Apply basic properties of real numbers using algebraic expressions. Simplify algebraic expressions (including the distributive property) and justify reasoning. Informally show that two triangles are similar using SAS. Use similarity properties to justify arguments (including AAA, SSS, SAS). Interpret relative and cumulative frequencies. Given data, select appropriate representations. Represent data using circle graph. Compare and contrast the use of mean and median.</p>

ATTACHMENT B: Language for Achievement—Language Demands—Academic English Language Functions

Academic English Language Function	Operational Definition—The language needed to <i>engage with and achieve</i> in the content (standard or item) consists of the use of:	
A	Identification	a word or phrase to name an object, action, event, idea, fact, problem, need, or process.
	Labeling	a word or phrase to name an object, action, event, or idea.
	Enumeration	words or phrases to name distinct objects, actions, events, or ideas in a series, set, or in steps.
B	Classification	words, phrases, or sentences to assign/associate an object, action, event, or idea to the category or type to which it belongs.
	Sequencing	words, phrases, or sentences to express the order of information (e.g., a series of objects, actions, events, ideas). Discourse markers include adverbials such as <i>first, next, then, finally</i> .
	Organization	words, phrases, or sentences to express relationships between/among objects, actions, events, or ideas, or the structure or arrangement of information. Discourse markers include coordinating conjunctions such as <i>and, but, yet, or</i> , and adverbials such as <i>first, next, then, finally</i> .
C	Comparison/ Contrast	words, phrases, or sentences to express similarities and/or differences, or to distinguish between two or more objects, actions, events, or ideas. Discourse markers include coordinating conjunctions <i>and, but, yet, or</i> , and adverbials such as <i>similarly, likewise, in contrast, instead, despite this</i> .
D	Inquiring	words, phrases, or sentences to solicit information (e.g., <i>yes-no</i> questions, <i>wh</i> -questions, statements used as questions).
E	Description	word, phrase, or sentence to express or observe the attributes or properties of an object, action, event, idea, or solution.
F	Definition	word, phrase, or sentence to express the meaning of a given word, phrase, or expression.
G	Explanation	phrases or sentences to express the rationale, reasons, causes, or relationships related to one or more actions, events, ideas, or processes. Discourse markers include coordinating conjunctions <i>so, for</i> , and adverbials such as <i>therefore, as a result, for that reason</i> .
H	Retelling	phrases or sentences to relate or repeat information. Discourse markers include coordinating conjunctions such as <i>and, but</i> , and adverbials such as <i>first, next, then, finally</i> .
	Summarization	phrases or sentences to express important facts or ideas and relevant details about one or more objects, actions, events, ideas, or processes. Discourse structures include: beginning with an introductory sentence that specifies purpose or topic.
I	Interpretation	phrases, sentences, or symbols to express understanding of the intended or alternate meaning of information.
J	Analyzing	phrases or sentences to indicate parts of a whole and/or the relationship between/among parts of an action, event, idea, or process. Relationship verbs such as <i>contain, entail, consist of</i> , partitives such as <i>a part of, a segment of</i> , and quantifiers such as <i>some, a good number of, almost all, a few, hardly any</i> often are used.

Academic English Language Function	Operational Definition—The language needed to <i>engage with and achieve</i> in the content (standard or item) consists of the use of:	
K	Generalization	phrases or sentences to express an opinion, principle, trend, or conclusion that is based on facts, statistics, or other information, and/or to extend that opinion/principle/etc. to other relevant situations/context/etc.
	Inferring	words, phrases, or sentences to express understanding of implied/implicit based on available information. Discourse markers include inferential logical connectors such as <i>although, while, thus, therefore</i> .
	Prediction	words, phrases, or sentences to express an idea or notion about a future action or event based on available information. Discourse markers include adverbials such as <i>maybe, perhaps, obviously, evidently</i> .
	Hypothesizing	phrases or sentences to express an idea/expectation or possible outcome based on available information. Discourse markers include adverbials such as <i>generally, typically, obviously, evidently</i> .
L	Argumentation	phrases or sentences to present a point of view with the intent of communicating or supporting a particular position or conviction. Discourse structures include expressions such as <i>in my opinion, it seems to me</i> , and adverbials such as <i>since, because, although, however</i> .
	Persuasion	phrases or sentences to present ideas, opinions, and/or principles with the intent of creating agreement around or convincing others of a position or conviction. Discourse markers include expressions such as <i>in my opinion, it seems to me</i> , and adverbials such as <i>since, because, although, however</i> .
	Negotiation	phrases or sentences to engage in a discussion with the purpose of creating mutual agreement from two or more different points of view.
M	Synthesizing	phrases or sentences to express, describe, or explain relationships among two or more ideas. Relationship verbs such as <i>contain, entail, consist of</i> , partitives such as <i>a part of, a segment of</i> , and quantifiers such as <i>some, a good number of, almost all, a few, hardly any</i> often are used.
N	Critiquing	phrases or sentences to express a focused review or analysis of an object, action, event, idea, or text.
O	Evaluation	phrases or sentences to express a judgment about the meaning, importance, or significance of an action, event, idea, or text.
P	Symbolization & Representation	symbols, numerals, and letters, to represent meaning within a conventional context (e.g., +, -, CO ₂ , >, Δ, π, cos, y=3x+4, c ² =a ² +b ² , h/2(b ₁ +b ₂), <i>cat</i> vs. <i>cat</i>).
Z	No Academic Language Function	Item or standard does not contain <i>any</i> academic language functions; may contain linguistic skills (e.g., phonemic awareness, syllabication).

Note: This taxonomy focuses on academic language functions and does not address the identification or definition of linguistic skills (e.g., phonology, morphology).

Language for Achievement—Language Complexity

Language complexity is influenced by both density and construction as defined below.

Density

Low	High
<ul style="list-style-type: none"> • Length ranges from a word to paragraphs • No/little variation in words and/or phrases in sentences/paragraphs; consistent use of language • Repetition of key words/phrases/sentences <i>reinforces</i> information • Language is used to present critical/central details • No/little abstraction; language reflects more literal/concrete information; illustrative language is used; language is used to define/explain abstract information • Graphics and/or relevant text features reinforce critical information/details 	<ul style="list-style-type: none"> • Length ranges from a word to paragraphs • Some variation in words and/or phrases in sentences/paragraphs • Repetition of key words/phrases/sentences <i>introduces new or extends</i> information • Language is used to present critical/central details, but non-essential detail also is presented • Some abstraction; language <i>may or may not</i> be used to define/explain abstract information; illustrative language <i>may or may not</i> be used; technical words/phrases are used • Graphics and/or relevant text features <i>may or may not</i> reinforce critical information/details

Construction

Simple	Complex
<ul style="list-style-type: none"> • Mostly common/familiar words/phrases; no/few uncommon words/phrases, compound words, gerunds, figurative language, and/or idioms • Language is organized/structured • Mostly simple sentence construction • No/little passive voice • Little variation in tense • Mostly one idea/detail per sentence • Mostly familiar construction (e.g., 's for possessive; s and es for plural) • Mostly familiar text features (e.g., bulleted lists, bold face) 	<ul style="list-style-type: none"> • Some common/familiar words/phrases; some uncommon words/phrases, compound words, gerunds, figurative language, and/or idioms • Language <i>may or may not</i> be organized/structured • Varied sentence construction, including complex sentence construction • Some passive voice • Variation in tense • Multiple ideas/details per sentence • Some less familiar/irregular construction • Some less familiar text features (e.g., pronunciation keys, text boxes)

This Language for Achievement Taxonomy focuses on academic language functions (as opposed to social language, linguistic skills, academic lexicon). It is theory and research based, and it has been used in the evaluation of standards and assessments in a number of states. For more information, please contact Dr. Edynn Sato at WestEd.

Definition from the *Framework for High-Quality ELP Standards and Assessments* (AACC, 2009):

Academic language, broadly defined, includes the language students need to meaningfully engage with academic content within the academic context. This should *not* be interpreted to suggest that separate word lists and/or definitions of content-related language should be developed for each academic subject. Rather, academic language includes the words, grammatical structures, and discourse markers needed in, for example, describing, sequencing, summarizing, and evaluating — these are language demands (skills, knowledge) that facilitate student access to and engagement with grade-level academic content. These academic language demands are different from cognitive demands (e.g., per Bloom’s taxonomy). Although there may not be just one accepted definition of academic language, there are a good number of resources available that address the issue of academic language and may be considered in the development of state ELP standards and assessments. For example: Aguirre-Munoz, Parks, Benner, Amabisca, & Boscardin, 2006; Bailey, 2007; Bailey, Butler, & Sato, 2007; Butler, Bailey, Stevens, Huang, & Lord, 2004; Chamot & O’Malley, 1994; Cummins, 1980; Cummins, 2005; Halliday, 1994; Sato, 2007; Scarcella & Zimmerman, 1998; Schleppegrell, 2001.

ATTACHMENT C: Linguistic modification guidelines and strategies

Desirable characteristics	Notes on approaches and criteria
<i>Item context</i>	
<ul style="list-style-type: none"> • Familiar to students. • No cultural or linguistic bias. • Minimal construct (no irrelevant words or phrases). 	<ul style="list-style-type: none"> • The context situates the problem (and may include description of relationship or interaction between location and time). • In the body of the report, context is often described in relation to its complexity and as part of biased or construct-irrelevant information that should be pruned out. Recommendations: <ul style="list-style-type: none"> ○ Remove passive voice construction in original item. ○ Remove past tense and conditional in original item. ○ Break stem into shorter, less complex sentences (sometimes a series of shorter sentences can create a story line or present a more familiar context/situation to students). • Context can provide description that helps make abstract or highly generalized situations more concrete and relevant. Simply stated, it helps to ground the content being tested. Context that facilitates access for English language learner students is expressed in concrete language, illustrative language, and illustrations/graphics.

Desirable characteristics	Notes on approaches and criteria
<i>Item graphics</i>	
<ul style="list-style-type: none"> • Familiar to students. • No cultural or linguistic bias. • Symbols, legends, and key vocabulary relevant to the construct and familiar to English language learner students. • Consistent graphic and labeling/naming conventions • Supportive of English language learner student understanding of assessed content. 	<ul style="list-style-type: none"> • Graphics include diagrams, tables, charts, drawings, graphs, pictures, and maps. • Student knowledge about certain graphics is required and assessed in mathematics. • Graphics allow for reduced amount or complexity of language in a test item. Use of graphics in test items should serve a clear purpose. Otherwise they may be misleading or distracting. For example, graphics may be used to: <ul style="list-style-type: none"> ○ Clarify key aspects of the content/construct assessed. ○ Clarify construct-relevant context. ○ Clarify a mathematical operation. ○ Indicate what the student is expected to do. ○ Help students shift from one context to another within an assessment (for example, from one type of test item to another). ○ Allow students to reinforce or verify understanding of key information in test item. ○ Simplify the structure of a test item that requires a number of operations or steps (for example, through bulleted lists or a diagram of the complete problem that accurately reflects the problem in its totality). • Some criteria that can be used to evaluate the need for a graphic include: <ul style="list-style-type: none"> ○ Does the graphic clarify construct-irrelevant information? If so, it may not be necessary. It might be better to revise or delete the construct-irrelevant information. ○ Does the graphic support the test item context without requiring additional written text? ○ Does the graphic accurately represent the full complexity of the problem? If not, it may be misleading. ○ Is the graphic consistent with the key content/construct of the item?

Desirable characteristics	Notes on approaches and criteria
<i>Item vocabulary/wording</i>	
<ul style="list-style-type: none"> • High-frequency words • Common and familiar words • Relevant technical terms that reflect language of the content standards and academic English language. • Technical terms defined, as appropriate • Naming conventions consistent with graphics/stimuli • Construct-irrelevant vocabulary/phrases at or below grade level 	<ul style="list-style-type: none"> • Careful selection of vocabulary and phrases can simplify sentence structure. The amount and complexity of language should be balanced with the amount of information necessary for student to understand/access the item. The goal is to make the language as clear and straightforward as possible, while still providing the amount and complexity of information necessary to communicate the targeted content of the test item. • Some general guidelines: <ul style="list-style-type: none"> ○ Use precise language. Appropriate language modification does not simply mean using common or familiar vocabulary. ○ Consider language used in the content standards and academic English language . ○ Repeat key words/phrases in the test item that students need to understand the item and respond to it. ○ Do not automatically provide synonyms for a key word. This may not be helpful, especially if a test item is already long or complex. Although providing synonyms may be helpful during instruction, it may not be useful in assessment items. ○ Use words/phrases consistently within the context of the item and consider consistency of terms within a strand—for example, reading or measurement). Support this use with context-familiar content-based abbreviations and make explicit connections between terms/abbreviations. • If possible, avoid using: <ul style="list-style-type: none"> ○ Ambiguous words or unnecessary words with multiple meanings. ○ Irregularly spelled words. ○ Proper nouns that are irrelevant or not meaningful to the population. ○ Words that are both nouns and verbs (for example, carpet, value, cost); however, if a choice needs to be made, use the word only as a noun. ○ Hyphenated and compound words ○ Gerunds. ○ Relative pronouns (for example, which, who, that) without a clear antecedent.

Desirable characteristics	Notes on approaches and criteria
<i>Item sentence structure</i>	
<ul style="list-style-type: none"> • Familiar, common sentence structure. • Complexity of sentence structure at or below grade level. • Key information presented first or early in the test item. • One sentence per idea for complex test items. 	<ul style="list-style-type: none"> • To reduce the complexity of a sentence in a test item: <ul style="list-style-type: none"> ○ Identify the agent (that is, the person or object carrying out the action) to construct sentences that use active voice (and avoid passive voice). ○ Make sure that the verb in a sentence follows the subject as closely as possible. ○ Remove introductory phrases that are irrelevant to the construct being tested. ○ Use conventional constructions (for example, apostrophes for possessives and “s” or “es” for plurals). ○ Use proper nouns that students are familiar and are grade-level appropriate. ○ Use clear grammatical structures. • To reduce language load: <ul style="list-style-type: none"> ○ Change past or future tense verb forms to present tense. ○ Change passive verb forms to active verb forms. ○ Change complex sentence structure to subject-verb-object structure. ○ Shorten any long nominals/names/phrases (for example “last year's class vice-president” to “a student leader”). ○ Replace compound sentences with two separate sentences, especially when making comparisons. ○ Shorten or delete long prepositional phrases. ○ Replace conditional clauses with separate sentences. ○ Change the order of a clause within a sentence. ○ Remove or rephrase relative clauses. ○ Rephrase questions framed in negative terms. • Make sure the following are clear. <ul style="list-style-type: none"> ○ Noun-pronoun relationships. ○ Antecedent references.

<i>Item format/style</i>	
<ul style="list-style-type: none"> • Clear parts of the item/question. • Explicit order of operations. • Relevant and appropriate distinctions. • Segmented or shortened long problem statements. 	<ul style="list-style-type: none"> • Place test item elements in the following order: (1) text that introduces the graphic; (2) graphic; and (3) the test item stem. • Format for emphasis of key words/terms (highly construct-relevant), using bold, ALL CAPS, and <u>underline</u> to call English language learner students' attention to them. • Consider whether blocks of text (that is, a paragraph) may be necessary and appropriate for presenting a test item. This depends on the construct assessed, the complexity of the information needed by the student to respond to the item, and the centrality of the context to the construct. Suggested strategies to help English language learner students process such text include: <ul style="list-style-type: none"> ○ Bulleted lists. ○ Indenting key information. ○ Emphasizing key words/terms. ○ Using graphics.

Source: Sato (2008)



Testimony to the U.S. Department of Education Race to the Top Assessment Program

Public & Expert Input Meeting on Assessing English Language Learners

Denver, Colorado – December 2, 2009

Good morning. My name is Jerome Shaw and I'm currently a faculty member in science education at the University of California's Santa Cruz campus. Previously, I was a teacher of science to English Language Learners (ELLs) in California K-12 public schools. My comments to you today come from my combined experience as a classroom teacher and educational researcher. Thank you for the opportunity to share these thoughts.

Cutting to the chase, my recommendation is this: in the interest of fair and accurate assessment of English Language Learners, the Race to the Top Assessment Program *must* include carefully crafted performance assessments. I'll contextualize my rationale for this recommendation by way of describing a particular assessment and some findings on its use with ELLs.

A few years back, here in Colorado, I had the privilege of observing several fifth grade classrooms as students engaged with a trio of curriculum-embedded science performance assessments. With support from the National Science Foundation, these tasks were developed locally by teams that included assessment and science content experts as well as classroom teachers. Teachers also received professional development on the assessments and their affiliated instructional units.

One of these assessments was the culminating task for a unit titled Food Chemistry during which students learned about nutrition and ways to test foods for various nutrients. For the assessment, students examined a previously untested group of food items and determined which one they felt was the most nutritious snack. Students worked in pairs or small groups to conduct the investigation, then shared their findings and defended their choice in an oral interview with their teacher. From the start, students were provided with rubrics to help guide their actions toward desired outcomes. In addition to teacher scoring, students also used the rubrics to rate their own performance at the end of the assessment.

In many ways, this assessment embodies what I mean by "carefully crafted." Key criteria are that an assessment:

- Be aligned with curriculum and instruction.
- Involve teachers in the development process.
- Support teachers in proper implementation.
- Allow multiple forms of expression (e.g., written, oral, graphic).

- Engage students in self-assessment.

An analysis I conducted of student scores on the aforementioned group of fifth grade science performance assessments showed ELLs performing on par with their non-ELL peers. This study was limited in scope and additional research needs to be done to investigate the generalizability of these results. Nonetheless, these preliminary findings point to the promise of performance assessments; they have the potential to level the playing field for diverse students.

With respect to ELLs, there are caveats. Contrary to what may seem popular belief, performance assessments are not necessarily more accessible to or fair for ELLs (see for example Shaw, 1997). The language that such tasks call upon students to comprehend and produce can present serious challenges. To better understand these challenges, colleagues at UC Santa Cruz and I recently developed an analytical framework that focuses on the functional and interactive uses of language inherent in science performance assessments (Bunch, Shaw, & Geaney, in press). Rather than seen only as daunting, we stress that these demands also represent opportunities for ELLs to develop language and demonstrate their understanding in myriad, authentic ways.

This work complements other analyses that examine lexical or syntactical features of assessments such as difficulty of vocabulary and sentence complexity. Together, these various approaches can provide a comprehensive understanding of language demands and inform classroom practitioners and assessment developers alike.

In sum, a dearth of findings indicate the potential for, and ways in which, performance assessments may contribute to fair and accurate assessment of ELLs. More work needs to be done. The Race to the Top Assessment Program offers a timely opportunity to move forward on this important front.

Thank you for your time.

References

- Bunch, G. C., Shaw, J. M., & Geaney, E. R. (in press). Documenting the language demands of mainstream content-area assessment for English learners: Participant structures, communicative modes, and genre in science performance assessments. Accepted for publication in *Language and Education*.
- Shaw, J. M. (1997). Threats to the validity of science performance assessments for English language learners. *Journal of Research in Science Teaching*, 34(7), 721-743.
- Shaw, J. M. (2009). Science performance assessment and English learners: An exploratory study. *Electronic Journal of Literacy Through Science*, 8(3).

From: Teri Siskind [mailto:tsiskind@ed.sc.gov]
Sent: Wednesday, December 02, 2009 4:54 PM
To: Race To The Top Assessment Input
Cc: Janice Poda; Karla Hawkins; Teri Siskind
Subject: Race to the Top Assessment Program

The South Carolina Department of Education (SCDE) welcomes the opportunity to comment on the Race to the Top Assessment Program. Our comments address 1) future assessment programs as detailed in the notice of Friday, October 23, and 2) race to the top grant funding for assessments.

Assessment Programs

- SCDE supports the national adoption of a common core of academic content standards and aligned assessments with common academic achievement standards.
- SCDE supports assessment systems that include the summative assessments aligned to a common core, interim assessments, and formative assessment techniques as part of ongoing teaching and learning.
- Ideally, SCDE supports the use of technology to support assessment and instructional systems. South Carolina is piloting the use of various technological media as textbooks and would support a system that would enable each student to use technology in the classroom, as a delivery system for textbooks and other resources, and as a mechanism for assessment.
- SCDE supports the appropriate inclusion and assessment of students with disabilities and English language learners and would like to see continued support of research into the valid assessment of these students. Although universal design principles appear to provide access for many of these students, the SCDE recognizes the complexity of assessing SWD and ELL meaningfully due to a compendium of factors including cultural and psycho-social.
- New assessment systems require time for development and the SCDE supports exemptions from double-testing under ESEA and new development.
- SCDE believes the summative measures should be utilized for state accountability while the interim and formative components focus on instruction and learning. To support this stance, summative measures should be aligned to grade level academic content standards while interim measures could be adaptive in nature. Not every student need be tested on the same identical content for a summative state accountability system and, in time, students should be sampled for a range of subject matter. A broad array of measures - including long-term projects - eliciting complex responses should be emphasized for interim and formative assessment. Extensive support for teachers in developing and scoring more complex measures is essential.

Race to the Top Grant Funding

- \$350 million is not an adequate amount to support development and sustain systems over time, especially if half of the money goes to LEAs.
- Development of a new system based on a common core of standards adopted by a consortium will be a long-term process and the period of

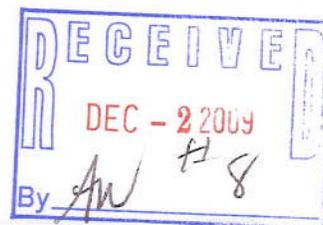
award is not sufficient. The timeline is critical however. Assessments cannot be developed until the academic content standards are adopted, yet once academic content standards are implemented, the assessments that are administered to students should be aligned to the standards that are being implemented.

- SCDE supports SCDE supports the national adoption of a common core of academic content standards and aligned assessments with common academic achievement standards, thereby supporting a single award for a summative accountability measure. However, SCDE would support competitive awards for more inclusive assessment systems as those previously described.
- Management of a consortium project is crucial and the details should be required and highlighted in proposals.
- USED staff should maintain constant contact, communication, and oversight of all awards.

Theresa Siskind
Deputy Superintendent
Division of Accountability
State Department of Education
1112 Rutledge Building
1429 Senate Street
Columbia, SC 29201
tsiskind@ed.sc.gov
803-734-8396 (Phone)
803-734-4480(FAX)

This message is intended only for the use of the individual or entity to

which it is addressed and may contain information that is privileged, confidential and exempt from disclosure under applicable law. If the reader of this message is not the intended recipient, you are hereby notified that any dissemination, distribution, or copying of this communication is strictly prohibited by law. If you have received this communication in error, please notify me immediately.



Assessment of English Language Learners

Guillermo Solano-Flores
University of Colorado at Boulder

Race to the Top Assessment Program Public & Expert
Input Meeting
Denver – Assessment of English Language Learners
Wednesday, December 2, 2009

Major threats to valid ELL testing

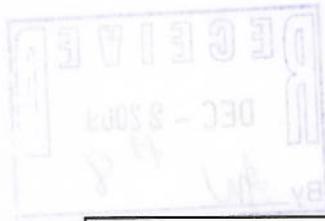
- Population misspecification
- Measurement error introduced by assessment systems
- Overgeneralization

Language variation and score variation

- Each ELL has a unique set of strengths and weaknesses in L1 and a unique set of strengths and weaknesses in L2
- Each item poses a unique set of linguistic challenges in each language
- Linguistically, ELL populations vary tremendously as to their English proficiency—even within a given broad linguistic group
- Linguistic variation is responsible for a considerable amount of measurement error in ELL testing

Language in testing

- Testing is a communication process
 - assessments system ask questions (items)
 - students answer those questions
 - assessment systems interpret those answers
- Critical in this communication process are
 - the students' proficiency in the language of testing
 - the ability of the assessment systems to communicate with students



Language of testing

- Even within the same broad linguistic group, communities may vary as to whether more dependable measures can be obtained by testing ELLs in L1 or in L2
- To a great extent, successfully testing ELLs in L1 or providing them with accommodations in L1 depends on the linguistic resources available in the assessment systems (e.g., proficiency in ELLs' L1 among the professionals who provide certain accommodations)
- Localization has a great potential as an approach for addressing language variation due to dialect (in both L1 and L2)

Design and implementation issues

- State assessment design should be informed and supported by empirical studies that address:
 - To what extent do ELL groups vary as to the minimum test length needed to obtain dependable measures?
 - To what extent is measurement error in ELL testing introduced by the assessment system (e.g., poor implementation of testing accommodations)?
 - How defensible are the testing accommodations used with ELLs?

Thanks!

Guillermo.Solano@Colorado.edu

Language variation and score variation

- Each ELL has a unique set of strengths and weaknesses in L1 and a unique set of strengths and weaknesses in L2
- Each item poses a unique set of linguistic challenges in each language
- Linguistically ELL populations vary tremendously as to their English proficiency—even within a given broad linguistic group
- Linguistic variation is responsible for a considerable amount of measurement error in ELL testing

Who Is Given Tests in What Language by Whom, When, and Where? The Need for Probabilistic Views of Language in the Testing of English Language Learners

Guillermo Solano-Flores

The testing of English language learners (ELLs) is, to a large extent, a random process because of poor implementation and factors that are uncertain or beyond control. Yet current testing practices and policies appear to be based on deterministic views of language and linguistic groups and erroneous assumptions about the capacity of assessment systems to serve ELLs. The question *Who is given tests in what language by whom, when, and where?* provides a conceptual framework for examining testing as a communication process between assessment systems and ELLs. Probabilistic approaches based on generalizability theory—a psychometric theory of measurement error—allow examination of the extent to which assessment systems' inability to effectively communicate with ELLs affects the dependability of academic achievement measures.

Keywords: assessment; bilingual/bicultural; generalizability theory; sociolinguistics; testing; validity/reliability

More than four decades ago, Joshua Fishman (1965) published his famous article "Who Speaks What Language to Whom and When?" In it, he provides an important notion in the study of bilingualism from a sociolinguistic perspective—that language choice among bilingual individuals is shaped by the interaction of multiple factors, such as the domain of the language behavior (e.g., at work, with family, with friends), the situation in which communication takes place (e.g., formal, informal), the mode of language used (conversation, reading, or writing), the role of language (thinking, comprehension, or production), and the topic being discussed (e.g., family, weather, job).

The notion that bilingual behavior is shaped by function and context provides a simple but powerful framework for examining why we have failed to create fair and sound testing practices for English language learners (ELLs). Fishman's article title inspired the title of this article.¹ The question *Who is given tests in what language by whom, when, and where?* is used as a conceptual framework for discussing why language in testing should be addressed as a communication process between an assessment system (e.g., National

Assessment of Educational Progress or a state's testing program) and ELL students. In this process, test writers write items that students read and interpret, students write responses to those items, and raters read and interpret those responses.

In this article, I contend that current ELL testing practices are limited in their effectiveness to produce valid measures of academic achievement because they are based on categorical, deterministic views of language and erroneous assumptions about the capacity of assessment systems to effectively communicate with ELL students. Examples of current practices driven by these deterministic views and these erroneous assumptions include classifying ELLs into a few categories of language proficiency, assigning them to treatment conditions as if they were linguistically homogeneous, looking for the form of testing accommodation that works for all ELLs, assessing language development without considering proficiency in the students' first language (L1), and assuming that all schools are equally capable of implementing testing accommodations properly.

I also submit that to be able to produce valid measures of academic achievement for ELLs, deterministic views of language in testing should give way to probabilistic views that address the fact that, because of their social nature, language and linguistic groups are dynamic, not static. Critical for this shift to occur is the recognition that, to a large extent, the process of ELL testing is shaped by factors that are unpredictable or factors that are beyond the control of simple testing policies. Different first languages, different migration histories, different kinds of exposure to formal instruction both in L1 and in the second language (L2), and dialect variation within both L1 and L2 are among the myriad of factors that make ELL populations considerably heterogeneous, even within a given classroom. Also, different tests used by states to measure English proficiency, different criteria used for defining ELLs, and different capabilities of schools to adequately provide testing accommodations for these students are some of the many factors that limit our ability to make valid interpretations of their test scores.

Who is given tests in what language by whom, when, and where? is used as a mapping sentence, a sentence that identifies six interrelated but distinguishable components of the process of ELL testing (Table 1). Mapping sentences have been used successfully in the past as conceptual tools that provide descriptions of a knowledge domain in the context of testing (e.g., Bormuth, 1970; Guttman,

Table 1
Who Is Given Tests in What Language by Whom, When, and Where? Mapping Sentence That Identifies Six Interrelated but Distinguishable Components of the Process of English Language Learner (ELL) Testing

<i>Who</i>	<i>Is Given Tests</i>	<i>in What Language,</i>	<i>by Whom,</i>	<i>When,</i>	<i>and Where?</i>
The student—the object of measurement in the process of ELL testing.	Procedures used to develop, adapt, and administer tests; ways in which tests are administered to ELL students.	Linguistic properties of tests; the language or dialect of a language in which tests are administered.	Linguistic skills of individuals who develop, adapt, and administer tests for ELLs; linguistic skills of individuals who score ELL students' responses to tests.	Time in the process of second-language development at which ELLs are tested; number of occasions on which ELLs are tested.	School communities; sociolinguistic contexts in which ELL students learn.

1969) or as guides for examining how the effectiveness of intervention programs is shaped by multiple contextual factors (e.g., Cunningham & Fitzgerald, 1996).

The first part of this article examines the six components of the process of ELL testing: *who*, *tests*, *language*, *by whom*, *when*, and *where*. The second part discusses how probabilistic views of language can improve the capability of assessment systems to properly address language in ELL testing.

The Process of ELL Testing

Who

Unfortunately, *who*, the ELL student, does not appear to be properly understood and is, to a great extent, an unknown object of measurement. Some difficulties for properly defining and classifying ELLs derive from the intrinsic complexity of the condition of being bilingual.² Multiple patterns of language dominance result from different kinds of language development in L1 and L2 (see Aguirre-Muñoz & Baker, 1997; Baker, 2006; Mackey, 1962; Stevens, Butler, & Castellon-Wellington, 2000). As a consequence, each ELL has a unique set of strengths and weaknesses in each language mode (i.e., listening, reading, speaking, and writing) in L1 and in L2. These strengths and weaknesses also vary by context (e.g., at home, at school, with friends, with relatives; Bialystok, 2001; Mackey, 1962; MacSwan, 2000) and are shaped by schooling (e.g., bilingual or full immersion programs) and the way in which language instruction is implemented (e.g., by emphasizing reading or writing in one language or the other; see Genesee, 1994; Valdés & Figueroa, 1994). Because of practical constraints, rarely are assessments of language development for ELLs comprehensive enough to provide an accurate picture of proficiency in both L1 and L2 and in the four language modes. As a result, language tests may provide fragmented and inconsistent information about the linguistic proficiency of ELLs.

Other difficulties for properly defining and classifying ELLs derive from inaccurate ways in which they are viewed as learners (Lee, Deaktor, Hart, Cuevas, & Enders, 2005). Testing practices reflect cognitive models of both the nature of the learner and the nature of the target skills and abilities being assessed (see Leighton & Gierl, 2007; Pellegrino, Chudowsky, & Glaser, 2001). A sign

that appropriate views of language are not always reflected in the ways in which ELLs are assessed is their overrepresentation in special education programs and their underrepresentation in talented and gifted education programs (Harry & Klingner, 2006). Another sign is that language assessment practices neglect L1 as a source of information about language development. For example, most of the evaluations of bilingual programs are based on data on L2 development but fail to capture information on L1 (Brisk, 2006). By focusing exclusively on L2, tests fail to provide important information about an ELL's language development, and they limit, among other things, the capability of educators to make informed decisions on the kinds of testing accommodations that are suitable for each student.

Inappropriate views of language are also reflected in the ways ELLs are defined and language proficiency is measured. First, legal definitions of ELLs (e.g., No Child Left Behind Act, 2002) focus primarily on demographic factors (age, grade of enrollment, place of birth, migration status) that are associated with the characteristics of the population of ELLs in public schools in the United States but are inaccurate indicators of language proficiency. Although these approaches operationalize decision making for ELLs (e.g., by providing criteria for deciding who should be tested), they are likely to produce inaccurate classifications of students.

Second, mandatory tests of English proficiency are not based on benchmarks of progress toward the development of a second language (Shin, 2004). Rather, these tests are aligned with English language arts standards created with monolingual, English-speaking populations and related to grade level (see Kopriva et al., 2004). Underlying these practices is the assumption that the communicative skills ELLs need to function in school and to benefit from instruction in English can be developed in the same way an individual learns English as a foreign language. But the student learning English as a foreign language is different from the student who is developing it as a second language (see Baker, 2006). The former does so voluntarily by taking elective courses (Valdés & Figueroa, 1994) and is usually an adult who has completed the development of L1. By contrast, an ELL learns L2 as a necessity and uses it as a medium to learn subject content matter while still developing L1 (Solano-Flores & Trumbull, 2008).

In sum, flawed definitions of ELLs and the assessment of their linguistic proficiencies are likely to produce inaccurate classifications of these students.

Tests

Some challenges to properly testing ELL students derive from the intrinsic linguistic features of tests. First, it is widely recognized that a test of any content area is, to some extent, a test of proficiency in the language in which it is administered (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Bernardo, 2002; Thurber, Shinn, & Smolkowski, 2002). Test items tend to consist of relatively small amounts of text, use prose styles that are different from other reading materials, provide little contextual information, contain words with unusual meanings, and have scant continuity of ideas across sentences (see Ferguson & Fairburn, 1985; Noonan, 1990). These seemingly subtle differences can in fact affect the ways in which students read and understand items (De Corte, Verschaffel, & Pauwels, 1990; Shorrocks-Taylor, & Hargreaves, 1999; Solano-Flores & Nelson-Barber, 2001).

Second, certain characteristics of tests and the ways in which they are administered favor communication styles that are not necessarily universal. For example, scoring open-ended items with rubrics that overvalue long written responses to test items may affect the dependability of scores for children from cultural groups in which giving long responses to questions asked by adults is regarded as impolite (see Heath, 1983).

Third, some linguistic demands derive from register—the specialized ways in which concepts and things are referred to within a given context or discipline (Halliday, 1978). They involve much more than technical terms and include sophisticated ways of socializing in a discipline, such as building arguments, expressing disagreement, constructing discourse, or using notation systems (Chamot & O'Malley, 1994; Scarella, 2003). At the lexical level, properly using the register of a discipline poses the challenge of negotiating the meanings of words with dual meanings, terms that differ in meaning and use within a discipline and in everyday language, and terms that are not exclusive to a particular discipline but are critical to reasoning and building arguments (see Wellington & Osborne, 2001). Needless to say, this challenge may be even greater for ELLs who have to deal, among other things, with false cognates and cognates that are used differently across languages in everyday life and in the academic context.

Other concerns stem from the fact that procedures used to develop tests or to adapt tests for ELLs do not properly address language. First, although a great deal of the process of test development has to do with refining the wording of items (Solano-Flores, Trumbull, & Nelson-Barber, 2002), ELLs are frequently underrepresented in the samples of pilot students.

Second, the process of test adaptation tends to change the constructs measured by items (Sireci & Allalouf, 2003), to the extent that a considerable number of translated items function differentially across language groups (Ercikan, Gierl, McCreith, Puhan, & Koh, 2005) because of subtle inaccuracies in test translation that cannot be identified and addressed unless appropriate interview methods are used with samples of the target population

(Ercikan, 1998, 2002). Thus, if ELLs are to be tested in L1, methods of cognitive validation (see Kopriva, 2001) should be part of the procedures used to adapt tests for ELLs.

Third, testing accommodations—changes in tests or test situations made with the intent to address students' linguistic needs without altering the construct being measured (George Washington University Center for Equity and Excellence in Education, 2005)—may be ineffective if they are assumed to be equally appropriate for all ELLs. For example, side-by-side dual-language versions of a test (e.g., García-Duncan et al., 2005; Sireci & Khaliq, 2002) may not be effective for ELLs whose reading proficiency in L1 is limited.

Fourth, other accommodations may be questionable because they do not directly address language issues (e.g., the accommodation of working in small groups; National Assessment of Educational Progress, 2005) or they have been transplanted from the field of special education (e.g., enhanced lighting conditions; see Ferrara, Macmillan & Nathan, 2004). Other accommodations assume that students possess certain skills that are needed to benefit from these accommodations, as is the case with bilingual dictionaries, whose proper use involves word-searching skills (e.g., Stansfield, 2003).

The limited extent to which testing accommodations can address language makes it unlikely that they contribute significantly to producing valid test scores for ELLs. Indeed, there is evidence that providing ELLs with accommodations that do not address their specific needs is no more effective than randomly assigning them to accommodations (Kopriva, Emick, Hipolito-Delgado, & Cameron, 2007). Because states vary tremendously as to the types of testing accommodations they use with ELLs (Rivera, Collum, Willner, & Sia, 2006) and the fidelity with which they implement the accommodations permitted by national assessments (Stansfield & Bowles, 2006), it is difficult to make generalizations about the effectiveness of testing accommodations.

The studies conducted by Abedi and his associates (see Abedi, Hofstetter, & Lord, 2004; Abedi, Lord, Hofstetter, & Baker, 2001; Abedi & Lord, 2001; Buder & Stevens, 1997) show that the linguistic simplification of items can reduce significantly (although moderately) the performance gap between ELL and non-ELL students. An explanation for the effectiveness of this form of accommodation may be not only that the accommodation directly addresses language but also that the fidelity of its implementation cannot be compromised.

In sum, existing procedures for developing, adapting, and administering tests for ELLs are difficult to implement and are not consistently or properly implemented; in addition, their effectiveness is limited by multiple language factors that are beyond control.

Language

Decisions on whether to test ELLs in either L1 or L2 appear to overlook the social dimension of language. Perhaps the most neglected aspect of language is dialect. Dialects are varieties of the same language that differ on such features as pronunciation, vocabulary, grammar, spelling conventions, and the like (Wardhaugh, 2002). Dialects may also differ on such discourse features as ways of organizing oral or written narratives or

responding to questions (Heath, 1983). Because dialects are associated with people from different geographic or social groups (Wolfram, Adger, & Christian 1999), the term *dialect* is sometimes used to characterize a variety of a language as "incorrect" or "bad." However, research in sociolinguistics reveals that non-prestigious dialects are often as sophisticated and complex as the prestigious, standard dialects (Farr & Ball, 1999).

Dialect is a natural occurrence and is not specific to ELLs. Dialect variation may take place among monolingual English speakers from different geographical areas and socioeconomic status, and this variation may be important enough to influence how students interpret test items (Solano-Flores & Trumbull, 2003). In the case of ELLs, important dialect variation may occur within the same broad linguistic group (e.g., native speakers of Spanish) for both L1 and L2 because of the strong influence of such factors as language contact and the diversity of individuals' places of origin.

In current research and practice in testing, the use of standard American English is often invoked as a strategy for minimizing the effect of dialect in the student's comprehension of items. The underlying assumption is that the standard dialect acts as the common dialect that everybody understands and uses. However, current thinking in the field of sociolinguistics holds that "standard" is not a common dialect but the dialect used by the privileged segment of a society (Halliday, 1978; Wardhaugh, 2002; Wolfram et al., 1999). Because the dialect used in the enacted curriculum may not be the same across schools (see Cummins, 2000; Moschkovich, 2000), using a "standard" form of English may unintentionally privilege some students and penalize others, even if they are native English speakers. Thus, the dialect used in national tests may differ considerably from the dialects used in certain classrooms and communities (Solano-Flores, 2006).

Testing ELLs in L1 raises a similar set of concerns. Even if qualified professionals translate tests into the students' native language, the translations may privilege the "standard" dialect of that language. Also, tight translation timelines prevent test developers from trying out translated items with ELLs and refining the wording of those items on the basis of the observed student interpretations (see Solano-Flores, in press). Moreover, subtle but important differences in word usage and word frequency and in the use of certain idiomatic expressions may limit the capacity of standard dialect versions of a test to properly assess ELL students in L1. This is especially the case for ELLs who are still developing their L1 (Sandoval & Durán, 1998).

In sum, whether ELLs are tested in L1 or L2, the validity of the measures of academic achievement may be affected if dialect variation across communities and the particular ways in which language is used in the enacted curriculum are not properly addressed.

By Whom

The effectiveness of ELL testing practices can be limited seriously by the implementation capacity of assessment systems (e.g., the skills of the individuals who develop, adapt, and administer tests). Flawed generalizations about the academic achievement of ELLs can result from failing to consider this capacity. For example, as part of their efforts to reduce the effect of limited English proficiency on test performance, states have used alternative

assessments such as portfolios of classroom work to assess ELLs. However, not all the states have the expertise that is required to properly develop alternate assessments and assessments in L1 (see General Accounting Office, 2006).

Unfortunately, seldom is this capacity questioned or examined. Rather, testing policies appear to be based on the assumption that all assessment systems are equally capable of properly implementing ELL testing procedures. For example, one of the accommodations permitted by the National Assessment of Educational Progress in the 2007 mathematics assessment is that the student has "test materials read aloud in native language." Differences in the qualifications of the individuals who read test materials aloud to students and differences in their degree of familiarity with the target population may produce important differences in the ways in which such accommodation is implemented.

The qualifications of the individuals who participate in the process of test development or test adaptation for ELLs are often misunderstood. For example, teachers of English as a foreign language and bilingual educators are likely to be mistaken as individuals with equivalent skills. Yet these two types of professionals have different kinds of training and views, and their work addresses the needs of very different kinds of learners (see Brown, 2007; Rabinowitz & Sato, 2006).

Also, literature on ELL testing rarely reports detailed information on the linguistic qualifications of the individuals who participate in testing projects (e.g., the characteristics of raters who score student responses to open-ended items or the criteria used to screen and select translators; see Solano-Flores & Li, 2006). Given the fact that bilingual individuals vary tremendously on their strengths and weaknesses in each language mode in L1 and L2, labels such as *bilingual* do not provide sufficient information on the linguistic skills of the professionals who participate in the process of testing ELLs.

The dearth of research on the qualifications of the individuals involved in the process of ELL testing suggests that the ability of assessment systems to communicate with ELLs needs to be examined carefully. There is much uncertainty about the characteristics of *by whom* in the process of ELL testing.

When

When, in their development of L2, ELLs should be tested and how many times they should be tested are issues that affect the validity of generalizations that can be made about their academic achievement on the basis of test scores. The former issue appears not to be properly addressed by legislation on the matter, and the second is not sufficiently investigated.

Testing policies for ELLs are not consistent with current knowledge on language development (see August & Hakuta, 1997). Including ELLs in high-stakes testing after a short period of schooling in L2 is not supported by evidence from research on language development. ELLs can develop basic conversational skills in a relatively short time after being immersed in an L2 environment, but they need considerably more time to develop the academic language in L2 that is necessary to learn at school (Cummins, 1981; Hakuta, 2001; Hakuta, Butler, & Witt, 2001). In addition to the vocabulary that is specific to a given content area, this academic language involves the set of oral and

written skills necessary to succeed in school (Scarcella, 2003); semantic and syntactic knowledge; and a critical set of functional language skills such as negotiating meaning, asking for clarification, confirming information, arguing, persuading, and expressing disagreement (Echevarría & Short, 2002).

An additional factor related to the appropriate time to test ELLs has to do with familiarity with testing. Immigrants from some countries may not be familiar with testing as it is practiced in the United States and may need some time to become familiar with the register of tests before they are able to understand those tests. This register includes, among other things, the format of multiple-choice items (e.g., a syntax in which sentences are cut off and have several complements) or certain discursive styles that are almost specific to tests (e.g., "Of the following, which one is . . .?"; Solano-Flores, 2006).

Although research has been conducted with monolingual students on the number of occasions needed to be able to make valid generalizations about the students' knowledge and skills, researchers have yet to investigate these questions for ELLs. Research with native English speakers shows that performance on tests is extremely unstable across testing occasions (Ruiz-Primo, Baxter, & Shavelson, 1993; Shavelson, Ruiz-Primo, & Wiley, 1999). Test scores vary considerably across occasions. The same students may obtain different scores on similar or even the same tasks or items administered on different occasions. Ideally, students should be tested on several occasions if we are to make appropriate generalizations about their knowledge and skills. Because of the complexity of dual-language development, it is quite possible that instability of performance across testing occasions is even a more serious issue for ELLs.

In sum, testing for ELLs takes place at a time in which they have not developed the academic language in L2 that they need to benefit from school. The kinds of generalizations that can be made on the basis of test scores may be further limited by performance instability across testing occasions.

Where

Research and practice in ELL testing have failed to examine the extent to which certain procedures can be effectively used in different contexts. This limitation may result from the fact that, in the field of testing, bilingualism is usually thought of solely as the condition of an individual who is able to use two languages.

Bilingualism can also be thought of as a phenomenon that involves communities that use two languages to communicate (Fishman, 1965; Mackey, 1962; Wei, 2000). This perspective makes it possible to appreciate that every school's sociolinguistic context is unique. What works in a given school context may not work in another school context because of subtle but important differences in the dialect and the register used in the enacted curriculum (Solano-Flores, 2006). Also, a given educational program for ELLs may be implemented in many different ways (Cummins, 1999). "One size fits all" approaches in ELL testing are often ineffective (see Guerrero, 1997).

The case of linguistic modification illustrates how strategies intended to support ELLs may vary tremendously in their effectiveness or in the ways in which they are implemented. Attempts to minimize the linguistic demands of test items can be based on

either simplifying their linguistic features (see Abedi et al., 2001) or ensuring that their linguistic features match the characteristics of the local (e.g., school, district) language usage (Solano-Flores, Speroni, & Sexton, 2005). These two procedures can be clearly defined and distinguished. Linguistic simplification is performed by committees of specialists and practitioners. In contrast, test localization is performed by teachers who teach in the specific communities and are assumed to be familiar with the dialects used in their schools. However, simply knowing the procedure used to modify a test item does not allow one to predict its characteristics (see Solano-Flores et al., 2007). Thus, if different teams of professionals were asked to modify the same item using the same procedure, each team would come up with a unique version of the item.

Stating that a given approach in the testing of ELL students is *the right* approach in many cases is likely to be an overgeneralization that ultimately affects the validity of their test scores. Improved approaches to ELL testing should be based on research that examines the effect of context on the effectiveness of testing strategies for ELLs.

Probabilistic Views of Language in the Process of ELL Testing

ELL Testing as a Probabilistic Process

In an ideal world, ELL students would be classified on the basis of accurate measures of proficiency in both L1 and L2 and in the four language modes, and the process of test development and adaptation would include appropriate samples of ELL populations. Also, testing accommodations would be customized to meet the set of linguistic strengths and weaknesses of each ELL student in each language mode in both L1 and L2. In addition, accommodations would be consistently and properly implemented; all individuals involved in the process of test development, translation, adaptation, or administration would have appropriate linguistic competencies; and ELLs would be tested in English only after they had developed a minimum competency in the academic language of L2. In that ideal world, the definitions of ELLs would be consistent across school districts and states, and the implementation of testing approaches and testing accommodations would be consistent with the characteristics of the individuals and their school contexts.

However, as the first part of this article shows, the effectiveness of the process of ELL testing is limited by such factors as tremendous linguistic variation among ELL populations, flawed definitions of ELLs, inaccurate or insufficient measures of language proficiency, and uncertainty about the fidelity with which testing accommodations are implemented.

When the behavior of the multiple components of a system is difficult to predict or control, the behavior of the entire system resembles a random process even if, in principle, it is intended to be a deterministic system (Szolovits & Pauker, 1978). In the field of education, in which complex phenomena interact and are difficult to isolate (Snow, 1968), effective research approaches can be devised that are based on recognizing and modeling uncertainty and, more specifically, on dealing statistically with measurement error (Lehrer, Serlin, & Amundson, 1990).

The limitations of deterministic models in testing and the advantages of probabilistic models were clearly discussed by Torgerson in 1958:

In the deterministic approach, the model itself is stated in terms of the ideal case, where all the variation in responses is accounted for by the variation in subjects and stimuli. No provision for unsystematic error of variance is made *in the model itself*. Since things are virtually never ideal, the practical problem in this approach consists not of determining whether or not the model fits any particular set of data—it almost never will—but rather whether the model can serve as an adequate approximation to the data. In contrast, the probabilistic models have built into them provision of one sort or another for a certain amount of unsystematic variance. (p. 300)

Torgerson's reasoning provides a good basis for thinking about the characteristics of effective probabilistic approaches to language in ELL testing. Such approaches should not depend excessively on (questionable) categories of language proficiency; they should allow examination of the amount of measurement error due to language factors—including the capacity of an assessment system to effectively communicate with ELLs—and they should provide a basis for examining how context shapes the effectiveness of ELL testing practices.

Theories in the field of psychometrics vary on their capabilities to produce approaches with the characteristics mentioned. Classical test theory treats measurement error as the discrepancy between examinees' observed test scores and their true scores (see Crocker & Algina, 1986). However, it treats error as undifferentiated and cannot disentangle language from other sources of error.

Item response theory (IRT; e.g., Hambleton, 1989; Lord & Novick, 1968; Tatsuoka, 1985)—a theory of scaling—examines the probability of an individual's responses to items as a function of the characteristics of the items and the characteristics of the person (see Marshall, 1990; van der Linden & Hambleton, 1997). Thus the theory does not allow detailed examination of assessment system–related language factors. Although IRT analyses allow for examination of language bias and differences between linguistic groups (e.g., van de Vijver & Poortinga, 1997), they are based on the assumption that these groups (e.g., "limited English proficient" and "English proficient"; Shepard, Taylor, & Betebenner, 1998) are homogeneous.

This leaves us with generalizability theory (G theory) as the only theory that allows detailed examination of multiple language–related sources of measurement error in ELL testing.

G Theory and Measurement Error in ELL Testing

Created by Cronbach and his associates as an extension of analysis of variance (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) and developed and disseminated by other authors (Brennan, 2001; Cardinet, Torneur, & Allal, 1976; Cronbach, Linn, Brennan, & Haertel, 1997; Shavelson & Webb, 1981, 1991), G theory allows examination of the extent to which test scores based on a sample of observations (e.g., a sample of test items scored by a sample of raters in a sample of testing occasions) can be generalized to a universe of all possible observations.

G theory distinguishes two types of sources of score variation: student, or the object of measurement, and facets (factors), or

sources of measurement error (see Shavelson & Webb, 1991) such as item, rater, and occasion. As Table 2 shows, student belongs to *who*, whereas the facets belong to the other five components of the process of ELL testing. (Needless to say, only some of the many possible facets are included in the table.) Although many of the facets listed in the table are specific to ELL testing, only item, rater, language, and dialect have been investigated in research using the perspective of G theory (see Solano-Flores & Li, 2006).

Table 3 shows the results from a G study—an analysis of the amount of score variation due to multiple sources of score variation—from a first study in an investigation with native Haitian-Creole speakers. In this first study, fourth- and fifth-grade students were given the same set of mathematics items in English and in L1 (Solano-Flores & Li, 2006). The table shows the estimated variance components and the percentage of the score variation produced by the main and interaction effect of student, item, rater, and language.³ The ϵ in the last term (sirl, ϵ) indicates error that cannot be accounted for and that is confounded with the score variation due to the interaction of the four sources of variation included in the study). The column on the left indicates the components of the sentence represented by the sources of score variation. In this G study, no *when* or *where* facets were included.⁴

The largest score variation observed is that due to the interaction of student, item, and language—which indicates that, in addition to the content and cognitive demands that are intrinsic to each item, each ELL has a unique set of strengths and weaknesses in each language and each item has a unique set of linguistic demands in each language.

G theory allows computation of ρ^2 and ϕ , coefficients that express the extent to which measures of academic achievement can be generalized, depending respectively on whether they are intended to rank students or to index the absolute level of knowledge of a student of a given domain (Shavelson & Webb, 1991). ELLs are given the same set of items in two languages, not necessarily with the intent to test them in two languages but rather with the intent to determine the language of testing with which higher ρ^2 and ϕ coefficients can be obtained. In the investigation with native Haitian-Creole speakers, higher coefficients were observed when students were tested in English than in L1. But this finding cannot be generalized beyond the specific community in which the investigation was conducted. As discussed next, there are limits to the generalizations that can be made across groups of ELLs.

Measurement Error and the Capacity of Assessment Systems to Communicate With ELLs

Because of its short history, the use of G theory in ELL testing has focused on examining score variation due to item and rater (see Solano-Flores & Li, 2006). G theory, however, can also be used to examine the effect of *tests* and *when* facets and the effect of *by whom* facets other than rater.

G theory distinguishes two types of facets, random and fixed. A facet is random when only some of its conditions are included in a G study (see Shavelson & Webb, 1991). Item is a random facet because the items in a test are assumed to be samples of a large knowledge domain (Hively, Patterson, & Page, 1968). Rater and

Table 2
Object of Measurement and Some Facets in the Process of English Language Learner (ELL) Testing According to the Six Components of Who Is Given Tests in What Language by Whom, When, and Where?

Mapping Sentence Component	Source of Score Variation	Specific to ELL Testing?	Categories
<i>Who</i>	Object of measurement: student	no	
	Facets:		
is given tests	test administration language mode	yes	printed form, test administrator reads items aloud
	testing time limit	no	with and without test completion time limit
	linguistic profile of test administrator	yes	different linguistic backgrounds and formal training in ELL's L2
	item	no	different items
	method of assessment	no	multiple-choice, open-ended, hands-on, etc.
	test translation model	yes	simple translation, parallel translation
	response format language mode	yes	student responds orally, student responds in writing
in what language	linguistic simplification	no	with and without linguistic simplification
	localization	no	with and without localization
	language of testing	yes	L1, L2
	dialect of the language used in testing	no	local dialect, standard dialect, nonlocal and nonstandard dialect
by whom	rater	no	different raters
	linguistic profile of raters	yes	different linguistic backgrounds and formal training in ELL's L2
when and where	linguistic profile of test developers	yes	different linguistic backgrounds and formal training in ELL's L2
	linguistic profile of test translators	yes	professional translators, teachers from the community
	occasion	no	different testing occasions
	school	no	different schools or school districts
	locale	no	rural, urban, suburban
	geographical area	no	different areas in the country

Note. L1 = first language; L2 = second language.

Table 3
Score Variation Across Languages: Estimated Variance Components in an $s \times i \times l \times r$ Random Model

Mapping Sentence Component	Source of Variability	<i>n</i>	Estimated Variance Components	%
<i>Who</i>	student (s)	49	.0299	20
is given tests	item (i)	10	.0097	6
in what language	language (l)	2	.0074	5
by whom	rater (r)	4	0 ^a	0
when and where	si		.0164	11
	sr		0 ^a	0
	sl		.0100	7
	ir		.0004	0
	il		.0016	1
	rl		.0001	0
	sir		0 ^a	0
	sil		.0589	39
	srl		0 ^a	0
	irl		0 ^a	0
	sirl,e		.0166	11

^aSmall negative variance components set to zero, following Brennan's (1992b) approach.

Source. Solano-Flores and Li (2006).

occasion are also random facets because any raters included in the scoring of students' responses or any occasion in which students are tested can be respectively assumed to be samples of all possible raters and all possible occasions.

A facet is fixed when all of its conditions are included in a G study or when generalizations are not made beyond the conditions being examined (Shavelson & Webb, 1991). The facet *testing time limit* (see Table 2) is fixed because there are only two conditions of interest: with time limit and without time limit for completing the test. Also, test administration language mode is fixed because it has only two categories: having the student read the test items and having the student hear the test items read by the test administrator.

Fixed facets are particularly relevant to examining conditions of measurement associated with context effects (e.g., the sequence with which items are administered in a test, the use of adaptive testing, or the use of testing conditions that move away from standardized measurement procedures in which the same form of a test is given to all students; see Brennan, 1992a). In ELL testing, many fixed facets originate from the characteristics of assessment systems and the actions taken with the intent to ensure the validity of academic achievement measures for ELLs. For example, although translated tests are intended to address ELLs' limited proficiency in English, test translation itself introduces measurement error. Another source of measurement error is translation model (defined, among other things, by whether the translation is made

by individual translators or by translation committees and by the number of translation review iterations intended to ensure that meaning is preserved across languages). A high percentage of score variation due to the main effect of the translation model may indicate a considerable difference of mean scores obtained by students when they are tested with tests translated using different test translation models. However, a high interaction effect of translation model with item and student may reveal that the effectiveness of one or another model of test translation is relative, as it is shaped by the characteristics of the items and the students.

Measurement Error and Contextual Factors That Limit Generalizations About the Effectiveness of ELL Testing Practices

Many results from research in the testing of ELLs should not be generalized beyond certain groups and contexts. Brennan (1992a) distinguishes between the universe of allowable observations and the universe of generalizations in testing:

The universe of generalization is the validity-defining universe, and systematic errors such as context effects arise in part because a standardized measurement procedure reflects a universe that is systematically different in some way or ways from the universe of generalization. For example, when all forms of a test use the same standardized conditions of measurement, the universe of allowable observations can be viewed as part of a universe of generalization—a part in which the standardized conditions are only one permissible set of conditions. If an investigator is interested in generalizing over a wider set of conditions than the standardized conditions, then inferences based solely on scores for standardized instruments are likely to involve context effects and/or other types of systematic errors. (p. 237)

A second study from the investigation with native Haitian-Creole speakers illustrates how G theory allows examination of the universe of generalization across contexts. Students from two communities, A and B, were given the same set of items in both the standard dialect version and the local dialect version of Haitian Creole (Solano-Flores & Li, 2006).⁵ On the basis of the estimated variance components obtained from the test scores in each dialect, we performed a series of decision studies—studies that make use of the information provided by a G study to design a testing model that minimizes error for a particular purpose (see Shavelson & Webb, 1991). This allowed us to determine the minimum number of items needed to obtain dependable scores by testing the students in each dialect.

Results showed that more than 15 items in the standard dialect and a little more than 10 items in the local dialect would be needed to obtain ρ^2 and ϕ coefficients of at least .80 for students in Community A. In contrast, about 20 items administered in either the local or the standard Haitian-Creole dialect would be needed to obtain ρ^2 and ϕ coefficients of at least .80 for students in Community B.

These results show how school communities vary in their sensitivity to approaches used to address language in the testing of ELLs. Reliability estimates are not necessarily the same for different groups (Li & Brennan, 2007; Solano-Flores & Li, 2008). What works for one school district may not work equally well for

another, even if the populations of ELLs tested are native speakers of the same L1 and are assumed to have similar levels of English proficiency, and the programs in which students are enrolled are deemed equivalent. Identifying the limits of the generalization of testing approaches is another form of incorporating probabilistic views of language in ELL testing.

Concluding Remarks

In this article, I have discussed the need for probabilistic views of language in the testing of ELL students. *Who is given tests in what language by whom, when, and where?* has been used as a conceptual framework and as a mapping sentence for examining the process of ELL testing from a perspective that takes into consideration the notion that bilingual behavior is shaped by multiple contextual factors. According to this approach, ELL testing involves a communication process (e.g., test developers write items, students respond to them, and raters interpret the students' responses). Valid testing for ELLs cannot be achieved if we focus solely on the proficiency of ELLs in English but fail to examine linguistic factors involved in the development, adaptation, administration, and scoring of tests.

A review of the six components of the process of ELL testing reveals that, although it can be operationalized on the basis of a system of categories of English proficiency and a set of testing procedures, in practice, the process of ELL testing behaves, to a great extent, randomly because of factors that are beyond control, cannot be measured accurately or sufficiently, have a high level of uncertainty, are instable or inconsistent, or involve strategies that are ineffective in addressing language proficiency.

Although ELL testing practice and research have used the probabilistic reasoning inherent to any psychometric theory, they are greatly influenced by deterministic views of language and linguistic groups. Enhanced approaches can and should address randomness in the process of ELL testing by accounting for measurement error due to both language factors and factors associated with the characteristics of an assessment system and the implementation of actions intended to address the needs of ELL students.

I have discussed evidence from recent research indicating that G theory allows for examination of the amount of measurement error due to the main and interaction effect of multiple facets (sources of measurement error). Because the use of G theory in the testing of linguistically diverse populations has a short history, evidence of its capability to address randomness in ELL testing is limited, for now, to only some of the many facets of the six components of the process.

Using probabilistic views of language in the process of ELL testing helps us to see the link between assessment system effectiveness and score dependability. By building into our analyses the provision of unsystematic variance due to the limited ability of our assessment systems to communicate with ELLs, we can do a better job of testing these students. Valid measures of academic achievement for ELLs cannot be obtained without seriously revisiting our testing practices.

NOTES

The research reported here was funded by National Science Foundation Grants REC-9909729, REC-0126344, REC-0336744, and SGER-0450090. I wish to thank Janette Klingner, Min Li, Nicole Sager, Rachel

Prosser, Chao Wang, and two anonymous reviewers for their comments on this article. I also want to thank Barry Sloane, Larry Suter, Elizabeth VanderPutten, and Rich Shavelson for their interest in this research. The opinions expressed here are not necessarily those of the funding agency or my colleagues.

¹Although "Who Gives Tests . . ." would be syntactically more consistent with Fishman's article's title, "Who Speaks . . .," I use the passive voice form, "Who Is Given Tests . . .," in which the grammatical object appears at the beginning of the sentence. The reasons for this will become clear later.

²*Bilingual* describes the condition of being capable of using two languages. The term refers to a wide range of degrees of proficiency in each language mode in each language. English language learners (ELLs) are regarded as bilinguals who have at least an incipient proficiency in English (see Bialystok, 2001; Valdés & Figueroa, 1994).

³In G theory studies, the score variation due to student—the object of measurement—is reported first, before the sources of measurement error. This is the reason that the title of this article is "Who Is Given Tests . . ." in which the object of measurement (and also the grammatical subject) appears at the beginning of the sentence.

⁴The design of this particular G study is a *crossed design*—it includes all the combinations of conditions of the sources of variability (see Shavelson & Webb, 1991). The design of a study involving *where* facets needs to be a *nested design*. For example, student is nested within school because a given student can belong to only one school. G study designs in which students are nested within a facet are rare and used in combination with other crossed-design studies (e.g., Lane, Liu, Ankenmann, & Stone, 1996). In a crossed-design G study with students nested within schools, the sequence and combinations of sources of score variation would not be the same as those in Table 2. As a component of the process of ELL testing, *where* is more likely to be investigated by an examination of *how* (because of different source variation) the patterns of score variability differ across schools, school districts, locales, or geographical areas, as in the case illustrated by Table 3.

⁵The students tested in two dialects are not the same as the students tested in two languages in the study described above.

REFERENCES

- Abedi, J., Hofstetter, C., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research, 74*(1), 1–28.
- Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*, 219–234.
- Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2001). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice, 19*(3), 16–26.
- Aguirre-Muñoz, Z., & Baker, E. L. (1997). *Improving the equity and validity of assessment-based information systems*. (CSE Tech. Rep. No. 462). Los Angeles: Center for the Study of Evaluation National Center for Research on Evaluation, Standards, and Student Testing; Graduate School of Education & Information Studies, University of California, Los Angeles.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- August, D., & Hakuta, K. (Eds.). (1997). *Improving schooling for language minority students: A research agenda*. Washington, DC: National Academy Press.
- Baker, C. (2006). *Foundations of bilingual education and bilingualism* (4th ed.). Clevedon, UK: Multilingual Matters.
- Bernardo, A. B. I. (2002). Language and mathematical problem solving among bilinguals. *Journal of Psychology, 136*, 283–297.
- Bialystok, E. (2001). *Bilingualism in development: Language, literacy, and cognition*. Cambridge, UK: Cambridge University Press.
- Bormuth, J. R. (1970). *On the theory of achievement test items*. Chicago: University of Chicago Press.
- Brennan, R. L. (1992a). The context of context effects. *Applied Measurement in Education, 5*, 225–264.
- Brennan, R. L. (1992b). *Elements of generalizability theory*. Iowa City, IA: American College Testing Program.
- Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.
- Brisk, M. E. (2006). *Bilingual education: From compensatory to quality schooling* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Brown, H. D. (2007). *Principles of language learning and teaching* (5th ed.). Boston: Addison Wesley & Longman.
- Butler, F. A., & Stevens, R. (1997). Accommodation strategies for English language learners on large-scale assessments: Student characteristics and other considerations (CSE Tech. Rep. No. 448). Los Angeles: Center for the Study of Evaluation; National Center for Research on Evaluation, Standards, and Student Testing; Graduate School of Education & Information Studies, University of California, Los Angeles.
- Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement, 13*, 119–135.
- Chamot, A. U., & O'Malley, J. M. (1994). *The CALLA handbook: Implementing the cognitive academic language learning approach*. Reading, MA: Addison-Wesley.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Holt, Rinehart & Winston.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: John Wiley.
- Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57*, 373–399.
- Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In *Schooling and language minority students: A theoretical framework* (pp. 3–49). Sacramento: California State Department of Education Office of Bilingual Education.
- Cummins, J. (1999, March). *Research, ethics, and public discourse: The debate on bilingual education*. Paper presented at the National Conference of the American Association of Higher Education, Washington, DC.
- Cummins, J. (2000). *Language, power, and pedagogy: Bilingual children in the crossfire*. Clevedon, UK: Multilingual Matters.
- Cunningham, J. W., & Fitzgerald, J. (1996). Epistemology and reading. *Reading Research Quarterly, 31*, 36–60.
- De Corte, E., Verschaffel, L., & Pauwels, A. (1990). Influence of the semantic structure of word problems on second graders' eye movements. *Journal of Educational Psychology, 82*, 359–365.
- Echevarría, J., & Short, D. J. (2002). Using multiple perspectives in observations of diverse classrooms: The Sheltered Instruction Observation Protocol (SIOP). Retrieved from the Center for Research on Education, Diversity, and Excellence website, October 23, 2006, <http://crede.berkeley.edu/tools/policy/siop/1.3doc2.shtml>
- Ercikan, K. (1998). Translation effects in international assessment. *International Journal of Educational Research, 29*, 543–553.
- Ercikan, K. (2002). Disentangling sources of differential item functioning in multi-language assessments. *International Journal of Testing, 2*, 199–215.
- Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2005). Comparability of bilingual versions of assessments: Sources

- of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education*, 17, 301–321.
- Farr, M., & Ball, A. F. (1999). Standard English. In B. Spolsky (Ed.), *Concise encyclopedia of educational linguistics* (pp. 753–757). Kidlington, UK: Elsevier Science.
- Ferguson, A. M., & Fairburn, J. (1985). Language experience for problem solving in mathematics. *Reading Teacher*, 38, 504–507.
- Ferrara, S., Macmillan, J., & Nathan, A. (2004, January). *Enhanced database on inclusion and accommodations: Variables and measures* (NAEP State Analysis Project report to the National Center for Education Statistics). Washington, DC: NCES.
- Fishman, J. A. (1965). Who speaks what language to whom and when? *La Linguistique*, 2, 67–88.
- García-Duncan, T., del Río-Parent, L., Chen, W.-H., Ferrara, S., Johnson, E., Oppler, S., et al. (2005). Study of a dual-language test booklet in eight-grade mathematics. *Applied Measurement in Education*, 18, 129–161.
- Genesee, F. (Ed.). (1994). Introduction. In F. Genesee (Ed.), *Educating second language children: The whole child, the whole curriculum, the whole community* (pp. 1–11). Cambridge, UK: Cambridge University Press.
- General Accounting Office. (2006, July). No Child Left Behind Act: Assistance from Education could help states better measure progress of students with limited English proficiency (Report to Congressional Requesters, No. GAO-06-815). Washington, DC: Author.
- George Washington University Center for Equity and Excellence in Education. (2005). *Accommodations*. Retrieved July 9, 2007, from <http://ceec.gwu.edu/AA/Accommodations.html>
- Guerrero, M. D. (1997). Spanish academic language proficiency: The case of bilingual education teachers in the United States. *Bilingual Research Journal*, 21, 25–43.
- Guttman, L. (1969). Integration of test design and analysis. In *Proceedings of the 1969 invitational conference on testing problems* (pp. 53–65). Princeton, NJ: Educational Testing Service.
- Hakuta, K. (2001, April). *Key policy milestones and directions in the education of English language learners*. Paper presented at the Rockefeller Foundation Symposium, Leveraging Change: An Emerging Framework for Educational Equity, Washington, DC.
- Hakuta, K., Butler, Y. G., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* (Policy Rep. No. 2000-1). Santa Barbara: University of California Linguistic Minority Research Institute.
- Halliday, M. A. K. (1978). Language as social semiotic: The social interpretation of language and meaning. London: Edward Arnold.
- Hamblen, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147–200). New York: American Council on Education/Macmillan.
- Harry, B., & Klingner, J. (2006). *Why are so many minority students in special education?* New York: Teachers College Press.
- Heath, S. B. (1983). *Ways with words: Language, life and work in communities and classrooms*. Cambridge, UK: Cambridge University Press.
- Hively, W., Patterson, H. L., & Page, S. H. (1968). A “universe-defined” system of arithmetic achievement tests. *Journal of Educational Measurement*, 5, 275–290.
- Kopriva, R. (2001, June). *ELL validity research designs for state academic assessments: An outline of five research designs evaluating the validity of large-scale assessments for English language learners and other test takers*. Paper prepared at the Council of Chief State School Officers Meeting, Houston, TX.
- Kopriva, R. J., Emick, J. E., Hipolito-Delgado, C. P., & Cameron, C. A. (2007). Do proper accommodation assignments make a difference? Examining the impact of improved decision making on scores for English language learners. *Educational Measurement: Issues and Practice*, 26(3), 11–20.
- Kopriva, R., Wiley, D. E., Chen, C., Levy, R., Winter, P. C., & Corliss, T. (2004). Field test validity study results: English language development assessment: Final report submitted to the Council of Chief State School Officers LEP-SCASS. College Park: Center for the Study of Assessment Validity and Evaluation, University of Maryland.
- Lane, S., Liu, M., Ankenmann, R. D., & Stone, C. A. (1996). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement*, 33, 71–92.
- Lec, O., Deaktor, R. A., Hart, J. E., Cuevas, P., & Enders, C. (2005). An instructional intervention's impact on the science and literacy achievement of culturally and linguistically diverse elementary students. *Journal of Research in Science Teaching*, 42, 857–887.
- Lehrer, R., Serlin, R. C., & Amundson, R. (1990). Knowledge or certainty: A reply to Cziko. *Educational Researcher*, 19(6), 16–19.
- Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*, 26(2), 3–16.
- Li, D., & Brennan, R. (2007, April). *A multi-group generalizability analysis of a large-scale reading comprehension test*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Mackey, W. F. (1962). The description of bilingualism. *Canadian Journal of Linguistics*, 7, 51–85.
- MacSwan, J. (2000). The threshold hypothesis, semilingualism, and other contributions to a deficit view of linguistic minorities. *Hispanic Journal of Behavioral Sciences*, 22(1), 3–45.
- Marshall, S. P. (1990). Generating good items for diagnostic tests. In N. Frederiksen, R. Glaser, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 433–452). Hillsdale, NJ: Lawrence Erlbaum.
- Moschkovich, J. N. (2000). Learning mathematics in two languages: Moving from obstacles to resources. In W. Secada (Ed.), *Changing faces of mathematics: Vol. 1. Perspectives on multiculturalism and gender equity* (pp. 85–94). Reston, VA: National Council of Teachers of Mathematics.
- National Assessment of Educational Progress. (2005). *U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2005 Reading, Mathematics, and Science Assessments*. Retrieved September 29, 2006, from <http://nces.ed.gov/nationsreportcard/about/inclusion.asp>
- National Assessment of Educational Progress. (2007). *NAEP 2007 Mathematics Assessment*. Retrieved July 14, 2007, from http://nces.ed.gov/nationsreportcard/about/inclusion.asp#accom_table.
- No Child Left Behind Act, 10 U.S.C. 6301 (2002).
- Noonan, J. (1990). Readability problems presented by mathematics text. *Early Child Development and Care*, 54, 57–81.
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- Rabinowitz, S., & Sato, E. (2006, April). *Technical adequacy of assessments from alternate student populations: Technical review of high-stakes assessment for English language learners*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- Rivera, C., Collum, E., Willner, L. N., & Sia, Jr. J. K. (2006). Study 1: An analysis of state assessment policies regarding the accommodation of English language learners. In C. Rivera & E. Collum (Eds.), *State assessment policy and practice for English language learners: A national perspective* (pp. 1–136). Mahwah, NJ: Lawrence Erlbaum.

- Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement*, 30, 41–53.
- Sandoval, J., & Durán, R.P. (1998). Language. In J. Sandoval, C. Frisby, K. F. Geisinger, J. D. Scheuneman, & J. R. Grenier (Eds.), *Test interpretation and diversity: Achieving equity in assessment* (pp. 181–211). Washington, DC: American Psychological Association.
- Scarcella, R. C. (2003). *Academic English: A conceptual framework* (Rep. No. 2003-1). Santa Barbara: University of California Linguistic Minority Research Institute.
- Shavelson, R. J., Ruiz-Primo, M. A. & Wiley, E. W. (1999). Note on sources of sampling variability in science performance assessments. *Journal of Educational Measurement*, 36, 61–71.
- Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973–1980. *British Journal of Mathematical and Statistical Psychology*, 34, 133–166.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.
- Shepard, L., Taylor, G., & Betebenner, D. (1998). *Inclusion of limited-English proficient students in Rhode Island's Grade 4 mathematics performance assessment* (CSE Tech. Rep. No. 486). Los Angeles: Center for the Study of Evaluation; National Center for Research on Evaluation, Standards, and Student Testing; Graduate School of Education & Information Studies, University of California, Los Angeles.
- Shin, F. H. (2004). English language development standards and benchmarks: Policy issues and a call for more focused research. *Bilingual Research Journal*, 28, 253–266.
- Shorrocks-Taylor, D., & Hargreaves, M. (1999). Making it clear: A review of language issues in testing with special reference to the mathematics tests at key stage 2. *Educational Research*, 41, 123–136.
- Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing*, 20, 148–166.
- Sireci, S. G., & Khaliq, S. N. (2002). *Comparing the psychometric properties of monolingual and dual language test forms* (Center for Educational Assessment research report). Amherst: School of Education, University of Massachusetts Amherst.
- Snow, R. E. (1968). Brunswikian approaches to research on teaching. *American Educational Research Journal*, 5, 475–489.
- Solano-Flores, G. (2006). Language, dialect, and register: Sociolinguistics and the estimation of measurement error in the testing of English-language learners. *Teachers College Record*, 108, 2354–2379.
- Solano-Flores, G. (in press). Successive test development. In C. R. Reynolds, R. W. Kamphaus, & C. DiStefano (Eds.), *Encyclopedia of psychological and educational testing: Clinical and psychoeducational applications*. New York: Oxford University Press.
- Solano-Flores, G., & Li, M. (2006). The use of generalizability (G) theory in the testing of linguistic minorities. *Educational Measurement: Issues and Practice*, 25(1), 13–22.
- Solano-Flores, G., & Li, M. (2008). *On the dependability of cognitive interview-based measures across cultural groups*. Manuscript submitted for publication.
- Solano-Flores, G., Li, M., Speroni, C., Rodriguez, J., Basterra, M., & Dovholuk, G. (2007, April). *Comparing the properties of teacher-adapted and linguistically-simplified test items for English language learners*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching*, 38(5), 553–573.
- Solano-Flores, G., Speroni, C., & Sexton, U. (2005, April). *The process of test translation: Advantages and challenges of a socio-linguistic approach*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.
- Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher*, 32(2), 3–13.
- Solano-Flores, G., & Trumbull, E. (2008). In what language should English language learners be tested? In R. J. Kopriva (Ed.), *Improving testing for English language learners: A comprehensive approach to designing, building, implementing and interpreting better academic assessments* (pp. 169–200). New York: Routledge.
- Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2002). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing*, 2, 107–129.
- Stansfield, C. W. (2003). Test translation and adaptation in public education in the USA. *Language Testing*, 20, 188–206.
- Stansfield, C. W., & Bowles, M. (2006). Study 2: Test translation and state assessment policies for English language learners. In C. Rivera & E. Collum (Eds.), *State assessment policy and practice for English language learners: A national perspective* (pp. 175–221). Mahwah, NJ: Lawrence Erlbaum.
- Stevens, R. A., Butler, F. A., & Castellon-Wellington, M. (2000). Academic language and content assessment: Measuring the progress of English language learners (ELLs) (CSE Tech. Rep. No. 552). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing; Graduate School of Education & Information Studies, University of California, Los Angeles.
- Szolovits, P., & Pauker, S. G. (1978). Categorical and probabilistic reasoning in medical diagnosis. *Artificial Intelligence*, 11, 115–144.
- Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics*, 10(1), 55–73.
- Thurber, R. S., Shinn, M. R., & Smolkowski, K. (2002). What is measured in mathematics tests? Construct validity of curriculum-based mathematics measures. *School Psychology Review*, 31, 498–513.
- Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: John Wiley.
- Valdés, G., & Figueroa, R. A. (1994). *Bilingualism and testing: A special case of bias*. Norwood, NJ: Ablex.
- van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.
- van de Vijver, F. J. R., & Poortinga, Y. H. (1997) (Eds.). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, 13, 29–37.
- Wardhaugh, R. (2002). *An introduction to sociolinguistics*, (4th ed.). Oxford, UK: Blackwell.
- Wei, L. (2000). Dimensions of bilingualism. In L. Wei (Ed.), *The bilingualism reader* (pp. 3–25). London: Routledge.
- Wellington, J., & Osborne, J. (2001). *Language and literacy in science education*. Buckingham, UK: Open University Press.
- Wolfram, W., Adger, C. T., & Christian, D. (1999). *Dialects in schools and communities*. Mahwah, NJ: Lawrence Erlbaum.

AUTHOR

GUILLERMO SOLANO-FLORES is an associate professor of bilingual education and English as a second language at the University of Colorado, Boulder, School of Education, 249 UCB, Boulder, CO 80309; guillermo.solano@colorado.edu. His research focuses on educational measurement, assessment development, and the linguistic and cultural issues that are relevant to both the testing of linguistic minorities and international test comparisons.

Manuscript received February 5, 2008

Accepted April 3, 2008

From: David Stevenson [dstevenson@wgen.net]
Sent: Wednesday, December 02, 2009 4:30 PM
To: Race To The Top Assessment Input
Cc: Larry Berger
Subject: Race to the Top Assessment Program

Submitter: Wireless Generation
Title: Feedback on RTTT Assessment RFP
Topic Addressed: General Assessment Input

We applaud USED's leadership on the crucial issue of large-scale assessment. We believe that the Common Standards conjoined with this opportunity for competition and collaboration will serve as the catalyst for a breakthrough alignment of curriculum, instruction, professional development, and assessment. As professionals who have devoted our lives to supporting better teaching, we are particularly enthusiastic about the prominent place that formative assessment practices hold in USED's plan for the RTTT Assessment RFP.

Our primary feedback is to encourage USED to be minimally specific about particular features of an envisioned assessment system and instead to focus on goals that the system will help schools and teachers achieve. A bold move would be to dedicate more than half of the competitive priorities in the competition to the instructional use of assessment data and the teacher improvement use of assessment data. A good model would be the language of 3C in the Race To The Top Program and its focus on "instructional improvement systems," which are to:

provide teachers, principals, and administrators with meaningful support and actionable data to systemically manage continuous instructional improvement, including such activities as: instructional planning; gathering information (*e.g.*, through formative assessments (as defined in this notice), interim assessments (as defined in this notice), summative assessments, and looking at student work and other student data; analyzing information with the support of rapid-time (as defined in this notice) reporting; using this information to inform decisions on appropriate next instructional steps; and evaluating the effectiveness of the actions taken.

The more flexibility permitted to states and their partners in proposing a new system, the more likely that surprising innovation will occur.

Our other strong suggestion is that states should be allowed to participate in multiple consortia so that they can explore both more traditional and more innovative approaches, thereby creating a better competitive dynamic in the RFP process. Otherwise, states may be forced to choose according to which consortium seems likely to win rather than which consortium has the most compelling vision.

Again, we applaud USED's leadership on this issue. This is a thrilling time to be working in education.

Thank you for the opportunity to comment on the proposed priorities, requirements, definitions, and selection criteria for the Race to the Top (RTTT) Assessment program. Wisconsin is in support of this program, which would provide for approximately \$350 million in grants to consortia of States for the development of common, high-quality assessments aligned with K-12 standards that are internationally benchmarked and that build toward college and career readiness by the time of high school completion. We are excited about the prospect of working with other states and securing additional funds to help us with our efforts to redesign our state assessment system in Wisconsin. Below are some comments related to the RTTT Assessment guidance for your consideration:

1. This program needs to remain a separate fund, and should not be rolled back into the larger RTTT grant.
2. The grant should allow for the proposal of a more comprehensive system of assessments, rather than being specifically limited to summative testing.
3. The Department should be clear about the proposed purpose(s) of the assessments that will be developed. That is, if the tests will be used for accountability, teacher/principal evaluation, etc. Please make this clear as it affects the characteristics of the system that will be developed.
4. The Department should strike a balance between psychometric rigor and the encouragement of innovation. (For example, consortia may want to consider ways to combine performance assessment items conducted locally with other more traditional assessment items, and if there is too much emphasis on comparability and reliability it may not allow for various models to be proposed that might be viable. Others may need to run two assessments simultaneously – one computer-based and one paper/pencil, and if the computer-based version has a different test construct it is not going to be comparable to the paper/pencil.)

Thank you for the opportunity to comment. We look forward to the final guidance.

A Design for an American Examination and Testing System

Marc Tucker¹

President, National Center on Education and the Economy

December 2009

The Obama administration has asked for advice as to what sort of voluntary national testing system the United States should have.

But it is impossible to design a testing system unless we know how and for what purposes it will be used. This includes what stakes will be attached to the tests for both students and teachers, what credentials based on the tests will be used for, whether the system is expected to produce information that teachers can use in real time to adapt their teaching to the actual needs of the students in relation to the standards, whether the standards and tests are to be strongly linked to curriculum (which is generally the case in other countries), whether the testing system is to be based on a common educational experience for all students that ends at grade 12, or at grade 10 (which is also the case in many other countries) and much, much more.

So we can probably agree that the design of the testing system should be intimately linked to the design of the larger education system of which it will be a part. But every aspect of those decisions is now in flux.

So we have two choices. We can assume that the design of the education system remains in the future as it is now, changed only by the introduction of the new Common Core standards and a system of aligned tests. Or we can assume that the new standards and the new tests are to be used in a system very different in important respects from the one we have now. This paper makes the latter assumption.

The reason is simple. The United States is now the second most expensive elementary and secondary system in the industrialized world, on a cost per student basis, and produces results inferior to those of more than 20 other countries. To cement our system in place would be tantamount to accepting our status as the country with the least productive school system in the industrialized world.

The testing system design offered below is based on the 21 years of research the National Center on Education and the Economy has done in more than two dozen countries in other parts of the world with much higher performing education systems than ours.

¹ The author is indebted to Howard Everson, David R. Mandel, Jim Pellegrino, Betsy Brown Ruzzi and Susan Sclafani for their comments on this paper and their contributions to the work on which it is based. None of these people, however, are to blame for whatever shortcoming this paper might have.

Much of that research has focused on these countries' academic standards, occupational skills standards, testing and examination systems and instructional systems.

Our reading of the best comparative research on successful education systems is the same as that of the best researchers elsewhere: The two most important features to be found in all of the most successful systems are, 1) they recruit their teachers from the top third in ability of their college graduates, and 2) they include complete, coherent and powerful national instructional systems for their students, systems that have at their heart well designed curriculum aligned with very high quality examination systems. The examination systems are designed at the high school level to support qualifications, which certify that the holder is qualified to go on to work or to the next stage of his or her education. This paper does not deal with the first of these two factors, but the second lies at the heart of the proposals made here.

The key elements in the larger system offered in this paper are the following: 1) a new *gateway* between high school and college defined by a certificate attesting that the holder has the knowledge and skill needed to be successful in the initial credit-bearing courses in our open-admission 2-year and 4-year colleges; 2) a new *system of instruction* through all the grades powerful enough to get virtually all our students to the new certificate standards before they leave high school; and 3) an accountability system for school faculty and students designed in such a way that *teachers have strong incentives to provide effective instruction* and *students have strong incentives to take tough courses and work hard in school*.

The testing system offered here is meant to support this larger design

•

The plan includes examining whether students have met internationally benchmarked standards for student accomplishment that are the same throughout the United States while at the same time recognizing American opposition to federalizing elementary and secondary education. Some of the elements of this plan are national, but very few are federal. Responsibilities for key aspects of the system are not concentrated at the national level, but are distributed up and down the system. And we have tried to conceive a system which offers choices rather than mandates wherever possible, especially when it comes to curriculum, while still insisting on common performance standards.

•

This paper is divided into sections. In the first, we describe at some length the criteria we think we should keep in mind in designing a national testing system. We've done this because many of the criteria we think are most important are not typically taken into account in designing American testing systems, but are often carefully considered in the design of assessment systems in those countries with education systems more successful than ours. A few are unique to this design, and are offered by way of acknowledging

some requirements that are derived from uniquely American values. All are central to the design.

In the second section, the design itself is described in detail. It would make a major break with the grade-by-grade requirements of No Child Left Behind. We offer a rationale for that approach in that sub-section.

The third section consists of some comments on the role of technology in the design.

In the fourth section, we offer the outlines of a plan for getting from where the nation is to the implementation of the kind of system we will have proposed in the earlier sections.

The kind of accountability systems that could be developed to take advantage of the standards, curriculum and assessment system described earlier are described in the fourth section.

In the last section, we sum up the advantages of the system design described in this paper.

One small note on definitions is important here. Throughout, we use the word “test” when referring to an assessment that is not based on a particular curriculum, and the word “examination” when the purpose is to assess the extent to which a student has mastered a particular curriculum. Wherever the words “test” and “examination” are used, you can assume that we are describing assessment for high stakes purposes unless the text explicitly states otherwise. We will generally use the phrase “formative assessment” when low or no stakes are attached to the assessment and the purpose is to provide information to teachers and others that is used to change the course of instruction in the light of the data produced about student achievement in relation to the standards.

Criteria for Design of Testing and Examination System

First, we want to offer the following set of explicit criteria for the adequacy of the design. This set of criteria assumes that the college-ready standards developed by the Common Core working group will serve as the basis of the work on test development going forward.

1. We should be aiming for a system in which all students complete the core program of studies on or about the time they are 16, or at the end of their sophomore year of high school. The standard students are expected to meet at the end of their sophomore year should be the level of literacy needed to succeed in the first credit-bearing courses in 2-year and 4-year open admissions postsecondary institutions

Most of the countries with the best performance in elementary and secondary education have defined what they expect of an educated person in their society and have

incorporated those expectations in a program of studies that they expect their young people to complete by the age of 16. After that, there are many pathways that students may take, depending on their demonstrated ability and interests. If we try to define a common program that all students are expected to complete by the age of 18, we will fail. The reason is well demonstrated by the problem that Achieve has had with its Algebra II program. The industries that depend on people with high competence in the STEM subjects will expect young people to have mastered Algebra II by the time they leave high school, because it is very important to them that these students be ready for calculus when they get to college. But probably fewer than five percent of working adults need the calculus in their work. If the common requirement is met by the age of 16, then those students who need Algebra II, and possibly calculus as well, can take it in their junior and senior year in high school. But if we require all students to take Algebra II in high school, very large numbers will fail, and will be denied a diploma because they were unable to master a subject they will never need as adults. Thus the testing system that becomes the basis of the accountability system of the states needs to be built on the assumption that the curriculum offered students can and will be delivered and its results can be assessed by the end of the sophomore year in high school.

Our research on the requirements of post-secondary institutions strongly suggests that it is entirely possible for high school students to complete a program of studies by the end of their sophomore year that will result in those students acquiring a level of literacy in English and mathematics sufficiently high to be successful in the first credit-bearing courses in our nation's public 2-year and 4-year postsecondary institutions. For example, the first credit-bearing course in mathematics at most community and technical colleges includes topics that would place it somewhere between Algebra I and Algebra II. It follows that students who have mastered Algebra I should be able to succeed in those college courses.

That being so, it would make sense to tell all our students that, when they have demonstrated that they have mastered the necessary skills and acquired the necessary knowledge, they need not hang out in high school, but should be able to go directly to the public 2-year or 4-year open-admissions institution of their choice the following fall, without having to take any remedial courses. They would then be able to pursue one of a number of defined pathways, depending only on their wishes. They could go into a technical program leading to 2-year or 3-year degree or certificate qualifying them for a career requiring such a degree (anything from nurses aide to dental technical to specialty welder to software systems manager). They could enter a 2-year college transfer program with the intention of going on to a four-year college. They could stay in high school to pursue a program of studies designed to prepare them for entry into a selective college. Or they could go to work, secure in the knowledge that they could go to college later, able to do college level work.

2. The standards and testing system should be designed to support a qualifications system

In much of the rest of the world, high school students work for qualifications. In the United States, they put in time. A qualification is a piece of paper conferred by the authorities that declares that the bearer of the qualification knows what he she needs to know and has the skills needed to do something in particular, for example, to go on to the next stage of one's education or to the next stage of one's career. A person who has passed her bar exam has a qualification. Qualifications systems are indifferent to when or in what institution the qualification was earned. In countries with qualifications systems, when we ask students where they are in their education, they tell us what qualification they are studying for. In the United States, they tell us what institution they are in: elementary school, middle school, high school, 2-year college, or 4-year college. Within very broad limits, it does not matter how well one does in any one of these institutions. One is expected to put in about the same amount of time as one's peers, and, when the time has come, one moves on. The result in the United States is that a great many high school students go on to college who are not qualified to do college level work, and it should not surprise us that they fail to complete. In our system, time is constant and the standard varies. In a qualifications system, the standard is constant and the time to reach it varies. The key to a successful qualifications system is that the standard is well known and widely accepted. The high school diploma is not a qualification, because the standard to earn it, in most states, is either not known or is so low as to be meaningless. The premise behind the principle stated in #2 above is that students would have to earn a qualification to go to college, and the standard required to earn that qualification would be a demonstration that the holder has the knowledge and skill needed to do college-level work. High schools would be held accountable for getting all students ready for 2-year and 4-year college work, whether or not they choose to go to college.

One might conclude that the principles behind the idea of a qualification are at odds with the American idea that it should never be too late to buckle down and succeed, the idea of a system in which one always gets a "second chance" and maybe many second chances to succeed. Actually, these two ideas are quite compatible. If a person has only one opportunity to take an exam and one shot at the learning that would enable one to succeed on that exam that is the gateway to the qualification, then the system becomes a sorting system. But what is presented here is the opposite of a sorting system. When it is never too late to retake the exam and earn the qualification and, and when the state has an obligation to those who do not succeed to give the learner another shot at learning the material, it becomes another, much more effective form of second chance system.

3. The tests or examinations at the heart of the assessment system should be standards-based.

The American style of testing was devised to sort students out along a distribution of ability or achievement. In this style of testing all students are compared to a norm. Because this is true, the ideal item in a test is one that, when field-tested, produces a normal curve of responses from the student responding to that item. Mathematically, it must be true that half of the students taking such a test will fall below the norm, and, in that sense, fail the test. If more than half succeed, the test will have to be renormed. In a

standards-based (or criterion-referenced) system of testing, the test-maker works against a specified standard of accomplishment and devises a test designed to report the degree to which students have achieved that standard. In such a system, it is theoretically possible for all students to achieve the standard. In a norm-referenced system, an item that all students could pass is deemed to be faulty by the test maker and is thrown out. In a standards based system, such an item should be included if it can be shown to measure the desired performance. If, as is now the case, society's interest is in getting all students ready for college, and "ready for college" can be incorporated in a measurable standard, then society requires a standards based system of testing, not a norm referenced system of testing, and the conventions and procedures that the testing experts use to devise the tests will have to reflect that requirement. American psychometricians have made progress toward this goal in recent years, and many professionals in the testing industry are very much aware of this issue and have built some high quality criterion referenced tests, but, on the whole, there is room for a lot of progress on this point.

4. The tests or examinations at the heart of the system should be curriculum-based

In most of the rest of the world, the purpose of the examinations is to determine whether the student has mastered the curriculum that student has studied. By curriculum we mean the courses to be taken, the topics to be studied, the instructional approaches to be used, and, often, the particular works to be studied and work to be undertaken. The process begins by deciding what set of courses at the high school level constitute a core curriculum which, if mastered, represent what it means to be a well education person. Once the decision as to the broad shape of the curriculum is made, the designers produce syllabi for each course, spelling out what the goals of the course are, what the student is expected to learn, what the student will be examined on, what the major assignments will be, what the student is expected to read, and what the final grade will be based on. The design of the examinations is derived directly from the syllabus. It is in this sense that the examinations are syllabus-based.

In the American system of testing, the ideal test is curriculum neutral, meaning that, when taking a test, no student should be advantaged by having taken any particular curriculum, notwithstanding the fact that many experts have shown that it is not actually possible to construct a test that is curriculum neutral. It is this feature of the American style of testing that has made American teachers hostile to the idea of teaching to the test, because in this country, teaching to the test means teaching to a test that is expressly designed not to test what the teacher is teaching. Teachers in most other countries cannot understand why our teachers do not want to teach to the test, because, in their countries, the examination is designed to determine whether the student has learned what the teacher was trying to teach. Curriculum based testing and examination systems have an enormous advantage over systems that are not curriculum based: They produce much higher levels of student achievement. It is a cardinal principle of test construction in the United States that, for a test to be valid, students must have had an opportunity to learn the material being tested. But it is not possible to get an opportunity to learn if the curriculum one has taken is not what is being tested. Conversely, and crucially important, if the student is examined on the specific course that has been taught, and the

student therefore knows the goals of the course, what he or she is supposed to read, and what assignments are supposed to be completed; and the student has been given a set of instructional materials that were chosen because they are perfectly matched to those requirements, and the teacher has been well trained to teach the material to students from many different backgrounds, and the design of the examination has been not just aligned to but actually derived from the design of the curriculum, then these students can be expected to do far better on the exam than they would if the test or exam was not based on that curriculum, the student was not exposed to a course matched to the exam, the instructional materials were not designed to support that particular course and the teacher was not prepared to teach that course well to students of that students' particular background. For all these reasons, students using curriculum based systems can be expected to perform much better than students in which the governing idea of the testing system is that tests should be curriculum neutral

5. Though the testing and examination system should be curriculum-based, that does not necessarily mean that there has to be one national curriculum. Other countries have found ways to have national systems that offer choices among different curricula for states and schools. And they have figured out how to do this in a way that makes it possible to have multiple curricula while still setting common standards for those curricula. The United States should learn from those countries and offer a choice of curricula, each with its own matching test or examination, all set to a common challenge standard, so that none are or are perceived to be easier than the others

It is very unlikely that the Congress or the states will ever agree to a single national curriculum or a single national test. Nonetheless, there is now strong aversion to a system in which the states can each set their own performance standards for their own accountability systems, with the result that some of the states with the strongest performance as judged by their own tests also show the worst performance according to the NAEP assessments. There must be some way, even if the states or even schools use different assessments, of holding them all to one common standard. If the principle stated in #5 above is observed, then the system must accommodate both multiple curricula and multiple assessments, each tied to its own curriculum, all set to a common performance standard.

6. The testing and examination system should encourage the development of a balanced curriculum

NCLB was designed to hold schools accountable for the teaching of mathematical and English literacy, and, to a lesser degree, science. The effect has, in many places, been a radical narrowing of the curriculum to this very limited menu of subjects. The national testing and examination system should be designed to correct this problem, without imposing a national curriculum.

7. The standards should be embedded in the instructional system and closely tied to the testing and examination system

The Common Core working group is developing what we would describe as narrative standards, that is, statements in the form of: Students should know this and be able to do that. Many of the highest performing countries have standards statements of this sort and they are an important anchor of their instructional system. But, in those countries, these narrative statements are only part of the standards system, and the standards are more closely tied to the assessment system than is the case in the United States. The appropriate components of the narrative statements of standards are also found in the syllabi for the core courses in the curriculum, where the goals of the course, the topics to be studied, and the statements as to what the student is expected to learn are all described. In addition, the questions asked in prior year exams are all made available to the students and teachers, as are examples of student responses to those questions that earned top grades. In countries that do this, it is understood that the standards consist of all of these elements, not just the narrative statements as to what students should know and be able to do. This form of standards helps the student and the teacher make the jump from the necessarily abstract narrative statements to much more vivid and concrete images of what is expected. As one young elementary school Black student from a low-income family who was achieving far above what his teachers expected from him said to me one day: “If only someone had told me that this was what they wanted, I would have done it before!”

8. The tests or examinations should be designed to capture students’ higher order skills, critical thinking skills, creativity and imagination and, insofar as possible, measure performances much like those they will be called on to perform as adults in the ordinary pursuit of their life and work.

The previously largely separate economies of the globe are now rapidly integrating. The consensus among economists is that the standard of living of the people of the United States will steadily decline unless the members of our work force are not only much better educated but educated differently, for jobs in which there will be a great premium on creative and innovative thinking, on learning quickly things one was not taught in school, on deep knowledge in several arenas and on the ability to apply what one has learned to complex, quickly changing problems unlike those found in the back of the chapter in the textbooks one used at school. The testing and examination system the United States develops for the next few decades must be able to measure these qualities in our students. Students who will be required as adults to write long analytical papers cannot be adequately tested by asking them to write short three paragraph essays. Students who will be required as adults to come up with original answers to complex questions cannot be measured by computer-based multiple choice tests in which the student is asked to select only from answers provided by the maker of the test. Students who will be asked as adults to come up with powerful arresting graphic arts images cannot be measured by tests that do not permit the student to do graphic art. By limiting our accountability tests to measuring things that can be measured by computer-based multiple choice tests and short answer essays of only a paragraph or two, this country is denying itself the opportunity to measure the very capacities on which the

competitiveness of this country is most likely to be determined over the next few decades.

Measuring what needs to be measured will require multiple forms of assessment, including assessment of extended assignments of many different forms. Much of what needs to be done will require the innovative and integrative use of technology to make the system effective and efficient. The best assessments will be performance assessments, calling for the performance of tasks that come as close as possible to the kinds of tasks the student will be called on to perform as an adult.

9. As much as possible, high stakes assessments should mirror the form that we want instruction to take and the tasks set for the students should call for responses as much like those they will be called on to produce in further education and work as possible.

When the stakes are high for teachers and students, the teachers have very strong incentives to teach the students what they need to do to produce the answers demanded by the tests, whether they actually understand the material or not. This is what leads to the most prevalent form of “test prep.” The students will be able to do problems of the exactly the same form as those they drilled for, but because they do not understand the logic of the mathematics they are doing, they will be at sea when faced with a question that calls for the same mathematical knowledge but which is presented in a different form. What is wanted is teaching that helps the student understand the underlying conceptual structure of the subject and develop a strong analytical capacity, and enables them to synthesize new insights from different perspectives on an issue, for example. If these are the goals of the curriculum and the standards that lie behind it, then the assessment must demand these abilities. Traditional testing regimes focus on the whether the student has the right answer, and are indifferent to the question as to whether the student understands why that is the right answer and whether the student could produce another right answer to much the same question if the form in which it is asked changes significantly. Whether the student can do that depends on how the student was taught. Whether the student will get the kind of teaching that will lead to real knowledge depends, in a high stakes testing environment, on whether the form of the assessment is designed to mirror the kind of teaching one is looking for.

Similarly, we should be looking for assessments which, as much as possible, demand that the student perform assessment tasks that are, as much as possible, similar to the tasks that they will be called on to do in their further education and the work they will do. The only way to find out whether a student will be able to write a high quality 20 page history research paper is to ask that student to write one and then assess its quality. The only way to find out whether a student can produce a high quality original work of graphic art is to ask that student to produce one and then judge its merits. The same is true of that student’s ability to construct a robot that is able to perform certain prescribed functions to a set standard. It is essential that the design of the assessment regime begin with a specification of the kinds of performances that we want to assess and a consideration of the most effective way to assess them than with the assumption that assessment will be

limited to certain cost effective techniques and then ask how we can assess our standards within those cost constraints.

This criterion, whether we have assessments that mirror the instruction we want and reflect the performances we most value, is among the most important of all the criteria for our new national assessment system.

10. The tests and examinations should be valid for the purposes for which they will be used

As we have written elsewhere, this is a very large issue. The first set of draft standards issued by the Common Core working group are very impressive in many ways, both with respect to the quality of the standards and the degree to which they have made possible an emerging consensus of standards that few thought possible until now.

If we have a concern, it is only by way of putting down a marker for future work. The team that put the Common Core standards together has probably done a better job of validating standards said to be for college and work. But that is not saying as much as one would hope. The authors do tell us what sort of college or what sort of work the students who meet these standards will be ready for, and, on the face of it, the demands of the first credit bearing courses in our colleges vary widely (if one considers the range from Harvard and Stanford to the weakest of our community and technical colleges), as do the demands of different jobs.

This should not surprise us. There is surprisingly little research that would enable educators to say with confidence what kinds of mathematics, for example, are required to do what kinds of work, but the research that has been done leads this observer to the conclusion that we spend enormous sums to educate students in the kinds of math they are never likely to need in their work and much less time educating them in the kinds of math they are most likely to need. The research that has been done makes it abundantly clear that asking college staff and workplace supervisors to describe the kinds and challenge levels of knowledge and skill required to succeed in college and work is a wholly unreliable method of determining these requirements.

When we ask supervisors to tell us what education is required to perform the work they supervise, they typically tell observers that the work requires the level of education they had when they do that job or the level of education that the current incumbents have. But they actually have no idea whether those levels of education are in fact necessary to do the work involved. This kind of research is the province of industrial psychologists. Because the methods they employ to do this kind of work properly are very expensive, not many jobs have been analyzed with the rigor required to properly inform educators as to the content and performance standards needed to prepare people for work. It is also true that the requirements are powerfully affected by the way work is organized. At any one time, by definition, only a minority of positions in a particular field are filled by workers who are employed by companies using advanced forms of work organization, which typically require higher order skills and greater knowledge than the jobs with the

same name in other organizations. These are the jobs of the future and the ones that should be used to define the requirements for current education programs, because it will be years before the current students are in responsible positions in the firms that are the key to future economic well being of the United States. These are serious issues in determining the validity of standards purporting to represent the demands of the workplace.

Much the same thing is true of the validity of “college ready” standards. When we gather college people in a room together to tell us what their standards are, it is important to understand that they do not have any standards, in the sense of a fixed level of proficiency below which they will take no one and above which all will be admitted and none need to take remedial courses. The reality is that they want to get the best freshman class they can get but they will do what they need to do to fill their seats. That means that their functional standards vary from year to year, with the fluctuations in the relationship between supply and demand. It is also true that there is a status hierarchy among postsecondary institutions and those lower in that hierarchy are reluctant to admit that their standards are lower than those higher in the hierarchy, so when they are around a table together in a standards-setting session, there is a natural tendency to exaggerate standards. An alternative is to look carefully at the actual course content in a sample of initial credit bearing courses in a carefully chosen sample of postsecondary institutions in a state and make an independent determination of the content and performance standards a student would have to meet to be successful in those courses.

Similarly, the draft standards are said to be internationally benchmarked. But this benchmarking appears to have been done by collecting the formal narrative statements of standards from a sample of advanced industrial nations. Just as in the United States, those standards may represent anything from aspirations to explicit requirements. They may determine the curriculum that is actually taught in the schools or may have little to do with it. They may be very closely aligned with the tests or examinations that are used, or may, as is often the case in this country, be only vaguely aligned with the tests and examinations. They might be indicative standards, that is, used for the most general guidance of teachers, and therefore quite ignorable, or they might be the basis of high stakes testing that will determine whether a teacher keeps her job or a student gets to go to college. The standards the researchers looked at might be for all students or only for a select few. To my knowledge, the benchmarking research done thus far has not addressed these questions, and so we do not have a very good idea how the standards we are developing relate to the standards actually used by other countries for purposes similar to the purposes we have in mind.

We do not in slightest believe that these shortcomings should be an excuse for failing to implement the Common Core standards. They are head and shoulders above the standards that most if not all of the states are using. But we have a long way to go before we are using standards that have the kind of validation against the actual demands of work and further education that they should have. And the same goes for the claim that the standards are internationally benchmarked. It is important that the country make the investment it needs to make to get these things right.

11. The tests or examinations should be fair and reliable

There is a vast literature on this and American psychometricians are probably the best in the world at assuring that these criteria are met. This is partly because American law creates substantial liabilities for organizations that allocate opportunities in our society based on the use of tests that are not deemed by the courts to be fair and reliable. The professionals in the field of psychology have developed and are now revising standards for validity, reliability and fairness. Whatever system of testing and assessment is developed to implement the new Common Core standards will have to meet those requirements.

12. The new system should be affordable and available soon

Perhaps this criterion is obvious, but it needs to be stated. A word of caution is in order here. The typical American state accountability system costs on the order of \$23 to \$25 per subject per student tested. The typical examination system in the countries that outperform us costs at least twice that. The difference is mainly the difference between systems that are mainly reliant, in our case, on computer-scored, multiple choice tests, and examinations in other countries that rely more heavily on extended essay-form responses that are scored by human beings.

As always, one gets what one pays for. If this country is content with measuring the limited range of things that can be measured in the way we typically measure them, then we will have to be content with the kind of national and personal incomes that will be the lot of people and national work forces that are limited to those skills. We would argue that one of the most productive investments this country can make in its future is the development and use of examinations that match our ambitions for our children.

What we are proposing here is going from a testing system that constitutes about .003 of total annual expenditures on elementary and secondary education to somewhere between .006 and .01 of annual expenditures. What a shame it would be if the United States continued to slip ever farther behind the education accomplishments of other industrialized nations because we were unwilling to spend a tiny fraction more for a measurement system adequate to our ambitions for our students. An even greater shame if we stopped to consider that all of these other countries have been investing in their examination systems at these slightly higher levels for a very long time.

Besides being affordable, the system should be available soon. As the reader will see in the section below on implementation, we believe that a sensible plan for the new testing and examination system would unfold over seven to ten years. Parts of this plan will require years of research and development before the products of that research and development can be field-tested, demonstrated and deployed at scale. But other parts of what is proposed here can be fully deployed at scale in three to four years. As the reader will see, we believe that it is possible to make enormous improvements in the system by using certain curriculum-based assessments that are available now, deployed in a

substantially redesigned system. That will make possible major gains in student achievement in the near and middle term, while creating a structure that will support even greater gains in the longer term, without having to wait for years before we see the fruits of an elaborate research, development, field testing program become available for widespread deployment.

Overview of the Testing and Examination System Design

The American College Qualification and the high school system for standards, curriculum and assessment

At the heart of the design is the creation of the American College Qualification (ACQ), a new credential that indicates the holder is qualified to begin studies in 2-year and 4-year open admissions postsecondary institutions without having to take any remedial courses. The examinations for this diploma would be offered to students as early as the end of their sophomore year in high school. The passing point on the examinations would be set to the level of literacy required to succeed in the initial credit-bearing courses in open admissions 2-year and 4-year open admissions colleges. Students who pass those exams and thereby earn their diploma would be eligible to enroll in any 2-year or 4-year open admissions postsecondary institution in their state the following fall without having to take any remedial courses.

Students could choose among different providers of these examinations approved by the National Examinations Board (see below), though all the examinations would be set to the same standards (also set by the National Examinations Board). Each exam system would be part of a program of study, selected from among the best such systems in the world, available in English for use in the United States. All offer a set of courses constituting a complete core curriculum, syllabi for each course, instructional materials aligned to the syllabus, high quality assessments, professional scoring and training for the teachers who teach the courses. Among those that would be suitable, with minor modifications, and available today for use at the high school lower division level are the ACT QualityCore program, the University of Cambridge International General Certificate of Secondary Education program and the Pearson/Edexcel International General Certificate of Secondary Education program.

While students who pass their lower division exams could elect to receive their ACQ and go on to an open admissions college, they could, alternatively, stay in high school and take another program of studies intended to prepare them for entrance into a selective college. Among the programs of study available now for such use would be a program made up of Advanced Placement courses, the upper division ACT QualityCore program or a similar program, the International Baccalaureate Diploma program, the University of Cambridge Advanced International Certificate Program and the Pearson Edexcel A Level program.

Programs of studies approved for use in preparing for their American College Qualification examinations would include, at a minimum, courses in English (including literature), mathematics, the sciences and technology, history and civics, art and design, and music. The program of studies would also include attention to certain cross-cutting skills such as critical thinking, higher order thinking skills, creativity and innovation, and the ability to apply what the student has learned to complex, real world problems.

The grades for these courses would be based on a combination of the students' scores on their final examinations and their grades on extended assignments given during year, scored by their teachers and moderated by the provider of the examinations in order to assure the validity of the teacher-given grades. These assignments could range from a 25-page history research paper to the design and construction of robot to meet states specifications to the painting of a work of art. The examinations themselves would consist largely of questions requiring responses constructed by the student, rather than responses to multiple choice questions constructed by the test maker. To the extent possible, these examinations would take advantage of the dynamic modeling and interactive capacities of modern computer technology. It is very important that the design of the assessments be driven by the constructs underlying the standards and the curriculum rather than by the conventions of what can be measured by conventional American testing systems.

Each high school would be required to offer at least one such program of study, selected from a list of such programs of study approved by the state. Each of these programs of study, in turn, would have to be chosen by the state from a list of such programs approved by a National Examinations Board. The National Examinations Board would be required to select only the best programs of study used anywhere in the world, and available in the English language for use in the United States, including, but not limited to, those developed in this country. States wishing to offer a program of study unique to that state, alongside others, could do so, provided that such a program of study meets the standards set for programs of study by the National Examinations Board.

The states would be empowered to award the American College Qualification to any student who achieves the necessary grades on their examinations. In order to receive their American College Qualification, all students in all states would have to demonstrate that they have achieved the nationally-set scores on English and mathematical literacy, as well as science and technology. But each state could establish its own passing scores in all the other required subjects and could add other subjects beyond those in the national core, at their own discretion. Thus each state would set its own requirements for the Qualification except with respect to mathematical and English literacy and science and technology, the passing scores for which would be set by the National Examinations Board. However all the examinations for subjects in the core would be set to common scales, so that valid comparisons could be made among the standards set by the states and student performance could be compared across the states in all the courses in the core. States would be encouraged to require their high schools to analyze the sub-scores of students who do not pass their examinations on the first attempt to determine what areas of the exams the students did not pass and put together a program for those students

directed at the areas in which they are weak, so as to improve their chances of passing on subsequent attempts.

The states would also be encouraged to waive the current course-based requirements for the high school diploma for students who meet the requirements for an American College Qualification, thus moving from a diploma based on time in the seat to a diploma based on mastery of the material they are required to study.

In this discussion of the programs of study and the examinations in which they culminate, we have provided examples of programs of study and examinations that already exist and shown how they could be used to greatly improve the performance of American high school students. The advantage of these programs of study and exams is that they already exist, which means that we can take advantage of the enormous amounts of time and money already invested in them and get a fast start on implementing a much more effective system than the one now in place. But there is no reason to stop there. Once this system is in place, it becomes possible to introduce other programs of study and associated examination and assessment systems that take full advantage of advanced technologies that could make possible remarkable advances in curriculum, instruction and assessment (about which more is said below). There is every reason to start investing in those advances now, but even more reason to put the basic structure described above in place today.

The National Examinations Board

The National Examinations Board would be constituted as a not-for-profit organization by and under the auspices of the Council of Chief State School Officers and the National Governors Association, under the terms of a Congressional charter. It would be given a Congressional charter, like the American Red Cross and the National Academies, which would make it eligible for Congressional appropriations, but it would not be part of the federal government. Its members would be chosen by the CCSSO and the NGA. A plurality of its members would be chief state school officers. Others would be leaders in higher education, general government, elementary and secondary education, business and the professions.

The Board would be responsible for producing and revising the content and performance standards for the subjects in the core curriculum at the high school level, and for certifying providers of programs of study as meeting the Board's standards. By content standards, we mean the content of what is taught. By performance standards, we mean the degree of mastery of that content expected of the students. It would also produce curriculum frameworks (see below) for those subjects extending from Kindergarten (where appropriate) through to the college-ready standard embodied in the American College Qualification. And it would produce assessments of school readiness for use by schoolteachers at the beginning of Kindergarten, as well as summative tests of English and mathematical literacy at the ends of grades 3, 5 and 8 and science at the end of grades 5 and 8, as well as resources for formative assessment at all grade levels. We will comment further on these assessment proposals in the section on K-8 assessment below.

The Board would adopt as the starting point for its content standards the standards for mathematical and English literacy now being created by the Common Core working group established under the auspices of the National Governors Association and the Council of Chief State School Officers; it would be responsible for converting those standards into criteria for certifying the offerings of potential providers of programs of study, including the criteria for judging whether the syllabi and examinations offered by those providers are acceptable to the Board. Thus the Board would have to specify the forms of assessment to be used in the examinations as well as the technical criteria the assessments would have to meet, including the criteria for the examinations as well as the criteria for assessments of other work products produced by the students on which the grades for the courses are to be based.

The Board would be charged with conducting continuing empirical studies on the demands of work and further education as part of its obligation to continually improve the validity of its content and performance standards. It would also be expected to conduct continuing research on the technical and practice requirements and resources for assuring that the tests and examinations produced under its auspices, as well as the system it uses for score moderation are fair and reliable. It would be empowered to conduct research intended to advance the use of technology in the delivery of curriculum and assessment. And it would be expected to adapt and extend the discipline of psychometrics to meet the demands of a standards and curriculum based assessment system that is itself expected to set the world standard.

The Board would do all the technical work necessary to assure that programs of study meet their standards (including standards of validity and fairness), that the examinations are set to a common scoring scale (assuring that a given grade on one is the equivalent of the same grade on another) and that there is a common passing grade for the examinations set to the mathematical and English literacy level needed to assure that those who pass are ready to do college level work in the initial credit-bearing courses in the nation's open admissions postsecondary institutions.

Every five years, The National Academies would be required to conduct independent validity studies of the content and performance standards established and revised by the National Examinations Board, and the Board would be required to respond to the observations of the National Academies in a public written response.

In this plan, the National Assessment Governing Board would continue to audit the performance of the American education system through the use of the National Assessment of Education Progress. It is very important that the agency charged with monitoring changes in the performance of the system not be same agency that is charged with providing or setting the requirements for the tests and examinations used as the basis of the national and state accountability systems. Nothing makes this need for separation clearer than the perennial and often fierce controversies in Britain over the interpretation of student achievement data coming from the British national testing system. Government takes great pride in the improving scores of students and the opposition

decries the lowering of standards that (it is obvious to them) lies behind those rising scores. Just as NAEP is now used as a check to monitor the performance of students in the states against the state standards, NAEP should be used in the future to monitor the performance of the system as a check against the data provided by the new testing system proposed in this paper, the one supervised by the proposed National Examinations Board.

A Curriculum Framework

Clearly, the design just offered will not work as well as it should unless students leave the 8th grade ready to do the work they would have to do to succeed in high school.

The first step toward that end is for the National Examinations Board to lay out a curriculum framework, beginning in Kindergarten and ending at the point at which the student has met the college ready standard,

By curriculum framework, we mean the specification of the progression of topics and sub-topics expected to be mastered in each subject, in the sequence in which they are to be mastered. That is the content. We also mean the specification of the performance level to be attained in each topic and subtopic by the students.

The progression reflected in the frameworks should reflect two related considerations. The first is the logical order of the unfolding of the subject as the student proceeds through the framework for that subject. By “logical,” we mean that each topic in the sequence should reasonably be seen as the logical prerequisite for the following topic. One cannot learn the later topic without first having learned the preceding topic. This is a judgment made on the basis of the intellectual structure of the disciplines underlying the subject (what a philosopher would call its ontology), and that judgment, to be well made, must be based on the underlying conceptual structure of the discipline.

The second consideration is related to the first, but is not the same. It has to do with what researchers are learning about the way students actually master these disciplines. This has to do with the way students construct knowledge as they learn. It is certainly related to the underlying conceptual structures of the discipline and with its internal logic, but it also has to do with the mechanisms of human cognition. Here we deal with the structures of knowledge that students carry around in their heads, the way they are built as the student interacts with his or her environment, the factors that affect the construction of accurate representations of knowledge, the factors that lead to the construction of mistaken structures of knowledge, and the other factors that affect the speed and efficiency with which humans add to the structures of knowledge they start with to build more complex and powerful structures.

Research on the developmental progressions of students through the curriculum is going on in many countries, but it is still in its early stages. Nonetheless, it is, we believe, very important that the construction of curriculum frameworks be one of the primary sources of data that we take into account as we build on the work of the Common Core initiative to construct the curriculum frameworks on which the K-8 curriculum and assessment

systems are built, building down from the Common Core college ready standards. Because this kind of research is still relatively undeveloped, we can expect that, as it matures, the frameworks with which we begin will have to be continuously modified as more knowledge becomes available over time.

The Common Core initiative put a lot of emphasis, rightly, we believe, on fewer standards. This was in part a reaction to the phenomenon, often noted, that typically is on display when the states create standards, in which everyone involved engages in a trading process in which we will support adding your standards if you support adding mine. It is this phenomenon that leads ineluctably to standards frameworks in which the topics are not logically related to one another as the students progress through the grades and to a plethora of topics so large as to make it impossible for any teacher to cover the waterfront of standards that is produced this way.

It follows from this analysis that it is not enough to have a framework that embodies a logical progression from topic to topic. It is also necessary to ruthlessly prune the progression of topics so that only those topics are included that are necessary for the students to have the knowledge and skill needed to meet the college ready standards at the end.

It is all of this that we mean by a framework, a framework that would unfold from Kindergarten to the end of grade 8, and on to the college ready standard. In this conception, though there would be a clear demarcation between the end of grade 8 and the beginning of grade 9, the progression in each subject would be continuous across the whole span of the framework.

The obvious question is whether this should be a grade-by-grade framework. My answer is that it should be an indicative grade-by-grade framework, but the conception of it as a grade-by-grade framework should be fluid and not rigid. Every teacher knows that different students progress at different rates through different subjects and even topics in the curriculum. By “indicative,” we mean that teachers and policy makers need to teach and to make policy in the knowledge that the framework, if it is well done, will be a reasonably accurate indication of where the average student should be if that student is on track to be college ready by the end of his or her sophomore year in high school. But there is no average student, and so the teacher and the policy maker need to make allowance for individual differences among students, most of whom will be ahead of the normal sequence in one subject and behind in another at any given time.

But we need to be careful here. One of the principles on which this design is based is the desirability of moving from time-in-the-seat systems to move-on-when ready systems, from keeping the time constant and the standards variable, to keeping the standards constant and time variable. The underlying assumptions are that all students can reach high standards, but that it takes some students more time to do so than others.

But this idea is easily abused. Teachers make judgments, sometimes unconsciously, that one student is more able than another, on the basis of which the student judged less able

is often given a less demanding curriculum. If that student achieves less at the end of the year in a move-on-when-ready system, the teacher could simply say that that student will need more time, and pass that student on to the next teacher with a clear conscience. After a few years of such experiences, that student will never catch up to his or her peers, no matter what.

The K-8 testing system offered below has an important premise, namely, that all students will, insofar as possible, begin each year ready to participate fully in the curriculum specified for the curriculum framework for that year. We are not assuming that all students will progress through all topics of every curriculum at the same speed. Some will need more time and small group and individual attention than others to begin the next year on the same footing as his or her peers. But most of that extra time needs to come before school, after school, on weekends and during the summer. It needs to come in the form of more intensive work on mathematical and English literacy during the regular school day if necessary. Students do progress at different rates and that fact has to be taken account of in a realistic plan for improving the performance of American students. But that does not mean that some students need to get moved from grade to grade, whether or not they are progressing at a satisfactory rate, falling ever further behind.

To some, this approach will sound like a worthy but unrealistic expectation. It is, however, precisely the formula that Singapore, with one of the most successful education systems in the world, has actually used for years.

All of that said, we will from here on out refer to the curriculum framework as being organized in the form of grade-by-grade standards. The reality, however, is that we have in mind something more complex than the image that is typically connoted by those words. The progressions for each subject will actually be continuous from topic to topic, and each will have its own associated performance standards. It is in the framework that curriculum (at a very high level) and standards become inextricably intertwined.

It remains to describe the narrative form the framework might take. For each subject in each grade, these content and performance expectations would be accompanied by examples of student work that meets both the content and performance standards. The content and performance statements (but not the student work needed to illustrate the standards) would consist of only a few very carefully written pages for each subject for each grade. From K through the college ready standard, they would not be intended to specify a complete, detailed curriculum and, in fact, would be intended to leave considerable scope for professional teachers to define their own curriculum, while at the same time giving them the tools needed to make sure that their students are on track to begin the next grade where they should begin it. These content and performance standards, in combination with the illustrative examples of student work, should be written in such a way that they support the development in the classroom of a culture based on standards, in which both students and teachers are constantly comparing the work being produced by the students to examples of work that meets the standards, thus providing the basis of a form of formative assessment that supports constant course

correction in the process of instruction if the student begins to fall behind the trajectory that student should be on.

What we have just described is the level of detail at which the curriculum framework should be described for mathematics and English literacy, and science and technology, grade by grade. As you will see in a moment, this framework would be used, among other things, to create a series of national tests under the auspices of the National Examinations Board in these subject areas.

In parallel with this system, the Board would also develop curriculum frameworks for each of the other subjects in the core curriculum for each grade. These would not be used as the basis for national tests, but would instead be used only as indicative frameworks by states, districts, teachers and schools that wished to use them as sources of ideas. The states could ignore them if they wished and be free to develop their own frameworks in these subject areas. It would, however, be in their interest to pay attention to these frameworks, because the specifications for these subjects for the board examinations would be derived from the frameworks for these subjects at the high school level, and so, if the states wanted to be sure that their students will succeed in the board exams at the lower division of the high school program, they would want to look carefully at the frameworks for those subjects for the lower grades.

The K-8 Testing System

It would certainly be possible to modify the current state accountability testing systems to reflect the sequence of topics contained in the curriculum framework described in the preceding section. One can easily envision improving that system with the use of more advanced testing technologies, including but by no means limited to, computer adaptive testing technologies. And one can imagine constructing greatly expanded test item banks to support such systems, of the sort that would be required for the full exploitation of computer adaptive testing. It may well be useful to include such capabilities in the next generation accountability testing systems, but, in my opinion, limiting our efforts to strategies of that sort would bring the country up far short of the opportunity we have to build a system that could support a major improvement in student performance.

Why a K-8 accountability testing system that is a straightforward adaptation of the current state accountability testing model would be a mistake

Why is that? First, because, as noted above, assessment systems that rely solely or largely on multiple choice, computer scored tests are strongest when it comes to measuring basic skills and weakest when measuring the kind of complex, higher order thinking and creativity on which the future of the American economy depends. The biggest mistake we could make is limiting what we measure to those skills that can be most easily, cheaply and quickly measured. This would certainly please our economic competitors, because nothing we could do would be more likely to deprive us of the skills we need to compete effectively in global commerce.

Second, because the most effective assessment strategy—that is, the one most likely to produce major gains in student achievement—is the one that is most likely to produce the kind of teaching and learning we want in our classrooms. Tests consisting mainly of multiple choice, computer scored items do and will continue to produce the antithesis of the kind of teaching we want in our classrooms. What is more likely to produce that teaching is assessment that is based on test items or prompts intended to produce student work of the sort that the best teaching produces: well drafted, substantial, carefully thought through papers; multi-step math solutions that require a thoughtful analysis of a real world problem to set up the solution; and so on.

Third, because the best classroom teaching is teaching that is set, for each student, to the precise point at which the material the student is asked to study is challenging, but not so challenging that the student gives up in frustration. It is not possible to provide that kind of instruction unless the teacher knows how the student is doing relative to the standard every day. Ideally, the teacher not only knows what the student knows and does not know, but also has some insight into the ways in which the student misunderstands the material and how those misunderstandings are getting in the way of learning the material correctly. The research clearly shows that this kind of formative assessment can contribute greatly to better teaching and student learning.

But timed multiple choice, computer scored tests are not the best way to produce this kind of knowledge about the student's understanding, if the aim is to go beyond basic skills to more advanced thinking skills and creative work, though technology can certainly help. A different approach is required.

There are two enormous advantages of assessment systems based on multiple choice, computer scored tests: they are cheap and they produce results virtually instantaneously. This makes them very attractive to those of us who hope to use such systems as the basis of hiring and rewarding teachers. Since they are cheap, one can imagine using them for every student and every subject, at every grade level, which would be necessary in any system in which they are used as the basis of teacher reward systems. Since the research shows that teacher effectiveness varies widely and nothing affects student progress more directly and powerfully than the quality of their teachers, the temptation to create a policy system in which teacher rewards are directly tied to student progress are understandably enormous. But the technology of assessment, combined with the circumstances in which it is used, will simply not support the use of student achievement data in this way, according to Ed Haertle, speaking for The National Academies Board of Testing and Assessment, arguably the most respected source of professional advice on testing and assessment. It is, in those circumstances, hard to make the case that the use of such systems is necessary to produce high student achievement or to make the case that such systems, if used, would stand up in court when subjected to legal challenge, as they surely will be.

We would urge those who have been hoping to create such a system to consider that there are many countries with education records superior to ours and not one of them uses student achievement data from its testing system as the primary basis of a high stakes

accountability system for teachers, much less to determine the hiring, promotion, or compensation of teachers. We are persuaded that there are effective ways to create systems that reward effective teaching and effective teachers without depending on value added measures of teacher performance in turn based on universal high stakes testing of students, in all subjects, every year.

The advocates of continued use of the NCLB model of accountability argue from a conviction that that accountability model is responsible for much of the improvement in the performance of low income, minority students reported in recent years. If this were true, it would be a powerful argument for keeping the current NCLB testing regime in place. But some of our most admired testing experts doubt that this is in fact true.

They argue that the design of the system has fostered an environment in which teachers of low-income, minority children have learned how to teach a curriculum dominated by repeated practice of problems exactly like the test items the students will be given on the accountability tests. As they get better at teaching these items, and students get better at learning how to respond to these items, their performance improves.

This would be perfectly all right if the test items captured the content that the student is supposed to be learning. But the test items and multiple choice format of the tests fall far short of capturing the material that the more advantaged students learn, material which is essential to learn if the students are ever to be truly college ready. Thus the improved scores present the illusion of improved performance by the student, but not the reality.

This effect of the current regime is reinforced by a technical feature of the usual test construction procedure. The tests administered each year are released and cannot be used again. But it is important to be sure that each successive test is set to the same standards. So the test makers include a significant number of items in successive tests that are the same as items in the previous year's test. These "anchor items" provide the means for making sure that the standards do not vary from year to year. But the prevalence of these items reassures the teachers that they can teach the students to solve a particular format of mathematics problem, for example, without really understanding the mathematics.

If the form of the problem is identical and only one or two of the variables in it are different, students can employ the same rote procedures to solve it that they practiced during the year. That will work on the accountability exam, but woe unto that student when he or she has to work problems that look different but actually require similar mathematical thinking. They do not know what to do.

Thus it is not at all clear that the NCLB testing regime is contributing in an important way to improved learning for low-income and minority students. This is not to argue that NCLB has failed in its intention. It has, in my view, been brilliantly successful at focusing the nation's attention on the actual performance of the most vulnerable groups of students in our society and it has created an environment in which it is not enough to provide increased resources to these students if the institutions that receive those resources cannot demonstrate that those resources are being used effectively to improve

the performance of the student for whom they were intended. Those achievements are hugely significant, and they will last. But it is important that we not persuade ourselves that real gains for the most vulnerable students depends on keeping in place the system of grade by grade and student by student testing that is a cornerstone of the NCLB design. The evidence for that proposition is very shaky.

We know of no country with a high performing education system that does grade by grade testing for high stakes. In fact, few do grade by grade testing at all. In most of the high performing countries, high stakes testing is high stakes mainly for the students, and it comes at the end of what we would call the sophomore year in high school, when they take their qualification examinations. Prior to that point, most national and state testing is done for the purpose of helping teachers to know where their students are with respect to the national or state curriculum, so that they can pitch and organize the instructional program to address their weaknesses and give them the support they need to succeed. A number of countries conduct national or state tests at the end of sets of grades, partly for the same purpose that NAEP is intended to serve (help policy makers and the public understand how the system is performing) and to hold schools accountable for their performance. There is no high performing country that we know of in which the decisions as to which schools are declared low performing and which of those schools will be sanctioned are made on the basis of student performance test data alone. If these data are used for these purposes, they are used to provoke a visit to the school by the authorities, and it is the information produced by that visit that becomes the basis for decisions about the future of that school.

In general, as you will see below, it is this pattern that we believe the United States would do well to follow. It is quite possible to argue on the evidence that adoption of a system of this sort is likely to lead to greatly improved student performance in the United States. It is impossible to argue on the evidence that radically different systems will lead to greatly improved student performance, because they have not been tried on a national scale. If the United States was out in front of the pack of nations, one could argue that we ought to try something different if want to increase our lead. Given that we are far behind, it seems reasonable to adopt the modal strategies that have been followed by other nations that have been far ahead for many years.

Thus the purpose of this system would be to provide information to teachers and principals to enable them to make adjustments to the education programs of students who are falling behind, so as to enable them to catch up, and to provide information to district managers and states to enable them to identify schools that need help, or, having failed to respond to that help, need either new management or to be closed.

With key exceptions, this design would not use the K-8 testing system for purposes of high stakes assessment, either for the students or for the teachers. While it would use the data for purposes related to first helping, but then, if necessary, changing their management or faculty or even closing them down, the information this system would provide for those purposes would be indicative and not conclusory. That is, the

information from this system would be combined with other information that would be as important, or even more important, in determining the fate of schools.

So the principal purpose of the testing regime should be to help teachers track student progress along the curriculum framework for a given subject at a given grade level and to enable the authorities to identify schools that are not succeeding in moving their students, or key groups of students within the schools, along at the rates indicated by those curriculum frameworks, so that they can take appropriate action.

The backbone of the K-8 testing system would be tests of mathematical and English literacy designed to be administered at grades K, 3, 5 and 8, and tests of science to be administered in grades 5 and 8. All would be keyed to the curriculum framework for those subjects developed by the National Examinations Board, and the Board would be responsible for the development of these tests. The states would be required to administer these tests to all students. There would be no stakes attached to these tests for the students. The results of these tests would be used by the teachers to formulate a plan for the students in their classes for the year following the year in which the test was administered, for informing the public about student performance in the school and by the authorities to identify schools in need of help (though this data would be used only in combination with other data, as the reader will see, to make decisions about schools). These tests would include more multiple choice, machine scored items than the board examinations, but they would also include substantially more and longer open-ended responses than is typically the case with current state accountability tests.

All would be given in the spring of the year, except for the Kindergarten test, which would be given at the beginning of the school year.

Some will be surprised by the suggestion that the Board produce assessments to be used in Kindergarten. We do not have high stakes assessments in mind, but rather assessments of Kindergarten readiness to be administered by Kindergarten teachers when they first receive their students, to enable them to understand where each student is in relation to the factors that research tells are most likely to affect the readiness of the student to profit from the Kindergarten experience in developmental terms. Among these factors is vocabulary, a critical determinant of a child's future educational development. The data from these assessments would be used both by the teacher to frame the instructional and support program for her charges, and also by the locality and state to assess the adequacy of the supports they provide to the Kindergartens in light of the actual needs of the students.

The 3rd grade test is intended to be a summative test of the primary years, and is intended again mainly to enable the 4th grade teachers to accurately gauge the literacy abilities of the incoming students in order to give them the program they need to succeed. Students who are not reading well at the end of third grade may have great difficulty ever learning to read at grade level, and will certainly need strong programs of reading assistance if they are to have a decent chance of meeting the college ready standard by the time they leave high school. At the same time, this test is intended to provide indicative

information needed by administrators to identify elementary schools that might be in need of on-site inspections to determine whether they need help, and, if so, of what kind. The 5th grade tests would come at the end of elementary school and the 8th grade tests would come at the end of middle school and would be used for much the same purposes.

The 3rd, 5th and 8th grade mathematics and English literacy tests and the 5th grade and 8th grade science tests would be administered as secure tests by the schools under the supervision of the state. The others could be administered whenever the state, district or school wished.

These summative tests, to be given at the end of grades 3, 5 and 8, would not look very much like today's state accountability tests. They would include multiple choice, computer scored items, but assessment of that sort would account for much less of the total assessment package than it does in today's state accountability tests. There would be much more reliance on performance items that mimic the kind of instruction the students should be getting, more emphasis on performance tasks that seek to capture the kinds of tasks that the student will be expected to do later on in their education and at work, particularly at the 8th grade level. Not all of these item types would be included in the timed tests. Some would be given during the course of the year, embedded in student coursework. Many other items of the same sort would be made available by the National Examinations Board as formative assessment items for teachers, all tied to the curriculum frameworks produced by the Board. Many items, both secure items intended for use in summative tests and public items for use by teachers to gauge the progress of their students against the curriculum frameworks, would be presented in digital form. This item bank should take advantage of modern computer technology and the internet, especially the capacity of the such systems to provide dynamic models of very complex systems and to provide manipulable environments simulating those systems that can be used by students to demonstrate their understanding of complex systems and to solve problems of design.

Thus the hard distinction we currently make between formative assessment and summative assessment would be blurred. If we were talking here about tests that were very high stakes for either teachers or students, all items intended for such use would have to be fully secure and could probably only be used in timed tests in secure environments. But we are not talking about such high stakes uses. For the most part, we are talking here about no stakes for the students and low stakes for the teachers, since the use of the data produced by these tests would never be the exclusive basis for making decisions about teachers' compensation, principals' compensation, or the future of a school.

That being so, some part of the score of student on the summative tests could in fact be their scores on assignments embedded in the curriculum and used to comprise some part of the final score or grade on the test. Such items would look just like most of the items produced by the National Examinations Board to support teachers' formative evaluation. One of the most important aspects of the K-8 system would be the development of a robust set of resources for teachers to assess their students' progress during the year as

they go through the curriculum framework. Here, the National Examinations Board should provide a rich assortment of resources for formative assessment in the form of an item bank keyed to the curriculum framework and standards that teachers can draw from to gauge the progress of their students at points of their own choosing, using scoring guides and rubrics provided by the Board.

Earlier, it was suggested that the National Examinations Board should construct indicative curriculum frameworks for the other subjects in the core curriculum. Here, too, we believe it would be useful for the Board to provide resources for those who choose to use them that could be used to examine students as they progress through the indicative curriculum frameworks for those subjects. This does not mean building tests or examinations in those subjects for K-8, but building item banks of prompts and questions along with rubrics for judging the adequacy of student responses that teachers could use if they wished at any point in the year to assess the progress of their students against the indicative curriculum frameworks and standards.

Notes on the Use of Technology

Much has been written on the potential for the use of technology in assessment. Many have also written on the possibilities for transforming instruction through the use of technology. And some have observed that a great deal might be achieved by creating systems in which instruction and assessment are almost inextricably intertwined using advanced forms of instructional and assessment technology.

We do not intend to summarize those literatures here, but only to bring them from the background of this discussion into the foreground. Australia now has underway an ambitious program to construct a form of Board Examination System that will be largely computer and web based. If our aim is, as it should be, to build a state of the art system of instruction and assessment for this country, we would do well to follow what the Australians are doing and to build a board examination system built on a similar model, as one of the alternatives we offer to our students.

There is no reason in principle why the syllabi described earlier could not appear on a computer screen held by the student, along with all the materials referenced by the syllabus, including all the links among those materials, wherever they might reside on the internet. There is no reason why the formative assessments could not be accessed by the same computer, or why those formative assessments could not include dynamic models of everything from economic systems to living organisms or ecological systems, which could be manipulated on the screen by the student in response to prompts in the exams administered on these computers. Conceived of this way, the computer or computer-like device becomes a portal to an unimaginable range of educational resources and assessment methodologies, all configured in a way that is designed and structured to produce learning.

My aim here has been to suggest a structure for an instructional system that could endure and prosper for some decades until it, too, is outmoded by the passing of time. Any such system will have to be framed in the high likelihood that the kinds of technological learning environments just described will come into widespread use in the not greatly distant future.

The system proposed here could do just that. The system would work for conventional organizational structures and settings for educating our young people but it does not require them. Indeed, the idea of setting high standards and then letting students reach them at their own speed and in their own way is highly compatible with the kind of technological vision just conjured up. The structure we are recommending is amenable to very traditional instructional forms, but just as amenable to high technology delivery systems in which a student never goes near a classroom. It can accommodate forms of assessment that have been around for a century or more, but it can also work for the most advanced forms of assessment.

It will take time to develop these systems and resources. But it is possible right now to implement systems that use technology to capture student work of many kinds in digital form so that it can be shipped by fiber optic systems and satellites to teachers and others who can score it while displaying on their screen the work to be scored as well as the rubrics and examples of student work that enable to score the work quickly and accurately. It is possible to take the results of the work of scorers and virtually instantly collate that work with the work of other scorers to produce reports that can be shipped digitally to anyone who needs the information, with appropriate password protection, and so on.

We will not leap into the future all at once. We will get there in fits and starts, some of us faster than others. The new national system should make maximum use of proven technology at every step of the way, in increments, as the technology becomes available. But, at each step, it will have to provide for those whose access to advanced technologies is limited, and that will not be easy.

Implementation

It will take seven to ten years to design and fully implement a system of the kind just described, perhaps more. The changes implied are too large and the infrastructure that must be created to support them too undeveloped to do it any faster than that.

Some of what has been proposed, however, can be done in the two or three years, other elements in four to six years, and some parts will take many years. In short, those who are tasked with laying out the plan for developing and implementing the new system of assessment need to conceive of it as unfolding over time, rather than being implemented all at once as a unified system.

One example might be useful. The National Center on Education and the Economy has been assembling a consortium of states willing to pioneer the development of the kind of board examination system for high schools described above. The states in that consortium are planning to initiate the first cohorts of demonstration schools in their states—10 to 20 schools in each state to start—in the fall of 2010 and 2011. They should be prepared to begin to scale up from the demonstration program to statewide operation beginning in some cases as early as the fall of 2012. In all likelihood, on the order of 10 states, some 20 percent of all American states, will constitute the initial membership of the consortium. Other states will be welcome to join later. One could reasonably expect that, assuming the initial demonstration is successful, many states will wish to do so.

NCEE will be constituting a governing board for the demonstration program structured along the lines indicated above. We plan to work closely with the National Governors Association and the Council of Chief State School Officers as we do so. NCEE will provide the staff support needed by the governing board to get the program off the ground. The Bill and Melinda Gates Foundation and NCEE have jointly agreed to fund the initial planning and operations of the program.

We will need to do the technical work necessary to 1) determine empirically the standard of literacy in mathematics and English needed to be successful in the initial credit bearing courses in open admissions colleges, and 2) use that information to set the passing standard for all the board examination systems our states will be using. Once that work is done, it would be easy to use the same technical strategies to set other instructional programs and their associated examinations to the same standards. In this way, it would be possible for us to set any well constructed high school instructional program, including those developed by other consortia, to the same truly college ready standard that our board examinations will be set to.

In time, the work now being done under the auspices of NCEE could continue to be done under the auspices of the NGA and the CCSSO, as proposed above. NCEE could continue to provide staff support or other arrangements could be made to provide the necessary staff support. Once the operation comes under the auspices of the NGA and the CCSSO, the organization could become a Congressionally chartered organization and become the direct beneficiary of the Congressional appropriations process.

Later, when the infrastructure for such a system is firmly in place in a growing number of states, attention could be turned to the next wave of development work to produce much more advanced board examination systems, incorporating, for example, much more sophisticated technology and more sophisticated assessment systems based on new technologies.

In the meantime, the Department of Education working alone or in concert with the CCSSO and the NGA could structure the competition for the funds carved out from the Race to the Top fund to announce and then competitively award a contract for the construction of a K-8 system of assessment along the lines recommended above, thus

assuring that the whole system would be aligned when the K-8 assessment program and the high school assessment program were fully implemented.

The K-8 system could not and should not at the outset incorporate all of the bells and whistles described above. It could get started with just a subset of them, working over time to put the rest in place.

Accountability

Much of the basic accountability system design has already been foreshadowed above. It remains to bring its elements together and to add a few points.

The design assumes a complex interplay between the federal government and the National Examinations Board. We can safely assume that the federal government will continue to have a strong interest in maintaining an accountability system that provides strong incentives to school professionals to work as hard as possible to improve the performance of low-income and minority students. If anything, the federal role in a national accountability system is likely to get stronger rather than weaker as more people over time come to see the performance of our schools as intimately linked to our national economic performance.

At the same time, this plan assigns to the National Examinations Board, an instrument of the states, the key role in defining and supporting the national testing system.

But the two have to operate hand in glove, as it were. We think this is perfectly possible, but it is important that the reader understand that this kind of partnership is a premise of the plan.

In this system, the federal government would focus mainly on school accountability for the progress of students against the national standards for mathematical and English literacy and science. It would do so against a high school standard for college readiness that is explicit and empirically determined. The tests administered at grades 3, 5 and 8 and the examinations administered at the conclusion of the lower division curriculum would be the principal measures used to enforce this accountability. The content and performance standards for these tests and examinations would be the same across the whole nation and would be set nationally by the National Examination Board. That Board would indeed be national but it would not be an instrument of the federal government since its membership would be controlled by the states and not by the Congress or the President.

The distortions in the curriculum that occur when the schools are held accountable only for literacy in mathematics and English would be averted by asking the National Examinations Board, on behalf of the states, to decide on what subjects are to be included in the broader core curriculum and by having the federal government require the states to decide on the standards to be set for all the other subjects in that core curriculum for

admission to their open admissions colleges. The high schools in the states would be held accountable for the performance of their high schools against those standards, and the federal accountability standards would stipulate a performance floor for a school's performance in mathematical literacy, while still taking into account the school's performance in these other subjects.

The states would be required to assess the other subjects in the core curriculum on a sampling basis in grade 5 for literature, social studies, art and music, and in all of the core subjects in grade 8. They would have to assure the federal government that the content and performance standards for these subjects at these grade levels had been set in such a way that students achieving passing grades were on a path likely to get them to the college ready standard set for high school in the these subjects by the National Examinations Board.

The states would be required to publish, every year, the scores for every school in all the subjects tested by this testing regime, and to publish the scores of each major designated subgroup for mathematical and English literacy and science.

The states would be further required to establish systems for regular inspections of schools by teams of experts with the expertise needed to accurately assess whether the board examination systems were being well implemented in the high schools and whether the core curriculum was being well implemented in the K-8 schools. The inspectors would be employees of or contractors to the states, not the federal government.

These inspection teams would be required to diagnose the nature of the problems in schools in which significant numbers of students were not making adequate progress against the curriculum frameworks (including the content and performance standards) and to make recommendations to the schools for correcting those problems. These inspections should be made on little or no notice to the school staff. All schools should be inspected on a five-year schedule, but the date from the board examinations, the national literacy tests and the state sample tests for the other subjects in the curriculum should be used to trigger more frequent visits by inspection teams when the data from those assessments indicates the likely existence of problems that need their attention.

The state accountability systems should be established in such a way that it is understood by all parties that failure to properly implement recommendations made by the inspection teams will lead to the replacement of the head of the school, replacement of other faculty members or even the closing of the school altogether.

The states should be required by federal law to create public accountability programs that make available to the public as much data on the background of the student body, the nature of the school program, the performance of the student body over time against state and national standards, pass rates on board examinations at each grade level, and so on as possible, including separate reports for each protected group of students. These reports should report the raw data, the data relative to other schools with similar student bodies, the rate of improvement over time relative to the rate of improvement for the state as a

whole and to other school similar students, and the performance of the students in relation to what the performance would have to be for the students to be on track to be ready for college by the time they leave high school. School inspection reviews should also be made public, as well as the written plans produced by the school spelling out the school's response to the inspectors' reports.

The Advantages of the Proposed System

Having laid out the outlines of the proposed system, it is important now to go back to the criteria with which we began and to assess the degree to which what has been proposed is responsive to those criteria.

The system described here is designed not simply to measure student progress but to radically improve that performance. The most serious problem in American education is the enormous failure rates in our high schools and the even greater failure rates in our postsecondary institution, especially our open admissions institutions. This plan is designed to reverse those failure rates, by making it very clear to our young people what kind of school work is required to get into college and succeed there and then by organizing a curriculum that will enable them to do that kind of work and by training their teachers to teach such a curriculum successfully. The United States has never ever had a system in place to do that.

The key is an examination system that is based on both standards and curriculum, not just standards. It is a system in which standards, curriculum, instruction and assessment are seen as inextricably related.

The plan does not make the mistake of treating all colleges as if they had the same standards. They don't. Access to a very large fraction of jobs in this country—from cosmetologist to computer systems manager and automobile mechanic—comes through our two-year colleges. But those same institutions can also open the door to four-year institutions and graduate study. The education required to succeed in the initial credit bearing courses in these institutions is not the same as that required to get into state flagship institutions, the Big Ten or the Ivy League. That is why we have devised a system that is pitched to this group of institutions. Doing so makes it plausible that we can get the vast majority of our students ready for college by the end of the sophomore or junior years in high school so that students have the choice of going directly to open admissions colleges or staying in high school to prepare for selective university.

If we can do that, then we can save enough money to pay for the additional help that most of our students will need to reach these standards. In the end, the costs of our elementary and secondary system will probably be about the same, but we will be using our funds much more wisely to get much better results.

Perhaps the most potent criticism of No Child Left Behind is the way it is left up to each state to determine the level of proficiency towards which its students would march.

But NCLB is also widely and justly criticized for narrowing the curriculum and dumbing it down, putting a lid on achievement for many students. This is very dangerous for this country at a time when the changing dynamics of the global economy demand that many, many more Americans demonstrate world class levels of achievement in the core subject in the curriculum as well as high levels of critical thinking ability, creativity, innovation and capacity to solve complex problems.

At the same time, NCLB has arguably raised the achievement of many low-performing students who had little chance of success in the system before NCLB. The object of policy ought to be to greatly broaden the curriculum beyond basic literacy and set standards as high or higher than those of the best performing countries in the world, while at the same time keeping in place a demanding accountability system that will enable the most vulnerable among us to make steady progress toward higher achievement.

That is just what this plan is intended to do.

The single biggest reason that the states were allowed to set their own proficiency levels in NCLB is because no one wanted the federal government to set national standards or a national curriculum or a national test. In this plan, the National Examinations Board, not the federal government would set a single proficiency level in literacy for all the states, and it would do so based on empirical evidence concerning the actual demands of work and college course taking. It would also superintend the process for creating national tests of literacy at four grade levels, from Kindergarten through eighth grade. And it would fix the passing standard in literacy for the national board exams. There would be no room for the states to set their own proficiency levels in this vital area or in science. The National Examinations Board would be a creature of the states, but, because it would be Congressionally chartered, would be eligible for direct funding from the Congress, and could therefore be assured of getting the funds in needed to do its work properly.

The data from the sampling tests and the census tests of literacy proposed in this plan should enable the states to assemble a rich picture of the schools' records in enabling student literacy, from Kindergarten through the gateway to college, and should provide the basis for an effective accountability system, to be incorporated in the reauthorization of NCLB or its successor.

The plan would not provide the grade-by-grade and student-by-student data needed to hold the teachers at all grade levels, from grade 3 to grade 8, responsible for student progress in literacy, in the way that NCLB does. Nor would it provide the data to support a value added approach to teacher accountability, as many have proposed. As we said above, other nations have managed to produce much higher levels of student achievement without either national or state grade by grade testing or value added systems of accountability. Most take the view that teachers and schools should have more freedom to determine the enacted curriculum on a year by year basis and it is sufficient, in their view, to set benchmarks every few years to make sure that students, particularly vulnerable students, are not falling behind. We believe the burden should be on those

who think otherwise to produce empirical data and analysis justifying their position as well as the costs associated with maintaining such a system relative to the purported benefits.

The idea of having the National Examinations Board agree on a nominal core curriculum, in combination with the idea of the use of board examination systems at the high school level to implement the core curriculum and the specification of a pass standard for the Qualification Exams at a true college ready standard, are all, taken together, designed to prevent the kind of narrowing of the curriculum attributed to NCLB.

The fear of federal control of the curriculum is still very real. Indeed, in some states, there is resistance to the idea of state control of the curriculum. This plan addresses these issues directly. First, the proposed National Examinations Board, in which the national policy decisions would be lodged, would be under the control of the states, not the federal government. Second, though the Board would decide which subjects are included in the core curriculum, and would provide an indicative curriculum framework for that curriculum, and resources for assessment for those subjects, states would be free to ignore the framework, the curriculum and the assessment resources in all of those subjects in grades K-8. At the high school level, the states would have to administer their own choice of board examinations in the subjects in the core curriculum, chosen from a list approved by a Board of their choosing. And high schools and parents and students would themselves be able to choose from among the board examinations offered by the state. Thus there would be choice of both curricula and examinations at every jurisdictional level of the system.

The use of the board examination system at the high school level would greatly change the nature of summative assessment in the United States. By moving from a system of state accountability testing based largely on the use of multiple choice, computer scored tests to one that emphasizes performances that require time and effort and could include multiple components, such as extended essays, designs, technology- or media-based artifacts, it becomes possible to assess a much broader range of skills and knowledge, especially higher order thinking, critical thinking, the ability to deal with unanticipated complex problems, creative and innovative activity and other skills and abilities crucial to the capacity of this country to maintain the standard of living of the nation as a whole and the individuals in it.

The emphasis in this plan on the role of the National Examinations Board in producing resources for formative assessment is important. Our aim, as we have said more than once in this document, must be to devise a system of assessment the primary purpose of which is to improve student performance. That cannot be done unless the instruction the student is getting at any given moment is matched to the needs of that student, day by day. It is well established that students make the most progress when the instruction they get is challenging but not overwhelming. To accomplish that feat for all the students in a class, the teacher must know a great deal about where the student is in relation to the standards, day-by-day, topic-by-topic. The first key to this state of affairs is having a clear curriculum framework, so both teacher and student knows what work looks like that

is meeting the standards as the student progresses through the curriculum. The second is having good data on where the student is with respect to that standard. That is why we have strongly recommended that the National Examinations Board work to produce these curriculum frameworks for the whole core curriculum and to produce also assessment resources that teachers can use to make accurate judgments as to where their students are relative to the framework.

This is especially important because of the imperative, advanced above, to provide more time and more assistance to students who need it, as soon as those students start to fall behind. That cannot be done unless those students are identified, and the assistance they need cannot be defined until the teacher knows what they are behind on and how far they are behind. This demands diagnostic assessment tools that differ markedly from the conventional tests and assessments we have now but that we are nonetheless capable of producing by drawing upon and further investing in contemporary research on the integration of curriculum, instruction and assessment.

Some readers will be disappointed or frustrated by our reluctance to offer a system that would produce numbers that, by themselves, could be used to enforce an accountability regime or be used as the basis of the compensation of teachers and school leaders. But we have always looked askance at such systems. One would not want to punish a school with a long record of low performance if an energetic, determined new principal with a good plan had just taken over. Nor would one want to shut down a school that had suffered a sharp decline in student performance if it turned out that that decline is the result of a wholesale change in the character of the residents of the school's catchment area. Nor would one want to lower the boom on a school if some sound advice from a wise counselor would have enabled the school to have turned the corner. This sort of thing is what school inspections are for. When they are conducted in a competent way, they can often avert more drastic measures. When they are triggered by examination of school data that is designed to identify schools in difficulty, they can be made to be much more efficient than if they are regularly scheduled for all schools on the same basis.

This is not an argument for a listless accountability regime. Quite the contrary. An inspection regime that is rigorous can be used to identify schools that have failed to the point that their head needs to be replaced, or key faculty need to be replaced or the whole school taken over by another management and staff. But better judgments are likely to be made about such things by a trained competent staff of experienced educators than on the basis of a formula driven by test data alone. The data can certainly be used to indicate a need for an inspection and it can be used to track progress after an inspection visit, but, in the end, only experienced professionals very familiar with the situation are in a position to make the kinds of judgments needed in such situations.

Lastly, we return to the conviction, shared by many, that we are on the edge of major advances in technologies that can support learning, technologies that could in fact produce a revolution in the organization and locus of American education. This plan invites those advances. For generations, board examination systems have been pencil and paper affairs, though more recently, the exams have been scanned and sent to scoring

centers electronically. But there is nothing inherent in the board examination system idea to prevent the resources for that curriculum from being limited only by the imagination of educators and software developers and wholly accessible through the World Wide Web, nor is there any reason why the exams themselves cannot be delivered by the web, as long as the proper measures are taken to keep the examinations secure and to be sure that it is the student taking the examination and not someone else.

The opportunity to design a new assessment system for the United States is really the opportunity to reconceive the nation's instructional systems, and, therefore, to leapfrog from our very low standing relative to the other industrial nations to the head of the pack. We cannot do that by concentrating on assessment systems as if they were disembodied from the rest of our system. The best way to take advantage of this extraordinary moment is to think hard about what sort of education system we want.



Jon S. Twing, PhD
Executive Vice President
Assessment & Information group of Pearson

Pearson
2510 North Dodge Street
Iowa City IA 52245 USA
Telephone: 319 339 6407
Fax: 319 358 4224
jon.s.twing@pearson.com

December 2, 2009

Office of Elementary and Secondary Education
Attention: Race to the Top Assessment Program—Public Input Meetings
U.S. Department of Education
400 Maryland Avenue SW
Room 3E108
Washington, DC 20202

To Whom It May Concern:

In many ways the future of assessment has nothing to do with the assessments themselves. If we are to achieve the common goals of education reform (improved learning, increased college readiness and true international competitiveness), we must design a **learning system** that uses assessment data as one component of a much broader and comprehensive information management model. Such a learning system must start with the premise that our fundamental objective is to facilitate personalized instruction and early interventions so that we prepare each student to compete in a global economy and thrive in a global society. This new student-centered learning system must use technology to reduce the burden on educators, students, parents, and the public. It must facilitate the flow of information for timely instructional interventions and continuous improvement to remove current barriers to student success.

Far too many American students are dropping out of high school or arriving at college needing remediation. Although standards-based reforms and No Child Left Behind (NCLB) have brought much needed heat and light to closing the achievement gap, what happens in schools and classrooms across the country still too often remains unknown. And, too often, we fail to provide timely interventions to help struggling students or advanced instruction for students who are ready for new challenges. The US needs better insight into effective teaching and learning to accelerate the replication of best practices. We need a transparent educational quality management system that facilitates meaningful comparisons across states and internationally.

Successfully transforming our education system will require the full integration of early childhood, elementary, middle, and high school education with college and workplace readiness. This new learning system must be designed to help students make more seamless transitions by routinely providing personalized feedback. Only in this way can education target the needs of individual students so that they learn the skills to master content, think critically, engage creatively in collaboration and problem solving, and become successful life-long learners.



Such a student-centered learning system must also explicitly be designed to build teacher capacity. By collecting and tracking teacher preparation and performance data, current challenges related to developing effective teachers and school leaders can be more directly addressed. Using technology to manage all aspects of education information, any gaps in teacher preparation and the needs of instruction can be documented. The system could then be used to plan for improvement through professional development. Improvement can be documented at an individual level showing outcomes and results for the teachers and other aspects of the system, such as various student measures, motivation and engagement.

To successfully transition to a student-centered learning system, we must first provide a stable and reliable bridge from our current context to the requirements of our future state. For example, we have invested large amounts of money and time into state accountability systems for NCLB. At the core of this current system are large-scale, paper-based summative assessments that predominately use traditional multiple-choice items. We will not be able to immediately replace this system with a new one. The bridge assessment system will need to incorporate both old and new attributes as we move forward toward truly world class standards. Some of these attributes should include the following:

- **Fair, legally defensible assessments** for all students, affording the opportunity to show what students know and can perform (i.e., reliable, valid, and fair)
- Truthful **indicators of college and workplace readiness** that may influence policy and accountability decisions for high school graduation, college placement, and college admission
- **Flexibility to incorporate on-demand assessments** to inform both high- and low-stakes decisions for improved instruction and system accountability
- **Security** so the integrity of not only the assessments, but the information and the outcomes, will be recognized as viable and worthwhile
- A **technology plan** that anticipates changing technology, supports open industry standards for interoperability, and facilitates the synchronization of local, state, and federal databases
- **Timely and accurate information** for teachers, administrators, counselors, parents, students, the public, and other stakeholders
- Efficient **information collection system** so traditional enrollment data, course records, and student growth trajectories are available for decision-making related to matriculation
- Ability to **differentiate personal information from information needed for instruction** (that might be scored locally or subjectively) **and from high stakes information** (that might be scored objectively even if not multiple choice)

This approach is both pragmatic and innovative. Diverse education stakeholders—state consortia, the US Department of Education, non-profit and for-profit entities, K-12 and higher education leaders, and assessment developers to name a few—will be required to collaborate, cooperate, and compromise in new ways. From our experience working in more than 30 states, Pearson has an understanding of the depth and breadth required to develop and implement such a system. Pearson currently works with all constituents and stakeholders in education across all levels and needs.



The breadth of Pearson support for large-scale assessment programs includes the following:

- Processing the National Assessment of Educational Progress (NAEP)
- Fulfillment of the ACT and SAT assessments and ACCUPLACER
- Scoring both the ACT and SAT essays in a distributed model using a network of more than 40,000 trained and qualified scorers, approximately 40 percent of whom are current or former teachers
- Scoring the Collegiate Learning Assessment (CLA) for the Council for Aid to Education (CAE) using automated essay-scoring technology
- Providing fulfillment services for the Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS) domestically
- Providing the Pearson Test of English to more than 800 colleges in the United States
- Providing teacher certification for more than half of the teacher education programs in the US
- Providing master teacher certification through our service to the National Board for Professional Teaching Standards
- Developing and implementing the American Diploma Project assessments for Algebra I and II in partnership with Achieve and a 15-state state consortium

Drawing on this experience and expertise, Pearson is responding to the Department's questions regarding assessment design and the future needs of a comprehensive learning system that incorporates international benchmarks, encourages problem-solving and critical thinking, and prepares students for college and the workplace. In the pages that follow you will find outlines of our vision based on your recent request for input in the pre-publication guidance on the Race to the Top Assessment Program. While it is too early to list and outline the specific requirements of this system, we have tried to outline the parameters to consider as we move toward an assessment system design. Similarly, the cost estimates included in Pearson's response are based on the assumptions and design parameters outlined, and are therefore subject to change as the specific requirements of the system are established.

If you have any questions or would like to discuss our response further, please contact me at 319-339-6407 or by email at jon.s.twing@pearson.com or my colleague Shilpi Niyogi, Vice President, National Services, at 202-434-0975 or by email at shilpi.niyogi@pearson.com.

Sincerely,

A handwritten signature in black ink, appearing to read "Jon S. Twing".

Jon S. Twing, PhD
Executive Vice President
Assessment & Information Group of Pearson

Table of Contents

General Assessment Questions	2
Specific Technical Assessment Questions	25
High School Assessment.....	31
Assessment of English Language Learners	35
Assessment of Students With Disabilities	39
Technology & Innovation in Assessment.....	44
Project Management.....	52
References.....	56
Appendix A: Definitions	58
Appendix B: Description of Performance-Based Items	59
Appendix C: Cost Estimates	68

General Assessment Questions

Question

1. Propose an assessment system (that is, a series of one or more assessments) that you would recommend and that meets the general requirements and required characteristics described in this notice. Describe how this assessment system would address the tensions or tradeoffs in meeting all of the general requirements and required characteristics. Describe the strengths and limitations of your recommended system, including the extent to which it is able to validly meet each of the requirements described in this notice. Where possible, provide specific illustrative examples.

Response

A Bridge to a Fully Integrated Student-Centered Online Learning System

To accomplish the ultimate goals of educational reform, we must begin by thinking of assessments as a critical component of a fully integrated, student-centered learning system. Such a learning system would incorporate assessments that draw on advances in cognitive science, psychometrics, technology, and effective instructional practices from across the US and around the world. The design of these new assessments would be innovative, online, integrated, and flexible. The biggest barrier to cost effectively implementing large-scale assessment innovations is the current paper-based system most states use.

As state consortia transition from the current paper-based system to a fully integrated online system, key considerations include the following:

- What innovations are possible, practical, and helpful?
- What methods are evidence-based and defensible?
- What drives costs?
- What are the trade-offs when making different choices?
- What timeframes are reasonable for implementation?

The American Recovery and Reinvestment Act (ARRA) and the Race to the Top (RTTT) present an unprecedented opportunity for states to establish the infrastructure and capacity for online assessments, online management systems, online content and instructional delivery and reporting systems. By moving to a technology-based assessment delivery platform we can accomplish the following:

- Facilitate wider use of performance-based tasks economically and reliably. Students can demonstrate their knowledge and skills through open-ended written responses, multi-step problems and inquiry-based investigations through simulations and interactive item formats, not just multiple choice responses
- Use new language evaluation technologies to automate the scoring of open-ended oral and written responses. These technologies are in practical use today in higher education and professional certification assessments
- Simplify the test administration process

- Speed the reporting of student results
- Improve the efficiency of the entire learning system
- Facilitate direct links to instruction

Technology-based assessments better reflects the world students live in outside of school today and the world of college and work they will live in after high school graduation.

By improving usability and speeding delivery of results, technology-based assessments can better integrate assessment data into longitudinal data systems and student information systems. This facilitates the integration of benchmark assessment and summative assessment and makes the data more useful to teachers and educators to inform instruction and improve decision-making. Technology-based assessments reduce cumbersome processes and the carbon footprint resulting from paper-and-pencil testing.

Preserving and Enhancing a Quality Management System for Public Schools

Both the federal government and the States have invested a significant amount of time and money in existing No Child Left Behind (NCLB) assessments and accountability systems. Therefore, we propose a bridging strategy that will allow us to first move from today's primarily paper-based, multiple-choice summative assessments to significantly enhanced summative assessments (online assessments emphasizing new item formats, such as performance-based tasks, but still given at the end of the year, primarily for accountability purposes) and ultimately to an integrated student-centered learning system with on-demand assessments throughout the school year that are linked with instruction and provide information for continuous teacher capacity building and professional development.

This bridging strategy recognizes two core functions of educational accountability that must be preserved in the transition:

1. **Provide individual student achievement data.** Annual assessment of student achievement is the foundation for a quality management system for public education. This is how we know what progress we are making in providing all children equal access to a quality education. Parents need this information to make appropriate choices about their children's education. Educators need this information to understand how their curriculum and instructional practices are working, what is effective, and what needs to improve. Policymakers and school administrators require this information to understand effectiveness of our public schools and their impact on students, especially traditionally underserved populations. Annual assessment data should be publicly available and disaggregated by subgroups to create both "heat and light" about what's happening in schools and who is accountable for results. Parents should have access to a universal annual Report Card on their children's school performance.

2. **Support meaningful comparisons.** Accountability systems should be based on objective information to support meaningful comparisons across schools, states and internationally. While student test scores alone may not drive accountability systems, and multiple measures should be used, the data elements of an accountability system must be fair, valid, and reliable measures to facilitate comparisons. Therefore, we must take rigorous steps to support verifiable and objective information from all measures used in the accountability system.

For example, if used in an accountability index, graduation rates and attendance need to be calculated using a standard formula and school climate should be measured using a standardized survey instrument. The data and calculations such as growth projections used for accountability should be transparent, replicable and audited (no proprietary “black boxes”). If teacher-generated data is used in addition to standardized test scores (for example, course grades or locally scored portfolios), quality management systems should be in place such as standardized rubrics, required teacher training, and periodic auditing of results. Since the focus of public education is to prepare students for college and careers, the data used in school accountability systems should focus on student outcomes.

Our motivation for proposing a bridging strategy is to first prioritize transition to an online platform with significantly enhanced summative assessments as the primary assessment of record. This assessment would most likely start as a fixed-form, online criterion-referenced test administered at the end of the year, but would become more flexible and adaptive as we move into the future. This bridge assessment would initially be comprised of English language arts (ELA) and mathematics for grades 3–8 and a series of end-of-course assessments for middle/high school. The content domain for the bridge assessment would be defined by the common standards, and operationally defined through specific item and test specifications. The assessment, however, would have a greater focus on performance-based assessment and problem-solving, with items and tasks that—thanks to ongoing research and innovation in both item design and technology—produce information regarding students’ academic knowledge and skills that until now have been difficult to assess with traditional multiple-choice tests.

The timeline for the development of the bridge assessment system anticipates that the RTTT Assessment grant awards are made in the fall of 2010, and test development begins in 2011 with field testing in 2012, followed by a full census field test in all participating states in 2013, and operational testing starting in 2014.

Strengths of Recommended Assessment System

The proposed assessment system meets all of the Department’s general requirements for the development of summative assessments:

- Individual student achievement as measured against standards that builds toward college and career readiness by the time of high school completion
 - The proposed assessment system will be aligned with a common set of K–12 standards that are internationally benchmarked and that build toward college and career readiness by the time of high school completion.
 - As described in the **Specific Technical Assessment Questions** section of this response, the proposed assessment system will be vertically aligned so that students, parents, and educators know if students are on-track toward meeting college-and career-ready standards by graduation regardless of their current enrollment status.

- Individual student growth (the change in student achievement data for an individual student between two or more points in time). Annual assessment of student achievement is a fundamental component of both the proposed bridge assessment and the vision for a fully integrated, student centered learning system.

Additional detail is also provided in subsequent sections to confirm that the proposed assessment system meets the Department's required characteristics (e.g., accessible to the broadest range of students, contains varied and unpredictable item types, produces results that can be aggregated, etc.).

Tensions Within the Recommended Assessment System

There are a number of inevitable tensions among the alternative approaches that might be taken with a common core assessment system. A primary attribute of our recommended system is the introduction of end-of-course assessments at the high school level. We address the reasons for this recommendation later in this document but recognize that an end-of-course system has significant implications for instructional practice and opportunity to learn. In addition, the ability to measure and track student growth is significantly complicated by the introduction of an end-of-course system. Nevertheless, we believe this approach best serves the goals of common assessments because of its direct link to instruction and college readiness and current research regarding the virtues of rigorous course selection.

An additional tension resulting from our proposed bridging strategy is the initial use of online, fixed-form assessments. We believe that an adaptive approach to an online common assessment is both desirable and feasible. However, adaptive testing requires a significant inventory of appropriately targeted test content. In addition, adaptive testing as it is commonly implemented today measures only a subset of what is important to test. The computer has significant capacity to personalize content and increase the precision of measurement for the individual student, but an adaptive testing system that can deliver only discrete multiple-choice items will not serve the future needs of a common assessment. Delaying the implementation of adaptive testing through the bridging strategy permits the development of both a sufficiently large test content inventory and the refinement of psychometric models that will be appropriately aligned to the purposes of common assessments.

Defensibility of the Recommended Assessment System

Anything is possible regarding changes to assessments. A successful change will involve well-articulated, research-based evidence that show the system produces valid scores. For example, one of the stated goals is college placement. For the system to be seen as fair and useful we will need to prove beyond a reasonable doubt that accurate and valid placement decisions can be made using the information provided by the system. Otherwise key education stakeholders, not just skeptics or critics, will be reluctant to embrace the new system. Therefore, the success of this system depends upon paramount research-based evidence of the successful uses of the resulting information for the purposes stated and the design of such evidence collection into the system from the beginning. This research-based evidence is even more critical given the varied purposes of the system, such as high school graduation, capacity building, and professional development.

Question

2. For each assessment proposed in response to question 1), describe the:
 - Optimal design, including:

Response to the Race to the Top Assessment Program Request for Input

- Type (e.g., norm-referenced, criterion-referenced, adaptive, other);
- Frequency, length, and timing of assessment administrations (including a consideration of the value of student, teacher, and administrative time);
- Format, item-type specifications (including the pros and cons of using different types of items for different purposes), and mode of administration;
- Whether and how the above answers might differ for different grade levels and content areas;

Response

Optimally new common assessments would be designed to be online, innovative, integrated, and flexible. As previously noted, because legacy NCLB assessments and accountability systems are in place, we have proposed a bridging strategy that will allow us to move first from the current primarily paper-based, multiple-choice summative assessments to significantly enhanced online, performance-based summative assessments and ultimately to a fully integrated learning system with assessments on-demand throughout the school year that are linked with instruction and provide information for continuous teacher capacity building.

As part of this bridging strategy we are proposing the development of an assessment system that is composed initially of significantly enhanced online assessments. There are many unanswered questions that will impact the optimal design of each component of the system. The optimal design of each component will be contingent on several considerations, including the following:

- Resolving the conflicts between the stated purposes and intended uses of the assessments
- Standards
- Reporting specifications
- Technology
- Size of the item or test pool and security
- Fairness and defensibility of the system

Pool size will itself be a function of several considerations, including the number of students, granularity of the information to be reported, item exposure criteria, release policies, and information that is valid for use in the classroom and elsewhere.

Despite these unknowns, we envision the bridge assessment to be a summative assessment administered at the end of the year, similar in implementation to the current NCLB assessments. This assessment would most likely start as a fixed-form, online criterion-referenced test but would become more flexible and adaptive as we move into the future. This bridge assessment would initially be comprised of ELA and mathematics for grades 3–8 and a series of end-of-course assessments for middle/high school. The content domain would be defined by the common standards, as operationally derived through specific item and test specifications. The assessment, however, would have a greater focus on performance-based assessment and problem-solving, with items and tasks that—thanks to ongoing research and innovation in both item design and technology—produce information regarding students’ academic knowledge and skills that until now have been difficult to assess with traditional multiple-choice-driven tests.

The common standards work currently underway is a key factor in the design of the bridge assessment, which must comply with NCLB and anticipate the most salient elements of the emerging system. Some of the decisions that will need to be made include the following:

- Which objectives/learning expectations should be part of a summative assessment and which are best assessed at the classroom level or in some interim assessment component?
- What are the opportunities to learn issues and how much time is needed to prepare teachers and students?
- What are the content parameters for each objective/learning expectation?
- What item type(s) best assess the content?
- What is the cognitive demand of each objective/learning expectation?
- What information is needed to assist teachers and other stakeholders (i.e., reporting categories)?

Ideally, the assessment objectives refined through answering these questions will transform into item and test development specifications. Practice has shown that trying out exemplar items may help produce stronger, more clearly articulated assessment standards—standards that inherently allow for coherent measures. Our bridging approach should appropriately allow for the exploration of these questions and help inform the longer term strategy for implementing a fully integrated learning system.

Using Online, Performance-Based Items to Enhance Assessments

Because technology is constantly changing, this new series of assessments should be based on the new opportunities presented as technological innovation makes possible new, more effective ways to assess student knowledge and skills across the curriculum and to capture, store, and use this information effectively. Some of these items would use current technology, such as drag-and-drop and hot-spots. Others would be more traditional constructed response items (requiring automated and/or human scoring), and yet others may be innovative item types (e.g., simulations) that are currently being researched and developed.

Multiple-choice items, where students select a response from those provided, have been well-suited to a paper-based large-scale assessment and provide the following advantages:

- Ability to gather a significant amount of information in a short period of time
- Can be scored quickly, easily, and objectively
- Have widely accepted development guidelines and psychometric models
- Provide effective measures of lower-level thinking skills

With advances in technology, it is possible to enhance multiple-choice items to include audio and/or video streaming and other interactive delivery features. However, for measures of college and workplace readiness, inquiry, problem solving, and critical thinking, performance-based responses need to be part of the overall assessment system design. No longer can we rely on multiple-choice assessment items where students may get the right answer for the wrong reason and where the appropriate skill may not be assessed.

Performance-based items commonly refer to test questions or tasks where students determine and create their own response. The task may consist of a single or multi-part response. These tasks are machine-scored, scored using an automated scoring engine, or are human-scored (supported via a technology-based management system) using a rubric. The human-scored tasks comprise the more traditional means to assess analysis, critical reasoning, and other aspects of higher level cognitive skills. Their primary advantage is the ability to see and evaluate students' thought processes and approaches to questions. Their primary disadvantage is the amount of time and resources required for scoring. Currently most scoring is done by humans, although automated scoring is gaining greater momentum.

The following table summarizes performance-based item types and indicates which of the following scoring options are appropriate. These scoring options are defined as follows:

- **Machine scoring.** Simple scoring rubrics are applied through fixed rules automatically by computer, as is currently done for multiple choice items.
- **Automated scoring.** Adaptive algorithms that require human-generated training sets are applied through dynamic rules automatically by computer, as is the case for automated essay evaluation.
- **Human scoring.** Complex scoring rubrics require trained teachers or other qualified scorers.

Performance-Based Item Types and Scoring Methods	
Student Response/Item Type	Scoring Method
Constrained Response <ul style="list-style-type: none"> • Drag-and-drop one or more elements • Select one or more elements • Mark one or more locations (“Hot spots”) 	Machine
Constructed Response <ul style="list-style-type: none"> • Written text (e.g., essay, short answer) • Graphing • Equation/formula construction 	Human readers and/or automated scoring
Simulations <ul style="list-style-type: none"> • Immersive, interactive problems • Multi-step problems • Outcome based responses 	Machine, human readers, and/or automated scoring

Scoring of Performance-Based Items. Performance-based items are grouped in the categories above and require different scoring methods appropriate for the different types of performance-based items.

We provide additional descriptions of performance-based item types, including sample items, in **Appendix B**.

Proposed Mathematics Test Design

Below is a possible test design for the grade 3–8 mathematics test as part of the bridge system. This design was also used to inform the development, maintenance, and administration costs in the **Appendix C** of this response. This test would be administered online and use technology in the design of the items. The presentation of information would not be bound by the paper world but would include innovative aspects through technology—videos, simulations, and other multi-media.

Response to the Race to the Top Assessment Program Request for Input

In our proposed test design, a large percentage of the score, approximately half would come from performance-based items. These items will contribute a range of scores to the total point value. Hot spot items may only be worth one point. Others may be a multi-part item where the parts can be either independent or dependent and worth multiple points.

Proposed Mathematics Test Design						
Grade	MC Items	MC Points	Performance-based Items	Performance-based Points	Total # Items	Total Points
3	36	36	12	30	48	66
4	36	36	12	30	48	66
5	39	39	17	40	56	79
6	39	39	17	40	56	79
7	42	42	18	42	60	84
8	42	42	18	42	60	84

Mathematics Test Design Option. This design was used as a baseline for the learning system bridge.

Mathematics performance-based items may require students to graph or solve an equation. Prior to testing students have an opportunity to practice using the online tools. In the following figure the student is prompted to use the online tools to write an equation. Mathematics constructed response items using the graphing and equation tools are currently scored by human scorers, however automated scoring of these item types may soon be available.

Try duplicating the number sentence $(x^2 + 3x + 2) \left(\frac{x^2 + 1}{x + 2} \right) = x$ using buttons from the menus and the keyboard. You can also practice using the menus and templates.

WWLastName, WWFirstName | Gr 3 | Section 2 | Question 10 of 13 | Section Review | Previous | Next

Sample Equation Constructed Response. In this student practice item in algebra, students are instructed to duplicate an equation using the buttons from the menus and the keyboard.

Proposed ELA Test Design

A similar design for an ELA bridge assessment using technology to enhance the ways we can assess the curriculum, is shown the following figure. The assessment would be composed of at least a reading and a writing section.

The writing component would have at least an editing section and an extended-response prompt scored on multiple traits. The writing prompt could be a traditional prompt at the lower grades, with a prompt based on a reading selection at the middle grades to multiple selections at high school that contributes to both the reading and writing components. Since this is a bridge, the ultimate ELA response would likely include multiple data sources, could include collaboration and would likely integrate a series of tasks to actually drive students to produce outcomes.

Proposed ELA Test Design							
Grade	MC Items	MC Points	Writing Prompt	Performance-based Items	Performance-based Points	Total # Items	Total Points
3	30	30	1	11	29	42	59
4	30	30	1	11	29	42	59
5	36	36	1	13	37	50	73
6	36	36	1	13	37	50	73
7	42	42	2	14	50	58	92
8	42	42	2	14	50	58	92

ELA Test Design Option. We propose this ELA test design as a way to use technology to enhance how curriculum is assessed.

Below is an example of a practice reading item that uses drag and drop functionality that can be delivered online and machine scored.

Reading: Re-order paragraphs

Instructions

For this question type you need to put the text in the correct order by selecting text boxes and dragging them across the screen.

There are two ways you can move the text:

1. Left-click on a box to select it (it will be outlined in blue), hold the left mouse button down and drag it to the desired location.
2. Left-click on a box to select it, and then left-click on the left or right arrow buttons to move it across. On the right panel, you can also use the up and down arrow buttons to re-order the boxes.

To deselect a box, left-click elsewhere on the screen.

Candidate Name

The text boxes in the left panel have been placed in a random order. Restore the original order by dragging the text boxes from the left panel to the right panel.

Source

He convinced Professor Fitzgerald of the University of Hull to set up a study into this matter.

Professor Fitzgerald and his team studied more than 47,000 women.

The women were asked to fill in a questionnaire about their diet and about their suffering from acne.

No link was found between acne and traditionally suspect food such as chocolate and chips.

Doctor Byron has long held that there is a link between diet and acne.

Target

Doctor Byron has long held that there is a link between diet and acne.

Next

Sample Drag and Drop Reading Item. This is an example of a performance-based reading item that is delivered online and can be machine-scored.

Response to the Race to the Top Assessment Program Request for Input

As with mathematics, the innovative ELA items will contribute a range of scores to the total point value. In this design the writing prompt is scored on multiple writing traits for all grades and starting at grade five a reading score is also applied.

Proposed EOC Test Design

One option for end-of-course assessments is provided in the following figure. We suggest beginning with 10 end-of-course assessments: Algebra I, Geometry, Algebra II, English I, English II, English III, Earth Science, Biology, Physics, and US History, but clearly this number will be driven by the common core standards and state policies. These tests would also be administered online and use technology to design items similarly to the 3-8 assessments. As with the grade 3–8 assessments, approximately half of the score would come from performance-based tasks. We recognize that a varying number of end-of-course assessments may ultimately be desirable and for this reason in **Appendix C** we estimate development costs on a per-test/per-subject basis.

Proposed End-of-Course Test Design							
Subject	MC Items	MC Points	Writing Prompt	Performance-based Items	Performance-based Points	Total # Items	Total Points
Math or Science	42	42	N/A	18	42	60	84
ELA or US History	42	42	2	14	50	58	92

Proposed End-of-Course Test Design. We propose the test design above for end-of-course assessments at the high school level.

When considering the length of any assessment, there are many considerations such as students' attention span, teacher and other stakeholder's need for information, and the time taken from instruction. These should not be taken lightly and often require tradeoffs. As we look to the future, the assessment should become an integral and non-intrusive part of the instructional cycle. For the interim, the bridge assessment would be a single assessment likely given in multiple sections over a series of days. With the emphasis on problem solving and college and workplace readiness, the time commitment for the assessments would be greater because more time is needed to process and think critically. The trade offs between security, testing time, and level of effort need to be considered. However, for the purpose of this response we have assumed the proposed design would likely require from 2-3 hours per content area for the lower grades and the end-of-course assessments.

The bridge assessment is only a small piece of the total solution and only an interim piece at that. With the need for accountability and high-stakes decisions such as college and career readiness, the need for a summative component will still exist. As such, our application will change from a single point in time to a more cumulative application with on-demand testing at multiple points in time, tailored to learning objectives throughout the year. Our system must also be agile enough to embrace new technologies and new requirements in the future.

We believe that assessment and instruction must be closely linked and envision a truly integrated system with each feature informing the other. Too often, the end-of-year test results are not used to inform instruction. As we look to the future, we envision a large and growing pool of items and activities that support the learning progressions encompassed within the common standards. Teachers could use these items at any time for formative, diagnostic, or summative purposes, gathering and documenting information about student academic progress.

Testing Across Different Grade Levels and Content Areas

Before discussing how tests should differ across grade levels in terms of format, item type, content, and other such considerations, it is worth noting some ways in which they should not differ. For example, tests at all grade levels should make full use of clarification/accessibility protocols such as universal design principles in test construction within any given grade level across all students regardless of gender, ethnicity, disability status, and other such key demographic variables (Dolan and Hall, 2001; Dolan et al., 2006; Ketterlin-Geller, 2005; Thompson, Johnstone and Thurlow, 2002).

Additionally, assessments at every level also need to span the full range of content and difficulty that is appropriate for that level and supported by the common core standards. The capability to track student progress across time depends to a significant degree on the fidelity with which we assess student capabilities thus making appropriate breadth and depth of content and difficulty necessary prerequisites for each assessment level.

Finally, as with all assessments, it will continue to be of crucial importance that the new assessments adhere to existing professional standards governing all aspects of test development, administration, reporting, and use, as described in such industry guidance documentation as the joint *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999). This adherence must be specific and evidence-based to defend the integrity of the resulting assessments. However, given the multiple purposes of improved learning, college and career readiness, measures of problem solving and critical thinking, high school graduation, college placement and admissions, the validity and fairness of these assessments (or this system) will need unprecedented evidence and defensibility.

In terms of differences between the common core assessments across levels, the proposed bridge assessments recognize increases in the numbers of questions and the anticipated time that students will take as they progress from the early grades through high school. In addition, the ELA test design recognizes an increasing emphasis on writing skills in the higher level assessments. In general, we expect these differences will reflect the progressive and cumulative nature of the common core standards.

Question

- Administration, scoring, and interpretation of any open-ended item types, including methods for ensuring consistency in teacher scoring;

Response

Test Administration

The vision for a seamless summative assessment is online test administration to promote tests that have the following features:

- Interactively engage students while being relevant to the way they learn
- Include more performance-based and innovative test items that assess problem solving and critical thinking within core subjects and in interdisciplinary contexts
- Can be administered, scored, and reported in a timely manner, so data can inform teaching and learning

Human Scoring of Performance-Based Items

Assessment reform is dependent on designing a system of testing and learning fundamentally structured around teacher participation, professional development, and assessment literacy. Teachers must be the backbone of the reform effort, and assessment design must include teachers in item development, review, and scoring. The system should be transparent and garner local involvement and buy-in, while achieving national scale and consistency.

Past barriers to teacher participation in the assessment process have included cost, schedule, logistics, time away from the classroom (causing a burden to districts that do invest in teacher scoring), and concerns about bias or inconsistency. However, these barriers can be addressed by innovative technologies.

Key components of a successful teacher scoring model should include the following:

- **Teaching experience requirements.** All scorers should have teaching experience relevant to the task they score. This requirement will foster credibility of scoring and maximize opportunities for teachers to gain insight into the testing and scoring process. Any scoring model that does not include teacher scoring may fulfill part of the scoring mission (application of scores on student responses) but will fall short of the full vision of teacher involvement, buy-in, accountability, and development. Therefore, teacher recruitment must be online, scalable, and easy for teachers and administrators to access and use.
- **Anywhere scoring.** The scoring system must be accessible nationwide and support teachers from any state or locale, including both urban and rural areas. The system must not create an advantage or disadvantage for one group of teachers over another due to location, infrastructure, or other technology barriers. For this reason, a web-based distributed scoring model for teachers is highly recommended, wherein a teacher with a very basic computer and Internet access can participate in scoring from home or school anywhere in the country. Distributed scoring is an operational model used by Pearson for scoring high-stakes assessments today, including college entrance exams and state assessments. Distributed scoring eliminates cost, schedule, and logistical barriers associated with “bricks and mortar” scoring wherein teachers have to travel to scoring centers to participate in scoring large-scale operational assessments.

- **System requirements.** The scoring system should support online responses, and be able to randomly and anonymously route student work to teachers to score. The ideal system should house student response data separately from student demographic data, so teachers have no indication of the student's identifying information, including class, school, gender, ethnicity, or other data that could bias scoring. The system should accommodate innovative student response formats, including inquiry, simulation, audio, video, and portfolio scoring. The system should also include standard security protocols, such as Secure Socket Layer (SSL) protection.
- **Online, interactive training.** To adequately train teachers to score, the training platform must support "anywhere training" in a consistent and predictable manner. The training must be online and asynchronous, so teachers can complete training on a flexible schedule while still facilitating an identical training curriculum completed by each trainee. The training should include rubrics and protocols designed to measure student performance against the common standards, along with samples of student work, so teachers better understand the successes and difficulties students have with the performance tasks and thus the standards. Scorer training can be expanded into a platform for professional development with direct and positive implications for the materials teachers develop for the classroom.
- **Professional development credit.** To encourage and reward participation, it is recommended that teacher training and scoring be accompanied by continuing education units (CEUs) to be granted to teachers for successful training, qualification, and scoring. Requiring teachers to complete scoring as well as training is critical to make certain that teachers have not only been trained on the standards, but also have seen and evaluated a range of responses to the items and prompts. Scoring expands the walls of the classroom, giving teachers a broader view of student performance, beyond their class and school to provide a rich context for understanding common standards and how students can perform against those standards.
- **Role of local educational agencies (LEAs).** LEAs can play a critical role in supporting teacher scoring by supporting teacher recruitment and supporting the CEU development and implementation process.
- **Schedule flexibility.** The system should be flexible to support teachers scoring from home or their schools, with daytime, evening and weekend hours.
- **Social networking tools.** Teachers must be supported with robust online tools for messaging, knowledge management, and information sharing, including interaction with scoring experts.

Methods for Consistency

For summative assessments, the scoring system should not allow a teacher to score his or her own students, but rather a randomized selection of student work. This eliminates the possibility of local bias and, equally important, broadens the range of student work the teacher sees and evaluates during the scoring process.

The training program should include item-specific training and qualification. This is an industry standard to make sure student work is only assessed by qualified scorers who have passed a scoring test and demonstrated that they can successfully and predictably adhere to common standards for scoring.

The scoring system should further include automation and reporting tools to track the performance of every scorer throughout the project. This promotes continued adherence to scoring standards. The system should also have the capability to lock scorers out of the system and/or reset their work if they fail to meet project standards.

Finally, the management system supporting scoring operations should have an industry recognized quality certification, for example, International Organization for Standardization (ISO) certification, to promote process rigor and consistency. Accurate, reliable, and timely scoring can only be achieved with process maturity governing the scoring life cycle.

Additional Professional Development Considerations

An integrated professional development program fostering assessment literacy should provide teachers an opportunity to participate in different activities that illustrate how tests are developed, built, and validated, as well as scored. A full professional development curriculum on assessment literacy could include the following components:

- **Standards**—Understanding the common assessment standards and the objectives connected to each standard
- **Item Writing**—Training educators on the criteria for high quality test items and involving educators in the item writing process
- **Item Review** (content review, bias and sensitivity review, data review, standard setting, etc.)—Training educators on and facilitating their participation in review cycles, so educators develop a full understanding of the test development process; how test items are aligned to curriculum standards; rubric development; and the rigor associated with screening and placing items on tests
- **Field Test Training and Scoring**—Facilitating teacher involvement in field testing and the item evaluation process
- **Rangefinding**—Convening groups of teachers to set scoring standards and to score, discuss, and select papers to be used in scorer training.

Interpretation of Performance-Based Items

Innovative multi-part and/or simulation-based, performance-based items allow collection of richer data on student performance. Interpretation of student responses need not be limited to simple ranked rubric scores, but rather can take advantage of multiple sources of data to support measurement of greater depths of knowledge and skills, including metacognitive and critical thinking skills, as well as multidimensional models of student learning.

Question

- Approach to releasing assessment items during each assessment cycle in order to ensure public access to the assessment questions; and

Response

One way to provide information about an assessment is to provide access to assessment questions. Given the secure nature of the assessment, it is not feasible to release all assessment items and tasks to the public. For example, linking items and embedded field test items on a full-length operational form cannot be released after a given administration. A sampling plan will need to be developed to select a set of items for release that provide an appropriately representative view of the assessment and stellar examples of how the common core standards are being measured. Because this information is needed before the assessment is operational, the sampling plan will need to be applied to the pool of field tested items prior to the first operational administration.

Released items may be used as part of several sources of information about the test. Item specifications include information about what is and is not assessed. Sample items are often included as part of the specifications to provide an application of the content parameters. These documents with sample items could be provided to the public. In addition, there are other opportunities for teachers and students to become familiar with the new assessment design before it becomes operational.

Educators from both K–12 and higher education will be involved in the test development process, which will likely begin in 2011 and be ongoing, to support the need for stand-alone field testing, operational testing, embedded field-testing, and item release. In the spring of 2012, some educators and students would have an opportunity to be involved in field testing—through a combination of cognitive labs and formal standalone field testing. This will allow examination of various item types and presentation and display options, and evaluation of their functioning and effectiveness with students and teachers. From the stand alone field test, sufficient items could be developed to be able to develop one “operational” form for use in a spring 2013 field test.

Whereas the spring 2012 field test would be administered to a sample of students from the participating states, the 2013 field test would be a full census field test and serve as a trial run for the first operational assessment in 2014. The 2013 assessment would be modeled after the future operational form in content representation, item type distribution, and non-scored items. This form, which would be available for all students, would provide students practice with the focus on critical thinking and problem solving. Following the assessment, we anticipate releasing all or most of the form so students, teachers, and the general public have access to a representative set of assessment items.

Ongoing item development will be required each year to produce enough new items to meet operational development and linking needs as well as a public item release policy that will support the annual release of some number of items, at least one form’s worth of items to be pulled primarily from the preceding year’s operational forms or item pool. This approach will meet the dual needs to provide appropriate, representative items for public examination and student practice needs while also providing us with the number and diversity of items needed to construct new forms and/or to replenish item pools for future administrations.

This annual release strategy will result in a growing pool of items that can be made available for a variety of uses. These items may be stored on a publically accessible web site. They may also appear as a practice form that students can access as part of a study program that includes links to further learning resources for specific items or other types of links back to the classroom. Other released items may be kept under a relatively more protected status and reserved for the use of teachers in classroom settings for diagnostic purposes or, with appropriate training or guidance, to facilitate instruction. Other items may find their way into item samplers with annotations or sample responses, study guides, professional development materials, and general information guides regarding the assessment. As the assessment matures and evolves into a truly adaptive assessment system with innovative item types, released items will continue to provide important information to the various stakeholders.

Question

- Technology and other resources needed to develop, administer, and score the assessments, and/or report results.

Response

Please see the **Technology & Innovation in Assessment** section within this response for recommendations regarding the technology needed to develop, administer, score, and report the assessment results. An overarching recommendation is that the new assessments need to be delivered online starting with bridge assessments and setting the stage for a fully integrated, student-centered online learning system.

In terms of other resources, local capacity building is needed to help provide consistency and longevity in whatever learning system is implemented. Teachers, administrators, and educational support personnel will not necessarily be ready to implement rigorous, college- and career-ready standards overnight. Similarly, administrators and support personnel will not necessarily be ready to support the data-driven decision making needed from a comprehensive data management system. As such, consideration must be made regarding the timeframes, expenditures, and requirements of such a system. A phased approach is recommended, starting with the best of what exists now and moving toward the vision of a fully integrated, student-centered online learning system encompassing all aspects of education (teaching, learning, assessment, professional development, and quality management).

The new assessments will also require greater collaboration between K–12 and higher education and the workforce so that the assessments are valid and useful.

Question

3. ARRA requires that States award at least 50 percent of their Race to the Top funds to LEAs. The section of this notice entitled Design of Assessment Systems—LEA-Level Activities, describes how LEAs might be required to use these funds. What activities at the LEA level would best advance the transition to and implementation of the consortium’s common, college and career ready standards and assessments?

Response

As stated above, States will be required to award at least 50 percent of their RTTT funds to local education agencies (LEAs). LEA funds should be used for one or more of the following activities:

1. Developing a Roll-Out Plan for the Implementation of Standards and Assessments. This might include:

- Covering the costs for participation in a full census field test of the new system of assessments
- Aligning the LEA's high school graduation requirements with the new end-of-course assessments
- Updating curricular frameworks and instructional materials
- Replacing formative and interim assessments with measures aligned to the common standards
- Enhancing professional development materials so that standards and data from the assessments are integrated into classroom practice

The funds could be used to support costs at each stage of the roll-out, and building capacity that can be sustained once the assessment system is operational.

Whenever possible the consortium of states or a state and partner LEAs—rather than individual LEAs—should develop or acquire formative and interim assessments and professional development training that can be shared. For example, as a member of the American Diploma Project Algebra II Consortium, Arkansas contracted with a vendor to create online professional development modules for the Algebra II Exam—instead of asking each district to manage its own professional development. The content is being rolled out to all districts online. In addition, there are regional, face-to-face professional development training sessions followed by a train-the-trainer model. Other states in the consortium are now interested in leveraging the professional development training that Arkansas has developed and using a consortium model to create additional training sets. Should a common professional development offering exist for the common assessment system, it could be made available for LEAs to purchase.

LEA funds might also be used to allow members from districts across a state to collaborate in addressing issues involved in the roll-out of the assessment system. For example, the Texas Consortium on School Research, which includes 19 geographically and demographically diverse school districts in the state, was formed to build capacity for research around school improvement. The group collaborates with local and national experts, and shares practices and knowledge to create solutions.

2. Investing in Technology and Infrastructure. One of the four assurances of the ARRA is to use data systems effectively. The states that have made the most progress in developing longitudinal data systems have integrated their statewide assessment programs with their data systems through the use of online testing. Together, the systems provide the state with enhanced quality, accessibility, analysis, and reporting capabilities for their preK-20 education agencies. However, many states have not started the process of transitioning from paper-based testing to online testing. And, while 22 states offer some form of online statewide student assessment, only 3 of those states have made online tests mandatory for students. Therefore, few states will be able to transition to the new online, common assessment system without a comprehensive roll-out plan that phases in technology and infrastructure by 2014.

As part of a roll-out plan, it is recommended that participating states form state technology teams to manage the transition from paper to online testing and administer statewide surveys to assess current technology and infrastructure within the LEAs. For example, a statewide technology survey in Texas found that only 6 percent of schools statewide have sufficient technology and infrastructure to allow all students to test within the current testing window (Texas Education Agency, 2008). However, if the

window was expanded to one week per test, approximately 65 percent of the state's schools have enough computers to support full online testing.

Based on the results from statewide technology and infrastructure surveys, state funds could then be allocated to LEAs for:

- Purchasing additional computers and supporting devices (keyboard, mouse, printer)
- Creating a computer lab that could be used for instruction and testing
- Upgrades to electrical power, Internet connections, bandwidth, and servers to support online testing
- Upgrades to technology to support special needs students participating in online testing, such as refreshable Braille displays, haptic devices, and other alternative augmentative communication devices
- Technology training for students and teachers
- Modifying curriculum and instruction to integrate technology and learning—to expand the capacity and potential of instruction

Prioritization for funding should be given to struggling schools and those willing to serve as “zones of excellence” in modeling the transformation to a digital teaching, learning, and data driven system. Students at “zones of excellence” schools would be the first to participate in pilot and field test administrations and would provide feedback and lessons learned to support late adopters.

Early adopters who already have technology plans (and the needed infrastructure) in place could be provided grants to update their plans to integrate them with the common assessment system and could serve as mentors to other LEAs. For example, Pamlico County Schools in North Carolina was recently awarded \$1.25 million to purchase laptops for high school students and other technology for earlier grades. The new grant will allow \$645 per student at the primary, elementary, and middle school levels and \$1,200 per student at high school. As a requirement of the grant, the school will be part of a study to see how technology is affecting education. The data from this study and feedback from the district will be helpful as other districts in the state purchase hardware and integrate technology into daily instruction.

While the Pamlico County School grant allowed up to \$1,200 per student (1:1 student to laptop ratio at the high school level), the statewide survey conducted in Texas found that if their testing window was expanded to one week and the student to computer ratio was reduced to 4:1 this would be sufficient—that 1:1 student to computer ratio was not required for online testing. The additional number of computers required for full statewide readiness capacity in Texas would be slightly more than 152,000. Including the costs associated with infrastructure and personnel readiness, the total estimated costs for full transition for online testing would be \$310 million, with additional ongoing operational costs of \$151 million. Given the 4:1 ratio this equates to \$510 per student in upfront costs and an additional \$248 per student in ongoing operational costs, given Texas' existing readiness for online testing.

3. Building Capacity and Support for Teacher Scoring. The Department is particularly interested in assessment systems in which teachers are involved in scoring of constructed responses and performance tasks in order to measure effectively students' mastery of higher-order content skills and to build teacher expertise and understanding of performance expectations. A teacher-scoring model that is effective both in terms of turnaround and cost will require LEA investments in labor, training, and technology to support online scoring and to turn the scoring task into part of a larger, more meaningful professional development experience. Questions that should be answered include:

- Will teachers be asked to score during regular work hours, as part of their regular salary, or will they be paid to score items during non-work hours?
- Will teachers be expected to score items from home and/or work? If from work only, technology implemented for online testing will be sufficient to support online scoring. If teachers need to be able to score from home as well, additional technology costs may apply (e.g., LEAs purchase laptops for teacher scorers).
- Are all mathematics and ELA teachers in grades 3–8 and in high school in the consortium states required to score, or will this be a professional development option, perhaps for new or inexperienced teachers?
- What training, monitoring, and feedback will be required of the teachers?

Question

4. If a goal is that teachers are involved in the scoring of constructed responses and performance tasks in order to measure effectively students' mastery of higher-order content and skills and to build teacher expertise and understanding of performance expectations, how can such assessments be administered and scored in the most time-efficient and cost-effective ways?

Response

The most time-efficient and cost-effective models will involve online test administration coupled with online scoring processes. A combination of online, human-scored, and automated scoring will:

- Promote scoring efficiency, quality, and consistency
- Support significant improvements in test item quality, including expanded use of performance-based items, similar to international models
- Involve teachers in the development and scoring of test items

Many states have started the transition to online testing but are still developing and administering their exams in both modes—paper and online. In order for online testing to be truly time-efficient and cost-effective and in order for it to support the problem-solving and critical thinking skills implicit in the common core standards, it needs to be the single mode of administration. As long as states are supporting both paper and online testing, the true savings will not be recognized—as evidenced in duplicate processes for item development, review, and administration.

Teacher Scoring

Key considerations for the scoring platform include the following:

- Using a nationwide pool of teacher-scorers, so the pool of teachers scoring student responses is as demographically diverse as the students taking the tests. A distributed, no bricks-and-mortar scoring model will enable local participation but is scalable to a national level. It is highly cost effective, because states will not have to pay for facility and equipment costs, and very efficient because of the large numbers of teacher-scorers who can be hired once the constraints of location-based scoring are removed.
- A scoring system that routes student responses anonymously, and uses industry standard security protocols such as Secure Socket Layer (SSL) technology.

Note: See our response to **General Assessment question 2** for more detailed recommendations regarding recommended scoring requirements.

Automated Scoring

Automated scoring is key to scale, but this does not have to occur at the expense of teacher involvement. Systems can be set up so items are double scored, with one score assigned by a human (teacher) scorer and one assigned using artificial intelligence. For lower complexity items that do not require double scoring, the artificial intelligence engine can assign the score with an audit conducted by teachers. Moreover, teacher-scored papers will be used to train the engine thereby creating a system where teachers are driving the scoring activity and setting the standards for scoring, even when the engine assigns one of the scores.

Automated scoring is particularly well suited to ELA test items, including passage-based reading items and essay responses to writing prompts. Automated scoring engines combine background knowledge about English in general and the assessment in particular along with prompt- or item-specific algorithms to learn how to match student responses to human scores. The scoring algorithm is adapted for each prompt based on the types of answers students write in response to that prompt and to the way in which human readers score those answers. Research has demonstrated both the accuracy and efficiency of automated scoring engines.

Machine Scoring of Online, Innovative Performance Items

Similar to automated scoring, online, innovative items are machine scored. These items are part of the online testing platform and scored immediately within that platform. Innovative items are particularly well suited to science and mathematics; are interactive; and are very engaging to test-takers, because they leverage the “gaming” qualities that can be achieved through online delivery.

Innovative items scored directly and immediately in the testing engine result in accurate, reliable, and very efficient scoring. Like automated scoring, this model does not have to preclude teacher participation. Teachers can be involved in item and rubric development, review, and field testing activities, thereby setting the standards against which the items will be scored.

In contrast to automated scoring, in which adaptive algorithms that require human-generated training sets are applied through dynamic rules automatically by computer, machine scoring involves simpler scoring rubrics that are applied through fixed rules automatically by computer. As such, they require no human training once initial item and rubric development, review, and field testing has been completed. This allows accurate, reliable, and cost-effective use of innovative items, such as simulations, which are particularly well suited to assessing students' higher-order science and mathematics knowledge and skills.

From Summative to Formative Assessment

Another critical element to meaningful teacher involvement is complementing summative tasks and assessments with formative, interim, and classroom-based assessments. These can also involve teachers scoring locally. The same national, web-based scoring platform can be used to support local scoring where tests are delivered online. Released items from summative tests can form the basis of interim, classroom-based tests, and offer professional development for teachers and hands-on learning activities for students. Further, a number of off-the-shelf solutions exist that can help integrate testing into the learning cycle. The ideal end state of a seamless and transparent cycle of testing and learning, where curriculum and the measures of its effectiveness are embedded into the same delivery system, can be fully achieved by online test delivery.

Question

5. Given the assessment design you proposed in response to question 1), what is your recommended approach to competency-based student testing versus grade-level-based student testing? Why? How would your design ensure high expectations for all students?

Response

As articulated in other sections of our response, our vision is an integrated learning system that uses technology for assessment delivery, allows for meaningful assessment content including performance-based measures of critical thinking and problem-solving, and one that coordinates and links assessment, enrollment, and learning information into actionable data useful to teachers, students, administrators and others. Furthermore, this system needs to be linked directly to instruction, not just by measuring the same common core content standards but by allowing teachers access to professional development, capacity building, and formative assessment linked directly to daily instruction. Such a system will efficiently link instruction, professional development, and all aspects of instruction (formative, interim and summative) to measures of preparedness, such as high school credits, college placement, or college and workplace readiness.

One reality such a system must take into account, however, is that fact that currently in the US, our entire educational system is predicated on the structure of essentially age-based enrollment classes. While research has shown that this design is not optimal (i.e., differences in personal development, differences in starting support structures, differences in starting experiences, etc.), it is nonetheless not likely to change in the near future. Therefore, a learning system must be able to personally adapt both instructional and assessment protocols to fit into such a system.

Recent efforts documenting learning progressions and individual growth trajectories seem to conflict with this grade-level organizational structure. We propose to work within this structure but propose flexibility in

adapting assessment toward student learning. Simply stated, assess students on the material after it has been taught (specific grade-level content, cognitive attributes like critical thinking and metacognitive skills like problem solving that may require across grade information).

This can be accomplished and based on our individual learning progressions or growth trajectories perspective, without watering down the requirements to stay on track for graduation or college readiness within the grade-based system. If an enrolled fifth grade student is reading at the third grade level, it seems absurd that we should measure them on the reading standards of grade five knowing full well they are not able to attain those standards. Rather, we need to instruct and measure this student progressively so that by the end of a defined period of time they are reading at the appropriate level. For example, perhaps the goal will be to have this fifth grade student reading at the eighth grade level in two years. As such, this would define the target at the end of a growth trajectory and take into account the progressions to that point. During these two years, the student should use learning materials at the appropriate level to reach this target.

Such a personalized system based on individual learning progressions or growth trajectories will require an integrated learning system such as the one we have described in order to know where each student is on their progression or growth track relative to where they ultimately need to be for graduation, college placement, or college admissions. Off-grade-level instruction or assessment is not the enemy—rather the key is to unlock the door to timely identification and implementation of a plan for catching up and surpassing this below grade level paradigm. It is precisely this type of timely intervention that can be highlighted immediately through a learning progression or growth trajectory system even if implemented with the existing on grade level classification.

Question

6. Given the assessment design you proposed in response to question 1), how would you recommend that the assessments be designed, timed, and scored to provide the most useful information on teacher and principal effectiveness?

Response

Using assessment data to provide information about teacher and principal effectiveness relies less on the design, timing, and scoring of assessments than on a comprehensive data system with unique student identifiers that also tracks students' course taking, their teachers, and their schools. The assessment itself should be designed to fairly and accurately measure students' acquisition of knowledge, skills, and abilities with fidelity to how students have been instructed and how they would be expected to demonstrate their mastery in the real world as defined by the common core standards. Given such an assessment, and a comprehensive student record system, aggregating assessment data to make inferences about teachers and principals is feasible. However, the specification of the statistical model to make such inferences is both technically and politically complex. And given this complexity, it is essential that the data and calculations used in such models be transparent, replicable, and audited (no proprietary "black boxes").

There is still significant debate over the validity and specification of various growth and value-added models. While most growth models, developed under the Department's pilot growth models program, were intended to alleviate the ramifications of NCLB adequate yearly progress in areas where significant

growth was being achieved even though progress targets were not, the technical aspects of the models are similar to those used to model teacher and principal effectiveness. The best known and longest implemented of these models is the Tennessee Value Added Assessment System (Sanders, Saxton & Horn, 1997). Even though this model has been used for more than a decade, there continue to be many criticisms of its implementation (Koretz, McCaffrey, and Hamilton, 2001; Kupermintz, 2002). In addition to technical considerations, such models are also politically controversial. Teachers and administrators are often wary of being held accountable for test score gains and/or having merit pay tied to such gains. They argue that the assignment of students across teachers is not equitable, that some are purposefully engaged in teaching students with academic challenges, and that some students come into a grade more unprepared than their peers.

Given the controversy and the potential impact of using student gains to evaluate teacher and administrator effectiveness, we propose a thoughtful approach similar to one adopted by the Department for projects such as the development of the National Educational Technology plan. Critical reviews of the research on value-added modeling and data analysis projects have shown that teachers have a discernible and persistent impact on student achievement (McCaffrey, Lockwood, Koretz and Hamilton, 2003). This, in itself, is enough to suggest that such modeling should be done to identify effective teachers and, thereby, effective teaching practices that can be shared with less effective teachers. Yet, specifying the model, identifying norms for gains, determining how student populations should be disaggregated, and determining the relative merits of a single model versus multiple models should be undertaken by a panel of stakeholders and experts. This can help to ensure that the data is used in a way that can do the most good and the least harm. This panel can be actively involved in modeling data, defining the technical characteristics of a model or models, and determining the impact of having a single model versus individually adopted state models. While such details are investigated, states without comprehensive data systems can build their capacity and those with existing capacity can use their data to begin to identify effective teaching practices for dissemination.

Specific Technical Assessment Questions

Question

1. What is the best technical approach for ensuring the vertical alignment of the entire assessment system across grades (e.g., grades 3 through 8 and high school)?

Response

Aligning Assessment System with an Integrated Approach

As a starting point, it is useful to consider Achieve's *Accelerating College and Career Readiness in States: Standards and Assessments* report, which recommends developing anchor assessments aligned to college- and career-ready standards that should be taken by all students statewide. These anchor assessments could be end-of-course exams and their primary purpose would be to determine if students have met the college- and career-ready standards in reading, writing, and mathematics for the end of high school.

Once the anchor assessments are identified, Achieve then recommends that all statewide assessments should be vertically moderated to the anchor assessments, including any other large-scale assessments given statewide earlier in high school and the tests for elementary and middle school. "The goal is for students, parents, and educators to know whether students in any tested grade are on-track towards meeting the college- and career-ready standards by graduation," according to Achieve's report.

Achieve's recommendation supports the concept that the common assessment system standards will be vertically aligned; K-12 standards will cascade down from the college and career readiness standards. However, while it makes sense to put some tests at grades 3–8 on a vertical scale, there may be some limitations when trying to place all high school assessments on a vertical scale.

From a measurement perspective so that assessment data can provide sound and useful evidence, aligning an assessment system requires an integrated approach to the various system components. These include the standards of the assessment system, the design and structure of the assessments, and the psychometric methods used to produce scores from the student responses.

Traditionally, vertical alignment in content standards is best achieved by focusing on the nature of content linkages from one grade to the next. These should be articulated clearly and, if done successfully, provide the basis for building assessments that are also vertically aligned.

A Council of Chief State School Officers (CCSSO) report by Wise and Alt (2006) provides one mechanism for considering a vertical alignment process by building on Norman Webb's (1997) framework and methodology for assessing the alignment of tests to content standards within a given grade. That is, the nature of content linkages across grades by determining:

- The level of concurrence between objectives between adjacent grades
- The extent to which comparable objectives increase in depth from one grade to the next

- The extent to which the range of content increases from one grade to the next
- How the balance of content representation changes from one grade to the next

If these elements are clear in a set of vertically articulated standards, it becomes possible to make specific content-referenced statements about how we expect students to progress toward college and career readiness as they progress through the educational system.

The common assessment standards can provide opportunities to implement new policies and new psychometrics that can capitalize on vertically aligned standards and better assess student growth. From a policy standpoint, permitting the assessment of “off-grade-level” standards can significantly improve the measurement of students who are performing significantly below or above the levels of typical students within a particular grade. This approach to assessment is consistent with a philosophy that focuses on personalized learning and recognizes the hierarchical nature of knowledge and skills that are required for students to be college and career ready. It is also consistent with the increased emphasis on measuring growth rather than just status in summative testing used for accountability purposes. It anticipates the blurring of grade levels that many predict will begin to occur in classrooms of the future as the US implements new instructional practices that have been proven successful in other countries. Such off-grade-level assessment does not need to be perceived as a “watering down” of grade level standards. In fact, if documented well, they will achieve just the opposite, a clear indication of where a student is currently functioning and whether they are ready for success.

From a psychometric standpoint, computerized adaptive testing (CAT) is a natural vehicle for permitting the measurement of off-grade-level standards in the assessment system. Given clear and coherent vertically articulated standards, CAT enables off-level items to be administered to students in flexible and non-intrusive ways. It avoids the labeling that was traditionally associated with students taking out-of-level forms in norm referenced testing because each student can begin at the same place and have their assessment branch as appropriate based on their performance.

As described in the **General Assessment Questions** section, a bridge assessment system that would begin as a fixed-form, online criterion-referenced test (at grades 3 through 8 and a series of end-of-course tests in high school) and move to an adaptive approach over time is recommended. A variety of adaptive approaches may be considered (e.g., traditional CAT, multi-stage testing, variable or fixed-length testing). The appropriate adaptive testing solution will ultimately depend on the content and structure of the exams.

There may be some tension between adaptive testing and the possible need to use open-ended items that require human scoring. Some of this tension may be lessened by assertive use of automated scoring technology. It should be noted that automated scoring still requires that human scored items be used to train the automated scoring engine; typically, several hundred scores from trained expert judges are required. However, in lieu of 100 percent automated scoring, one efficient way is to use adaptive multiple-choice testing in combination with constructed response assessment administered in stages. For example, stage one would consist of machine-scorable items administered adaptively. Stage two would include items/tasks that would use human scoring. Note that these items/tasks could be selected adaptively from an available pool based on the performance of each student in stage one.

The proposed use of fixed-form bridge assessments is in recognition that an adaptive test will require a period of development. There are challenges associated with different levels of motivation that might characterize student performance as items/tasks are tried out in the developmental period of the assessment. While an adaptive assessment will fit well into a vertically aligned system, adaptive testing may not be possible at the outset of the assessments. Alternatively, the initial year or years of the program could use some number of fixed-form tests, which would be randomly selected and administered to students. These tests would include sets of carefully selected anchor items administered in field-test positions across grades and used in vertical scaling work. In addition, standard setting studies could be conducted with reference to the vertically articulated standards, the underlying psychometric vertical scale, and comparisons of the standards to international benchmarks.

Question

2. What would be the best technical approach for ensuring external validity of such an assessment system, particularly as it relates to postsecondary readiness and high-quality internationally benchmarked content standards?

Response

External validity evidence for the assessment is to be contrasted with the validity evidence that will be exhibited in the development of vertically aligned standards and the assessments that will measure the standards. By all indications, the common core standards will be articulated by drawing on evidence of postsecondary readiness and will be benchmarked internationally. Effective vertical alignment of the Common Core assessment system should cascade this focus from high school to the 3–8 assessments. In this manner, a certain amount of validity evidence for the Common Core assessments as related to postsecondary readiness and high-quality internationally benchmarked content standards should be documented through the assessment development process itself. In other words, postsecondary readiness becomes an intrinsic part of the construct being measured and validated from the outset.

Establishing external validity evidence for the common assessment system will require a variety of methodological approaches. Careful plans will need to be in place early on to validate assessment scores and claims made based on them, as well as a long-term research agenda to continuously improve the efficacy of the assessment system. Even if consortia start with internationally benchmarked and college and career readiness standards, a research plan must be in place for checking and updating these standards, and making necessary changes to the test design. The research needs to extend beyond the standards. Because the high school tests will claim to measure college readiness, there should be a plan in place to validate that claim.

One model for this validity work is the recently developed American Diploma Project (ADP) Algebra II End-of-Course Exam. For this exam, three types of validity studies were conducted to support setting a college-ready performance level on the exam¹:

¹ See “American Diploma Project Algebra II End-of-Course Exam Standard Setting Briefing Book”, available at <http://www.pearsonaccess.com/cs/Satellite?c=Page&childpagename=ADP%2FadpPALPLLayout&cid=1205460857541&p=1205460857541&pagenam=adpPALPWrapper&resourcecategory=User+Documentation&start=20>.

1. **Concurrent studies**—Student scores on the ADP Algebra II Exam were matched to student scores to other state and national assessments to establish relationships, including those with existing measures of college readiness.
2. **Cross-sectional studies**—The ADP Algebra II Exam was administered to students at the beginning of the semester of their college mathematics course and compared to their final grade in the course to determine how well a student’s performance on the exam predicts his/her performance in the college math course.
3. **Judgment studies**—Feedback was gathered from more than 100 college professors who teach College Algebra and Pre-Calculus courses regarding the relevance of the ADP Algebra II Exam standards to their course, draft performance level descriptors, and recommended cut scores.

Similar research could be undertaken during the initial years of a common assessment, concentrating in particular on the high school assessments. Once the assessments were in place long enough, criterion-related validity studies could track students from high school to college and could be expanded to include non-assessment indicators (e.g., academic behaviors, cognitive strategies, contextual awareness and skills). As longitudinal student results on the assessments became available, research could be expanded to consider relationships between performances on assessments taken earlier, high school assessment results, and college or career outcome measures.

Obtaining validity evidence with respect to internationally benchmarked content standards can occur through similar studies. In particular, relationships between the performance on common assessments and international assessments (e.g., PISA, PIRLS, TIMSS) can be established. One way these data might be collected would be to embed items from these assessments in field-test positions on the appropriate corresponding common assessments. Another way might be to convene experts to compare test blueprints with those of corresponding international instruments (which could also include country-specific assessments). A third approach might be to administer the common assessments to selected samples of international students and compare their performance with US students taking the assessments.

Given the multiple purposes desired from the assessment system, a system of longitudinal or cohort analyses would also provide validity evidence that could also be used to monitor changes in the end-to-end system likely to result from its implementation. For example, if graduating college-bound seniors were tracked into college and the workplace, a powerful portfolio of evidence would exist that would not only shed light on the effectiveness of the assessment system as an index of college or career readiness, but would also serve to link the post-secondary and high school instructional systems as well. Such a profile would finally link high school performance and assessment scores with college courses and performance and insights into the work place.

Taken in concert, these approaches to gathering external validity evidence for the Common Core assessments will provide strong documentation related to college readiness and comparisons with internationally benchmarked content standards. As states develop their research plans they should partner with higher education leaders and faculty for participation in the research studies, and to determine what it would take for results from the high school assessment(s) to be used for placement into credit-bearing entry-level courses and/or admissions. For example, in California, results from the 11th grade Early Assessment Program (EAP) can be used to place students into college-level mathematics classes or signal if the students need additional mathematics preparation during their senior year.

The consequences of test score use offer an important source of evidence for the external validity of an assessment system from which information is useful for the following:

- Teaching, learning, and program improvement
- Determinations of school effectiveness
- Determinations of individual student college and career readiness, such as determinations made for high school exit decisions, college course placement in credit-bearing classes, or college entrance

The consequences of test score use serve as evidence for the external validity of an assessment system in two ways. First, consequences can serve as evidence for external validity when findings on test score use either confirm or question the intended meaning of test scores. For example, consequences would support external validity if student achievement increased after assigning students remedial instruction based on results from a formative assessment. In contrast, consequences would question the external validity of a formative assessment if student achievement remained the same after assigning remedial instruction based on assessment results.

Second, consequences of test score use can serve as evidence for external validity when findings on test score use confirm or contradict predictions. For example, consequences would support external validity if few college freshmen predicted to be college ready by a high school assessment failed first-year college courses. In contrast, consequences would question external validity if many college freshmen predicted to be college ready by that same high school assessment failed first-year college courses.

External validity is called into question when the consequences can be linked to a flaw in the conceptualization of test score interpretation and use. This flaw in the conceptualization of test score interpretation and use may be due to construct under-representation when the assessment fails to assess important aspects of a given construct. Alternatively, the flaw in the conceptualization of test score interpretation and use may be due to the inclusion of sources of construct irrelevant variance when the assessment measures not only the construct but also something irrelevant to the intended construct.

Question

3. What is the proportion of assessment questions that you recommend releasing each testing cycle in order to ensure public access to the assessment while minimizing linking risk?² What are the implications of this proportion for the costs of developing new assessment questions and for the costs and design of linking studies across time?

Response

It is both an ethical imperative and a standard of industry best practice (see, for example, AERA, APA, NCME, 1999) that testing organizations (and state and federal education agencies) maintain transparency and communication with the public, and remain accountable for their products and services. Periodic release of test items from previous administrations and item pools is an important aspect of that

² Michael J. Kolen and Robert L. Brennan, *Test Equating, Scaling, and Linking: Methods and Practices* (2nd ed), 2004, New York: Springer-Verlag. See especially: Chapter 6, "Item Response Theory Methods," Section 9, "Using IRT Calibrated Item Pools"; and Chapter 8, "Practical Issues in Equating," Section 1, "Equating and the Test Development Process" and Section 6, "Conditions Conducive to Satisfactory Equating." See also Hedges, L. V., and Vevea, J. L. (1997). A study of equating in NAEP. http://www.air.org/publications/documents/hedges_rpt.pdf

transparency, and the public deserves to see examinations and items that have the potential to play a significant role in the academic and career paths of their children.

At the same time, there is a need to continue to develop new forms of a test that are parallel in terms of content, difficulty, and the interpretation of student scores derived from them. To do this, a pool of items must be maintained that fully represents the appropriate content and psychometric characteristics needed to provide items with which to develop new forms. Further, sufficient items must be maintained in secure status so they can be added to future operational forms as linking items to enable the calibration and equating processes necessary to promote equivalence of score interpretations across forms and years. Additionally, secure items can also function in research situations such as investigating potential score drift.

These competing priorities—to release items to the public and to retain items for use as linking items, as research items, or in some situations, as re-used operational items—dictate the need for careful planning in determining item development needs. At a minimum, the equivalent of at least one operational form per year, per content area and grade (not including embedded field-test or linking items) should be released in order to provide public access to the assessment as well as to provide students and teachers with a representative sample of item contents and types of test form length and complexity, and to accommodate requirements of current state laws. This recommendation is based on a minimum of four operational test forms per year per content area and grade, or approximately 20 percent to 25 percent of the items.

Costs for item development are, of course, directly impacted by the structure of the release policy. In this case there would be an ongoing need to replenish the item pool to support attrition due to release and to support ongoing field-testing of new items. One approach may be to build a pool of released (and “semi-released”) items that may be used for a variety of purposes, including teacher professional development. Planning at the outset to build such an item pool geared to further productive educational use should help mitigate costs and increase value by leveraging high-quality test items to continue to provide meaningful information to teachers and students as well as the public at large.

The proportion of released items depends on the depth and breadth of the common core standards and how those standards are assessed. The goal of release is to provide examples of how the common core standards are being measured to enhance the meaning and intent of the standards. As such, it may be insufficient to conceptualize the release of a test form. It might be more useful to release examples of the tasks measuring common core standards (at some level of detail) periodically through out the school year, thereby greatly increasing the number of tasks released. Because of the field testing and bridging proposed, such a release can be accelerated or slowed based on the trade offs in cost, timeliness, and schedule. It is premature, however, to speculate how much of the pool needs to be released until the size and type is determined and clarified through the common core standards.

High School Assessment

Question

Provide recommendations on the optimal approach to measuring each student's college and career readiness by the time of high school completion. In particular, consider—

1. How would you demonstrate that high school students are on track to college and career readiness, and at what points throughout high school would you recommend measuring this? Discuss your recommendations on the use of end-of-course assessments versus comprehensive assessments of college and career readiness.

Response

Preparing Students for College and Careers

A growing body of research suggests a relatively clear root cause of why far too many students are graduating high school unprepared or under prepared for college—namely the lack of rigorous and relevant instruction in advanced courses (such as ELA and Literature, Algebra II, Chemistry and Physics). Moreover, cognitive science indicates that students need opportunities to acquire deeper content understanding and a context for developing higher order thinking skills like problem solving, analytical reasoning, and critical analysis. To create these opportunities, both instruction and assessment must create an environment for students to conduct complex performance tasks and sustained project work.

Of equal importance in achieving true educational reform is the need for a system of capacity building, instructional support, and implementation maintenance such that real instruction—not only in the content domain aspects of curriculum—but also in problem solving, information organization, query and decision making, can be realized. By implementing such an integrated and co-dependent system, schools can prepare students not only for success in a college of their choice, but also entry into the workplace, technical schools or community colleges, and the military.

The vision is based on the premise that discrete content knowledge and application will come from a focused and linked instructional/assessment model in which end-of-course measures are likely to play a predominant role. The fidelity, flexibility, and relevance of specific units of instruction associated with “course models” are a compelling reason to embrace end-of-course assessments. Most current statewide programs are calling for more advanced curricula that provide multiple years of core subjects such as in mathematics (Geometry, Algebra I, Algebra II, Pre-Calculus), ELA (English I, II and III and Literature), Science (Biology, Chemistry and Physics) and often the social sciences as well (US History, World History, and Government).

For example, in Texas in 2007 Senate Bill 1031 was passed, which called for the development of end-of-course assessment instruments for secondary-level courses in algebra I, algebra II, geometry, biology, chemistry, physics, English I, English II, English III, world geography, world history and US history. The end-of-course assessments for lower-level courses must include questions to determine readiness for advanced coursework. The assessments for higher-level courses must include a series of special purpose questions to measure college readiness and the need for developmental coursework in higher education. In addition, a student's score on each EOC assessment will be worth 15 percent of the

student's final grade for that course (<http://www.tea.state.tx.us>). Similarly, a subset of the ADP Algebra II Assessment Consortium states has also developed a common Algebra I end-of-course exam, with exam standards vertically aligned with the Algebra II exam so results will indicate student readiness for advanced high school mathematics courses (*Accelerating College and Career Readiness in States: Standards and Assessments*. Achieve Inc.).

If the instructional model in high school is course based, the research linking college readiness to advanced courses (ACT, 2005; Achieve, 2007; Hargrove, Godin & Dodd, 2008) is course-based, and the need for assessment linked directly to instruction is required, then why not use an end-of-course model for high school assessment? Working together, states and their partner LEAs in consortia share the burdens and benefits of developing high-quality common assessments, as well as share and replicate best practices to strengthen instructional capacity and ultimately increase productivity across US public schools.

Requiring all students to take advanced courses and designing these courses to link to learning progressions, growth models, and student growth trajectories creates a robust system to support students' progress toward the attainment of college- and career-ready standards. End-of-course assessments will make it possible to articulate learning progressions, individual student growth models, and growth trajectories to document empirically the probability that a student exiting high school is ready for college. This could be done periodically starting as soon as middle school (eighth grade, for example when many students are taking classes such as Algebra I) and through high school. This type of system would facilitate the use of early warning indicators and allow early interventions both for struggling students and for students ready to advance into more challenging courses.

For example, such a model could show a high school freshman who struggled with mathematics in middle school and who was signed up for a remedial math course in high school his probability of graduating ready for college. Similarly, this same model will show the much greater probability of success for a student who has the same attributes in middle school, but who signs up for Algebra I during their freshman year. Likewise, this model or projection of college readiness could be updated as students obtain instructional interventions and personalized instruction to increase their chances of exiting high school ready for college. Our proposed integrated system will allow tailoring individual instruction, remediation and/or accelerated work in middle school and high school to maximize the probability of success in college.

These models are not fiction or things of the future, they are used today (in Tennessee and Texas, for example) and can be used to counsel students regarding course selection, reward students, teachers and schools for fulfillment and can provide empirical measures of student progressions or growth toward college readiness.

Obtaining data and conducting research related to college readiness is relatively straightforward. For example, contrasting group models where students who are operationally defined as having been ready for college (for example, second semester returning freshmen with GPA of at least 2.0 and no remediation) can take the high school assessments. It is also possible to follow students from high school to college longitudinally. This permits direct empirical links between high school end-of-course assessments and college success. In addition, other profile information (such as courses taken and when, grade-point average, extra-curricular activities) can be used to round out college readiness criteria.

Question

Note: If you recommend end-of-course assessments, please share your input on how to reconcile the fact that college and career ready standards might not include all of the topics typically covered in today's high school courses.

Response

While the examples in the response above cite a traditional course sequence, the adoption of an end-of-course model is not predicated on states mandating a traditional course sequence. States may choose to take a more integrated and interdisciplinary approach within and across core content areas. They may also choose to build on the trend of not only expanding their Advanced Placement[®] offerings but also expanding opportunities for students to take dual enrollment college courses in high school. The goal is to dramatically increase the academic preparedness of students leaving high school for college—and to increase their options and opportunities to pursue a diverse range of careers in a dynamic technology-driven, global economy. To meet these goals, the high school experience must be designed to so that all students successfully complete four years of challenging course work including advanced math and science courses. An end-of-course model provides a modular architecture for states and their partner LEAs to create an environment in which students can cultivate rigorous content knowledge and complex skills through tightly coupled instruction and assessment.

Returning to Achieve's *Accelerating College and Career Readiness in States: Standards and Assessments*, "adopting common, and career ready standards, will impact a wide range of state and district policies and practices. Achieving true system alignment will likely mean that rigor must be increased across the board while at the same time there will be fewer, more streamlined expectations. Curriculum, coursework, and high school graduation requirements will need to be aligned to the common standards. Teacher training and support will need to be updated and upgraded."

Achieve highlights the following as some of the most important ideas for advancing college and career readiness:

- Ensure all students have access to college-and career-ready course of study
- Ensure all students have strong incentives to compete a college-and career-ready course of study
- Ensure the curriculum follows the standards
- Ensure that students have multiple pathways to learning the content knowledge and skills included in the standards through innovative pathways

However, exposure to content alone is not enough. Students must also learn how to learn. They must learn to take organized and accurate notes, to think, to reason, to debate, and to collaborate. They also must learn how to find answers to additional questions, learn discipline to complete homework on time, learn how to be inquisitive, and learn how to become motivated. In short they have to mature into thinkers. If we fail to challenge our students to do these things while we are teaching them the fundamental content skills associated with these courses of instruction, then we will fail in our mission to prepare these students for college.

It won't be enough to measure college and career readiness through the use of summative assessments alone. A richer, more integrated approach to assessment is required. This may include the use of interim

assessments, formative assessments, and other performance measures assigned by teachers (such as research papers, class projects, or experiments).

Lessons from high-achieving nations (Stanely, MacCann, Gardner, Reynolds and Wild, 2009; Darling-Hammond and McCloskey, 2008) and research on college readiness (Conley, 2007) highlight the need for an integrated curriculum, instruction, and an assessment system that allows teachers to build learning experiences that help students master all of the prerequisite skills that are necessary for college and career success. Teachers are involved in the development, administration, and scoring of the assessments to provide them with opportunities to deeply understand standards and information useful to teaching and learning.

There are significant potential benefits with supplementing a more traditional end-of-course testing model with locally-developed or locally-managed performance assessments. Teacher-administered, curriculum embedded assessments could promote the development of powerful curriculum and instruction in schools and districts. Curriculum-embedded assessments can also be transformative by providing diagnostic and formative information to teachers and administrators that is rooted in actual district and classroom practices. Including rigorous instruction and assessment targeting measurement of higher-order skills throughout the high school experience will also benefit students by focusing on critical college and workplace readiness skills and providing challenge and focus through the senior year (National Commission on the High School Senior Year, 2001). Although there are challenges with including such assessments in the system, particularly with respect to their role in accountability and graduation decisions, their inclusion will enhance the ability of the system to produce high school graduates who are prepared for college and career opportunities.

Assessment of English Language Learners

Question

1. Provide recommendations for the development and administration of assessments for each content area that are valid and reliable for English language learners. How would you recommend that the assessments take into account the variations in English language proficiency of students in a manner that enables them to demonstrate their knowledge and skills in core academic areas? Innovative assessment designs and uses of technology have the potential to be inclusive of more students. How would you propose we take this into account?

Response

Developing Fair, Valid Assessments for English Language Learners

Two guiding principles that are inextricably linked and must be addressed in the assessment of all learners take on additional significance in the assessment of English language learners (ELLs): 1) fairness and 2) validity. Assessments must be designed, developed and administered to be as fair as possible to allow individual students to accurately demonstrate what they know and can perform. And in order to do this, the construct—the relevant knowledge and skills to be assessed in a given domain—must be defined with precision.

Explicit inclusion of English language learner (ELL) students in instruction and accountability systems has been a priority of both the Individuals with Disabilities Education Act (IDEA) and NCLB legislation (Abedi, 2004). The emphasis on English language learning—and assessment—is appropriate so long as it does not interfere with ELL students' acquisition or demonstration of construct-relevant knowledge and skills. Thus, in assessing ELL students' proficiency in subjects other than ELA, such as mathematics, various strategies have generally been applied to minimize students' English language proficiency as a construct-irrelevant factor (i.e., don't interfere with students' ability to demonstrate the knowledge and skills). These fall into the following three general categories:

1. Testing in English with Linguistic Accommodations.

Linguistic accommodations are applicable when ELL students take the same test form as their non-ELL peers.

Specific linguistic accommodations for ELL students include:

- Clarification of test directions
- Breaks and extended time
- Reading assistance
- Bilingual dictionaries
- Bilingual glossaries
- English, ESL, or picture dictionaries

Since accommodations cannot assist students in construct-relevant ways, they may not include explanations, definitions, pictures, gestures, or examples related to subject-area terminology, concepts, or skills assessed. Accommodations should be consistent with the linguistic accommodations used with the student in routine classroom instruction and testing, and should be selected on an individual student basis, consistent with their background and needs. While students may need multiple linguistic accommodations, they should not be provided more accommodations than they require to access the text. In addition, accommodations should not be provided in a testing environment that will distract or disturb other students testing, nor should they make students feel self-conscious or stigmatized.

2. Testing in English with Linguistic Modifications.

Linguistic modification of test items involves alterations to the test language to lessen its linguistic complexity without affecting construct-relevance. These modifications are above and beyond the linguistic accommodations discussed earlier. Such tests reduce or eliminate linguistic features that might otherwise increase the construct-irrelevant reading load of test items. Items are developed and/or modified using simple, clear, grade-appropriate language and avoiding complex grammatical constructions and idiomatic speech that may be unfamiliar to ELL students.

Specific linguistic accommodations for ELL students include the following:

- Language structures/syntax
- Vocabulary
- Contextual information
- Formatting

3. Native Language Testing.

The options for developing and administering assessments in native languages are described in response to Question 2 in this section.

While there are advantages and disadvantages to linguistic accommodations and modifications in terms of validity, reliability, and fairness, it is impossible to make blanket decisions on which of these approaches is best for measuring all ELL students. Various student-specific factors, in interaction with subject area and item formats, affect which solution will best allow students to demonstrate their subject matter knowledge and skills.

In order to make valid decisions about the development and administration of content area assessments for ELL students it is necessary to:

- Better evaluate English language proficiency in terms of genre (e.g., expository vs. narrative) and subject area
- Track students' language(s) of prior instruction
- Understand ELL students' native language proficiency

This information can then support decisions that maximize test validity and reliability using grounded approaches, such as evidence-centered design and universal design, as described in the next section, **Assessment of Students with Disabilities.**

While linguistic accommodations and modifications may be used to assess students' proficiency in subject areas other than ELA, such as in mathematics, in both teaching and assessing ELL students in ELA, we must allow for variation in the amount of native language knowledge and skills each student brings to the process of learning English. Use of subject-area accommodations and modifications is much more complicated and potentially problematic for ELA. In order to move beyond retrofitted accommodations and the challenges to validity they potentially bring, learning standards must explicitly state conditions under which students, during testing, are to understand what is expected of them by test directions and stimuli and the conditions under which they demonstrate their knowledge and skills. For example, during item development and administration, it must be clear whether a high school ELA item intended to assess reading comprehension can be read aloud.

Role of Technology

Technology provides additional ways to provide linguistic accommodations and modifications to ELL students to increase validity and fairness. Translations of words or phrases can be readily accessed by the student simply by clicking on the word or words not understood. Pictures can also be provided to clarify meanings of words or provide additional context. Audio components to provide an oral version of test items can be added as an optional feature that the student may turn on or off as desired; these can be accomplished using text-to-speech and/or digitized human speech. Technology-delivered accommodations increase control over the testing situation, thereby helping to standardize the administration and support meaningful score interpretations and comparisons. In addition, technology can facilitate student access to accommodations as needed and help decrease test anxiety.

Additional Considerations

In addition to test development and administration considerations, translation of test score reports and/or parent guides is important for parents and guardians of ELL students, as their English language proficiency is often poorer than that of their children.

As we work to strengthen the quality of instruction and assessment for ELL students, it is worth considering many or most of these students possess something valuable: proficiency in their native language. Hopefully over time we can learn to better leverage these skills and competencies to expand and accelerate foreign language learning for all our students even as we work to improve the English proficiency of our ELL students.

Question

2. In the context of reflecting student achievement, what are the relative merits of developing and administering content assessments in native languages? What are the technical, logistical, and financial requirements?

Response

Assessing ELL students in their native language in some cases may allow them the best opportunity to demonstrate their construct-relevant knowledge and skills in subject areas other than ELA. Offering versions of a test in the students' native language is one way states can more accurately assess students' content knowledge, separate from their English language proficiency (Stansfield & Bowles, 2006).

Written translations are most appropriate for students who:

- Are literate in the native language
- Have had formal education in the home country/language
- Have been educated bilingually in American schools through a bilingual education program, but whose English language skills are not yet sufficient for testing in English (Bowles & Stansfield, 2008)

Three general options exist for developing native language versions of tests for ELL students:

1. Translation.

Translation consists of rendering content originally written in English to the native language of the ELL student. Only minor modifications can be made to account for linguistic or cultural differences, ones that have no impact on intended constructs. As a result, threats to validity and comparability are minimal. While the simplest and least expensive of the methods for creating native language test versions, translation is nonetheless a time-consuming and expensive process.

2. Adaptation.

Adaptation involves linguistic and cultural changes to content beyond pure translation. In many cases it is a necessary step to minimize construct-irrelevant linguistic and cultural factors that would otherwise threaten the validity of ELL student scores. Because adaptation is more likely to impact construct-relevant factors and hence impact validity and comparability, it is generally considered a modification rather than an accommodation (Stansfield, 2003). As such, adapted tests must be treated as separately test and so essentially doubling test development efforts, including review, field testing with target ELL student populations, standard setting, and linking. In addition, field testing and psychometric evaluations might be hindered by low numbers of students in target populations. As a result, adaptation is significantly more costly than translation.

3. Parallel development.

Parallel development, or concurrent development (Solano-Flores et al. 2002), involves simultaneous creation of construct-equivalent test forms in both English and alternate languages. It can be thought of as a universal design approach since it does not rely on after-the-fact adjustments of content in isolation of the authorial intent. As such, in principle it should result in better validity. However, it is a more costly process than adaptation because it involves a separate test development process.

Two additional options exist for providing native language supports for ELL students.

First, audio-recorded translations of can be provided in place of written native language test content. These can be developed in advance through scripted oral translation or in real time through sight translations. In general scripted oral translation is the better choice of the two, as it allow for greater standardization of student experience and hence comparability. However, translation involves additional efforts during test development. Also, through technology, students can have on-demand access to just the translations they need through text-to-speech and/or digitized human speech.

Second, students can receive bilingual test booklets. In the case of online testing, students can have on-demand access to native language translation of test items when taking the test in English, or vice-versa. While the use of bilingual test booklets has generally been successful and has been adopted in several states, it is challenging for many students to go back and forth between the English and native language content, especially in languages where there is poor one-to-one correspondence of linguistic constructs.

Assessment of Students With Disabilities

Question

1. Taking into account the diversity of students with disabilities who take the assessments, provide recommendations for the development and administration of assessments for each content area that are valid and reliable, and that enable students to demonstrate their knowledge and skills in core academic areas. Innovative assessment designs and uses of technology have the potential to be inclusive of more students. How would you propose we take this into account?

Response

Using New Technology, Research to Improve the Assessment of Students with Disabilities

Using grounded approaches such as drawing on the principles of evidence-centered design and universal design, lessons learned from states working to meet NCLB accountability mandates, new technologies and a growing body of research, assessment developers can significantly improve the assessment of students with disabilities—and in turn assessment of all students.

Inclusion of students with disabilities in statewide accountability systems had been mandated since the 1997 reauthorization of the IDEA. The IDEA Amendments stated, “children with disabilities must be included in general state and district-wide assessment programs, with appropriate accommodations, where necessary.” Where accommodations alone could not make these assessments accessible, the amendment required that the agencies develop additional assessments so that every child would be included in the accountability programs.

IDEA was primarily responsible for widespread development and implementation of alternate assessments; however, when states became explicitly accountable for the achievement of students with disabilities under NCLB, the attention on those assessments greatly increased. NCLB contained additional mandates requiring that all students be provided access to and assessed against the state curriculum, and that all assessments used for NCLB purposes had to meet the standards for technical adequacy. As a result, many states that implemented alternate assessments in response to IDEA 1997 had to re-design their assessments to reflect a curricular focus and improve technical rigor.

Under current NCLB regulations up to 1 percent of a state’s tested population of students can be considered proficient against alternate achievement standards, and an additional 2 percent of the tested population can be considered proficient against modified achievement standards. All states have developed alternate assessments with alternate achievement standards for students with the most significant cognitive disabilities. Far fewer states have developed or are developing alternate assessments with modified achievement standards.

The alternate achievement standards legislation included requirements for identifying the students for whom alternate standards would be appropriate. The Department refers to this population as the most significantly cognitively disabled students (MSCDs). According to the regulations these students had to

have been identified within one or more of the existing IDEA categories and have a cognitive impairment preventing them from attaining grade-level achievement standards, even with the very best instruction. States were made responsible for defining MSCDs and establishing guidelines for individualized education plan (IEP) teams to use in assigning students to assessments holding them to alternate achievement standards. Regulations for the development of alternate assessments with modified achievement standards identified the population for the assessment as students who, even when receiving appropriate instruction, are not yet on target for grade-level proficiency, even though growth may be demonstrated. There is an additional third group of students who do not participate in alternate assessments, but are identified as students with disabilities; many of these students participate in the general assessment with accommodations.

Identification and recognition of these three populations was based on significant amounts of developmental and educational research, yet, research on students with disabilities is still limited, particularly with regard to the cognitive strategies used by such students in knowledge acquisition and demonstration. However, there is research indicating that academic content is appropriate for students with disabilities across the spectrum of cognitive ability (Kleinert, Browder, and Towles-Reeves, 2009; Center for Education Policy, 2009; Towles-Reeves, Kearns, Kleinert and Kleinert, 2009; and Browder, Flowers, and Wakeman, 2008). The question, then, moving forward with common core standards, is not whether students with disabilities should continue to be involved in academic content and assessments based on that content, but rather what lessons can be learned from what has been done thus far under NCLB about how to create the most appropriate assessment systems for the most significantly cognitively disabled students, for students who do not achieve with the same level of mastery or in the same time frame as non-disabled peers, and for students who need accommodations to participate in the general assessment.

A majority of states developed extended learning standards, extended content standards, or some explanation of how academic content could be appropriately and effectively used in instruction for students with significant cognitive disabilities. The documentation of this extended or linked content was used by special education instructors, many of whom had not previously engaged with their state's academic curriculum, as an instructional and assessment resource. Creating extended or linked standards for the common core, and likely for the state curriculums that are developed based on the core, will be a necessary component of developing the alternate assessment. Indeed, it should be a part of the development of the core standards and curriculums. Several resources exist to support the development of such standards including a comprehensive alignment manual from the National Alternate Assessment Center (Flowers, Wakeman, Browder & Karvonen, 2009).

The assessment for students with the most significant cognitive disabilities, based on the linked curriculum, needs to be flexible enough to respond to the extremely varied disabilities in the population, yet robust enough to exhibit comparability, reliability, validity and other psychometric properties. As the development of alternate assessments with alternate achievement standards have evolved, many have begun to resemble one another. Increasingly, states are providing a bank of activities, tasks, or items that are explicitly tied to specific portions of the curriculum. Generally, there is some flexibility in adapting the tasks to the communication level and needs of the students, and a comprehensive rubric for levels of student mastery and expectations of independence. States differ as to whether a score is recorded or whether evidence of the task is collected for later scoring.

Developing a Comprehensive Online Task Bank for Instruction and Assessment

We recommended the development of a comprehensive online task bank, aligned to the core standards and curriculum, that could be used for both instruction and assessment. This allows assessment to be seamlessly integrated into the classroom experience. Accompanying tasks for each curricular area would include professional development materials and videos providing instruction in how to adapt the tasks for students with varying needs, and how tasks can be integrated into the classroom learning for the content area. Teachers would be required to upload evidence from tasks aligned to a portion of the student's grade level curriculum. Although costly, special education teachers from a district, or area of a state, could meet as a group to discuss a pre-selected sample of task submissions that could be used for scorer training and standard setting, and participate in group scoring using standardized rubrics. This experience would serve as a professional development opportunity for teachers to objectively discuss how academic content is being taught and assessed and to receive training in scoring. In addition this experience can serve as professional development, helping teachers calibrate their own scoring in conjunction with their peers. Following the group scoring, teachers would access and score the content they upload for their own students.

A 25 percent audit by LEA would be re-scored by professional scorers to provide external validation of the scoring and provide scoring reliability information. When results of the audits are returned to LEAs, special education teachers could reconvene to discuss the results and audited submissions. Understanding how the LEA scoring compared to the professional scorers can provide an additional moderating influence on individual teacher scoring for the following year. Additionally, the collaborative discussion of the LEA results engages teachers in one another's success, creates a support network for struggling teachers, and fights the sense of isolation that special education teachers feel, many of whom are often segregated from general education teachers in their schools.

There is clear recognition that a significant number of students, both with and without disabilities, are not succeeding on general education assessments, nor are alternate assessments with alternate achievement standards appropriate for them (Bechard and Godin, 2007; Gong, 2007). However, the best method for identifying these students, the most effective instructional interventions, and the most appropriate assessment methodology is still being debated. The Response to Intervention literature, suggests that approximately 2 percent of all students will not master grade-level content with their peers even after intense, research-based, individualized instruction and will not grow at rate commensurate with their peers (Fuchs and Fuchs, 2006). This literature suggests that a small group of students may continue to need modified achievement standards. However, it is possible that with technological advances, they may not need a fully separate alternate assessment. A comprehensive computer-based assessment, such as was described in the General Assessment response, may be sufficiently adaptive to allow nearly all students to participate in the same assessment program. "Adaptive" does not refer exclusively to traditional Computerized Adaptive Testing (CAT) although CAT may prove a valuable component.

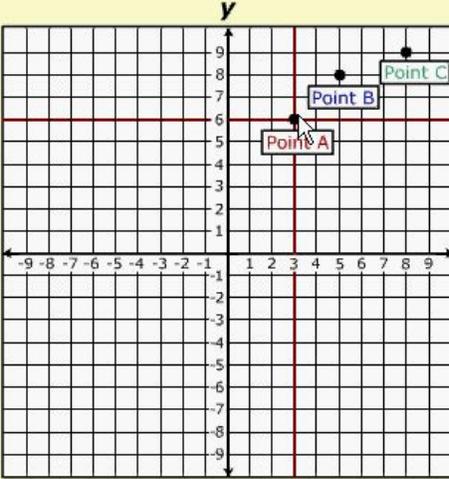
Below is a sample eighth grade mathematics item from the Virginia Modified Achieve Standards Test (VMAST), developed by Pearson for Virginia's Alternative Assessment based on Modified Achievement Standards (AA-MAS). Students can interactively click or mouse-over the points on the grid to have the guides drawn that help locate values on the x and y axes. This is an example of how we can go beyond simple accommodations to address students' challenges or disabilities in higher-order skills such as executive function. While this item was developed as a modification, elements of what was done here would be appropriate as accommodations, or better yet, built-in features of an online testing system.

Click on the points in the coordinate plane to find the x,y coordinates for points A, B, and C.
Which set is the domain of the function?

A. { 3 , 5 , 8 }

B. { 3 , 5 , 6 , 8 , 9 }

C. { 6 , 8 , 9 }



Hint:
Move the mouse over the points in the coordinate plane to highlight the lines of intersection.

Back Reset Go to... Question 2 of 6, Section 1 of 1 Review Next DLA

Joe S Student

Example of Modified Mathematics Item. This online item allows students to interactively click or mouse-over the points on the grid to have the guides drawn that help locate values on the x and y axes.

According to a 2002 Synthesis report published by the National Center of Educational Outcomes, assistive technology supports need to be considered in the design of computer-based tests, including text-to-speech technology, visual highlighting, and page navigation (Thompson, Blount and Thurlow, 2002). Other beneficial supports for students with disabilities that can be added to online assessments include simple scaffolds or structures that provide organizational supports embedded within the items such as highlights, underlines, outlines and other devices that draw the test takers attention to essential information. Current technology has allowed us to build many traditional accommodations directly into the computer-based testing platform, including text-to-speech or audio, flexible font sizes, zooming, flexible color combinations, masking, tagging screen elements and alternative navigation options. Yet, technology can be used to do so much more.

As a mechanism for meeting the needs of students with disabilities not participating in alternate assessments with alternate achievement standards, and the needs of many students in the general population who are not currently succeeding in the general assessment, a computer-based assessment with integrated accommodations may be used, with on-demand scaffolding, and designed from the outset with the principles of universal design. Universal Design for Learning is considered one of the most promising methods of providing access to the general education curriculum for students with disabilities (Hitchcock, Meyer, Rose & Jackson, 2002). However, to realize such an assessment a comprehensive research and development program is needed, focusing on student cognition and the psychometric impact of on-demand scaffolding within an assessment situation.

In looking toward the future of educational testing, the National Research Council (1999) stated that the field “is continually experimenting with new modes, formats and technologies.” O’Bannon, Puckett, and Rakes (2006) concluded that the new technology offers a variety of means for representing the structure of information and the ways in which concepts are related. This fact alone suggests that computer-based or online testing may hold significant promise for students with disabilities.

Moreover, the potential to deploy universal design for large-scale implementation increases with the support of technology—holding significant promise for all students. Technology-based assessments can provide appropriate tools for students to control the user interface. Envision a system of assessments where no longer will assessments have to be “adapted” to accommodate students with special needs as the technology will allow personalized learning experiences for all students regardless of their learning needs.

As a new mode of assessment, online tests have the capacity to incorporate features that will support students’ understanding of assessment items and provide options for them to express what they know. Burk (1999) concluded that test modifications such as large print, audio overlay, and extra spacing were relatively easy to accomplish in computer-based testing programs. Additional features such as hints, interactive graphic organizers, and providing reminders of strategies have been shown to be beneficial to the performance of students with disabilities in preliminary research (Burling and Susbury, 2009).

Understanding the cognitive processes involved in student learning, interacting with items or tasks, and demonstrating their mastery is key to creating an assessment system that can appropriately support each student. Further research is needed to understand these processes in successful students, in low achieving students and students with disabilities. Assessment researchers and developers can build upon cognition research to develop innovative methods of supporting struggling students in assessment situations. While the psychometric implications of such supports are not understood, the prevalence of online learning systems and intelligent tutoring systems could provide a wealth of initial data for exploration. From this starting point, targeted research studies can investigate appropriate psychometric models for such systems and implications for item and test level scoring. Combining such adaptive supports with adaptive algorithms to correlate item difficulty to student ability would create a truly individualized, accessible and appropriate assessment experience for all students.

Even with technology, the needs of all students may not be met in a single computer-based assessment system. There will always be a small number of students for whom the computer-based administration remains inaccessible. Additionally, even with the future potential of refreshable Braille displays, interactive and immersive computer-based environments may be difficult for blind and low-vision students. However, alternative modes should be available for the extremely small number of students who cannot participate in the alternate assessment with alternate achievement standards or the flexible general assessment, potentially with general and modified standards. Some states, such as Virginia, have such an assessment in place for their “gap” students. In these systems, a collection of classroom based evidence of student mastery aligned to the content standards can be submitted for professional scoring. Despite the time and intensity of creating the collection and the cost of scoring, the overall costs of systems are low because few students qualify to participate.

Technology & Innovation in Assessment

Question

1. Propose how you would recommend that different innovative technologies be deployed to create better assessments, and why. Please include illustrative examples in areas such as novel item types, constructed response scoring solutions, uses of mobile computing devices, and so on.

Response

Moving Assessments Online to Advance Instruction and Accountability

The single biggest barrier to developing and implementing the innovative system of common assessments the Department envisions is the current paper-and-pencil based system most states use. Moving online will enable states and schools to increase both the efficiency and the effectiveness of assessments—advancing both instruction and accountability. The Department should encourage and provide incentives for states to use new technologies to support learning and instruction in the classroom. By doing so, it naturally follows that the common assessments should be delivered online as well so that students are assessed in the mode in which they're instructed.

Technology can be better deployed to benefit assessment programs in three fundamental ways:

1. The methods used to administer the assessment
2. The nature of the assessment as a learning experience for students
3. The feedback provided to students, teachers, parents and other education decision makers and stakeholders

Web-Based Assessment Program

A first step in deploying technology for such an assessment system is to migrate the current system to a powerful, easy-to-use, and reliable web-based system. This comprehensive system should seamlessly integrate all major activities involved in the development, delivery, scoring, and reporting of interim, formative, and summative assessments. States such as Virginia and Oregon, have been early adopters of such an approach.

Development

Because the assessment program will be serving a consortium of states, the content management system should be deployed as a web-based service to allow real-time collaboration between and among content providers and reviewers within the consortium. This will allow for greater participation by the states' stakeholders while maintaining a secure and efficient process.

The content management system should do the following:

- Allow tasks such as content authoring, reviews, edits, alignment, and test construction to be performed remotely

- Contain industry-standard data security processes and enforce role-based user management to ensure the integrity of the content management process
- Represent test content in a non-proprietary format, such as the Question-Test Interoperability (QTI) specification, that allows portability to other systems (IMS Global Learning Consortium, 2002)

Delivery

A transition to online delivery is recommended to administer interim, formative, and summative assessments. Until now the choices for assessment design have largely been limited to the following formats

1. Multiple-choice and selected response question formats that can be inexpensively machine-scored
2. Teacher-scored assessments that have questionable reliability, consistency, and equity
3. Constructed response question formats (like essays or short-answer questions) that require human interpretation and scoring

With the next-generation of online innovative assessments this is no longer necessary. Assessments can be designed and developed to more authentically capture broader types of student performance and enable us to measure higher order thinking skills, critical thinking, writing, and the application of knowledge to solve problems—without losing the benefits of lower-cost delivery and scoring and timelier return of information. Admittedly, it is not currently possible to automate the scoring of all complex assessments, but the technology and expertise exists to score many types of assessments and will only improve over time.

For example, Pearson is working with several states to implement the next-generation of secure, online assessment technologies. We have developed a web-based assessment management and delivery platform that supports the next generation of innovative assessment content (PearsonAccess and TestNav 7.0). The system delivers Flash and XML-based interactive, open-ended problems and performance tasks that enable measurement of students' performance applying content knowledge, while using artificial intelligence software to automatically evaluate and score the assessments and provide immediate feedback. Such immediate feedback not only allows for more timely delivery of data in a summative assessment system, but greatly enhances interim and formative systems. For examples of innovative item types, see **Appendix B**.

Interactive items might include the following:

- Items that may have been delivered on paper in a multiple-choice format that can be redesigned more authentically as interactive performance-based items using Flash. Specifically, math content can be made interactive to allow students to graph a formula, interact with three dimensional objects, or demonstrate categorization—task designs that more closely resemble student learning and real-life applications. These items will be more authentic for the students, increasing their motivation and linking directly to their perceived usefulness of the instruction behind the measures.
- Questions requiring a written or constructed response can present a word processor interface to allow students to respond in an environment that more closely reflects real world situations.
- Multi-media elements can be included as stimuli so that the test content is richer and more relevant.

Additionally, the use of technology makes assessments more accessible and more inclusive:

- Test content can be enabled with software-based accommodations, such as tools to assist the visually impaired, to significantly increase the number of students who can access the content and to decrease construct-irrelevant barriers facing these students.
- English language learners can be provided augmented content that more accurately assesses their genuine linguistic abilities.

Assessments should make use of computer adaptive testing when possible to increase efficiency and to gather more precise data across a large range of performance. However, adaptive models will need to make allowances for the full range of item types needed to measure emerging constructs, including those that will be scored by humans. Therefore, computer adaptive assessments may need to be phased into the common assessment system and algorithms will need to be more sophisticated than those that merely select among discrete, multiple-choice items.

Universal design can direct the use of technology to develop tests that are usable, accessible, and accurate for a broad range of students, including those with disabilities and English language learners. Pearson, together with the Center for Applied Special Technologies, has developed initial guidelines (www.pearsonedmeasurement.com/cast/) for application of Universal Design for Learning principles to reduce construct-irrelevance inherent in traditional testing while identifying ways technology can help test all students to greater depths of knowledge and skill.

Portable technology platforms, such as PDAs and smart phones, have great potential for expanding the various environments in which testing can occur. For example, tests such as DIBELS can currently be administered on handheld devices (www.wirelessgeneration.com/solutions/mclass-dibels.html). While the use of such devices in large-scale assessments might not be relevant in the near future, classroom-based interim and formative assessments might benefit from the ubiquity and portability of such platforms, especially in conjunction with the use of equipment, such as equipment used in science laboratories or field projects.

Scoring

Once the foundation of online delivery and innovative content is in place, powerful new scoring options become feasible.

Through the use of a web-based scoring system, distributed teacher scoring can be accomplished using a suite of online scoring monitoring and management tools. These will help improve the overall assessment scoring consistency, accuracy, and efficiency. Pearson uses this technique today in the delivery of essay scoring for the College Board's SAT.

Automated text scoring can also be used to augment teacher scoring of constructed responses. This could further reduce costs and turnaround times, leading to more rapid reporting when time is critical. The automated scoring technology can be adjusted to analyze and evaluate text in various languages and subject areas.

As mentioned previously, Pearson is exploring a significant expansion of the types of constructed response items that can be scored automatically, such as those involving the writing of algebraic expressions and graphing of equations.

Reporting

After the student responses have been scored, innovative reporting solutions shorten the feedback loop and enhance the ability of educators and parents to make instructional decisions based on accurate and relevant data. For example:

- Reports detailing students' individual achievement can be made available via web-based portals that allow students, parents, and teachers to access information on how each student is performing. These reports can also take advantage of cutting-edge technologies to make them easy to interpret and to provide links to educational content that is personalized for individual students, effectively providing them targeted guidance for higher levels of achievement.
- With respect to formative assessments, predictive reporting can be used to project student growth over a period of time.
- Data warehousing techniques can be used to allow educators to mine the assessment data.
- Using the Student Interoperability Framework (SIF) to seamlessly connect the assessment data to the data contained in the school's student information systems facilitates not only reporting on student growth, but the identification of effective teachers, principles, and techniques.

Question

2. We envision the need for a technology platform for assessment development, administration, scoring, and reporting that increases the quality and cost-effectiveness of the assessments. Describe your recommendations for the functionality such a platform could and should offer.

Response

Technology enables assessment to evolve from the static world of paper-based test questions toward more innovative and comprehensive online assessments. A smart technology platform for the online assessment system should accomplish the following:

- Be **accessible** across the spectrum of client devices available within schools. A browser-based system that runs on hardware and software commonly found in schools levels the playing field for schools with limited technology resources (bandwidth, network infrastructure, number of computers, etc.).
- Take advantage of **cloud computing**, enabling school systems to access the assessment system without having to worry about server infrastructure or capacity.
- Provide the **flexibility** to allow year around, on-demand testing. The infrastructure supporting the system must be sufficiently architected and managed for uptime and reasonable response times under load.
- Offer a **single platform** for interim, summative, and formative testing, with **rapid results** that are directly **linked to instructional strategy**.

- Use **standards-based** formats and technologies, such as SIF2, XML, QTI, and Adobe Flash®. A standards-based solution confirms that the system is open and compatible with other systems and the data contained within the system is portable.
- Be **integrated** so that data need only be entered once (to be available throughout the system).
- Be supported by **full-time system and network monitoring** capabilities and dedicated technical support professionals so the administration of the assessments is problem-free.
- Be **secure** to prevent unauthorized access to sensitive information.

The following section outlines in broad strokes several key recommended capabilities of the technology platform with respect to test **development, delivery, scoring, and reporting**.

Development

The assessment system should facilitate the creation and management of the test content. The development solution should:

- Enable **tracking test content** throughout the item lifecycle
- Offer a **distributed workflow** that is web enabled and supports collaboration across geographically distributed participants
- Provide comprehensive data management of **metadata and exposure statistics** associated with the content
- Use a **standards-based content format** (such as QTI) to allow the content to be ported between vendors and delivery systems without requiring complete reformatting
- Allow **real-time editing and rendering**, which saves time by allowing content developers and educators to collaborate in real time to make edits and assess their impact on the content
- **Facilitate test construction** with tools and analytics to construct psychometrically valid tests
- **Be fully integrated** with the other functions of the assessment system to minimize the risk of scoring keys, for example, becoming corrupted when passed between systems

Administration

The assessment administration system should turn the enormous potential of web-based testing into reality. The delivery solution should:

- Use **Student Interoperability Framework (SIF)** technologies to transmit data between existing data sources within schools
- Support a spectrum of **item types**: multiple-choice, short answer, constructed responses, click-and-drag, multi-media (streaming video and audio), interactive Flash-based, etc.
- Provide a full set of **online tools** (e.g., calculator, protractor, ruler, etc.), test taking strategies (answer review, section breaks, etc.), and ancillary testing materials (e.g., math formulas, table of elements, etc.)
- Provide a high degree of **fault tolerance**. No student response data should be lost due to public internet slowdowns, outages, or other local network issues

- **Be secure.** Students testing in a high-stakes environment should not be able to access unauthorized resources (e.g., Internet). All data and test-content transmissions should be encrypted to prevent malicious users from accessing test content or manipulating the results
- Allow assignment and delivery of group-based or individual **accommodations**

Suggesting the new assessment system be available completely online (versus on paper) has significant operational implications. Even with expanded technology in the schools, testing windows will need to be open long enough for students to each have an opportunity to use a computer, and item pools will need to be large enough to protect test security. As the assessment system is phased-in, moving toward an on-demand testing schedule, where students test when they are ready, and not necessarily at the end of the school year, may help lessen technology constraints in buildings as students are testing throughout the school year. However, this will increase the need to have a large enough item bank so as not to compromise test security.

Scoring

Accuracy and timeliness are primary concerns of scoring. The system must be able to score the assessment accurately and consistently. Further, the faster the assessment can be scored, the sooner reports can be generated to facilitate the learning process. The scoring solution should:

- Be **integrated** with the content management system to prevent human errors from causing scoring issues. System-to-system communication of content metadata and scoring keys greatly reduces the likelihood of incorrectly scoring student responses
- Be capable of **scoring tests in real time** to allow for immediate feedback to the student and educators
- Use **machine scoring** to greatly reduce the turnaround time for providing feedback to the learning process
- Be capable of scoring constructed response items either by professional **scorers**, or by **automated text analysis technologies**
- For professional scoring, it is desirable for the system to facilitate scoring over the Internet so that scorers can be **geographically distributed**
- For automated text analysis, the technology be able to understand and evaluate text in **any language** or text in **any subject area**

Score Reporting

Reports should be timely and accurate and in a format easily understood by the target audience—score reports must provide information that is both instructionally actionable and useful for accountability decisions. To measure student growth, it is important that the system be capable of tracking student and assessment data throughout their school careers. The reporting solution should:

- Provide **real-time** and **on demand** reports
- Report student results at **all levels**: student, class, school, district, state, etc.

- Provide a **web-based longitudinal reporting** system that gives educators the ability to analyze test results for a specific administration, for multiple administrations within a year and for year-to-year results
- **Link student achievement to instructional content** in order to provide effective feedback and guidance to the learning process
- Warehouse relevant data to provide the foundation for **data mining** to determine things such as teacher effectiveness and student growth as well as identifying non-trivial factors that impact learning
- Use a consistent data model to enable **student tracking across states**

Question

3. How would you create this technology platform for summative assessments such that it could be easily adapted to support practitioners and professionals in the development, administration, and/or scoring of high-quality interim assessments?

Response

As defined in the RTTT Program regulations and guidance, the term “interim assessments” refers to assessments given at regular intervals designed to measure students’ knowledge of specific academic content standards. These assessments must be designed so that the results can be aggregated at least up to the LEA level and so results for various courses, schools, or LEAs can be compared either to each other or an expected standard of mastery. The results of such assessments could be useful to students, parents, teachers, and administrators for both formative and summative purposes.

The interim assessment should provide evidence of student mastery of academic content that will also be a part of the domain of an end-of-year or end-of-course assessment. Student achievement on the interim assessment, then, could be used to determine which areas of the academic content had been sufficiently mastered and which required more instruction or remediation if the student is to demonstrate mastery on a later assessment. Such information can be used formatively to shape subsequent instruction on an individual student basis. Concurrently, the interim assessment could serve as a summative measure of student mastery after a specific period or unit of instruction, analogous to a chapter test or a mid-term exam. Given sufficient comparability across schools, LEAs, and states, the results of such interim assessments may also be incorporated into accountability measures.

Many of the platform requirements for a comprehensive summative assessment system are the same as those needed for a system of high-quality interim assessments that are developed, administered and/or partially scored by practitioners and professionals at the local level. However, depending on the interim assessment model adopted there are various additional technological requirements for on-demand form development, item development, and instruction in item development and scoring. Interim assessment models can vary from available fixed forms aligned to sections of the core standards or a state’s curriculum, to a mega item bank aligned to standards and curriculum from which states, LEAs, schools, and/or teachers could be trained to assemble forms of varying length on-demand, to a system which supported teachers in writing new items and creating scoring methodologies. As the system moves from less to more individual involvement in the creation of the assessments and their content, the challenges to comparability, and therefore, appropriate aggregation of data, increase. That is not to say that an

individualized interim model with comparable and aggregated data is impossible, just more technically complex from both a technological and psychometric perspective.

Most central to both the summative and interim assessment system is a comprehensive data system that is able to collect and aggregate data across observations, time, students, classrooms, schools, and, at least, LEAs. At a minimum the technology infrastructure described above and a common standard for data would support fixed form interim assessments including both selected and constructed response as well as performance-based innovative interactive items.

To support a mega item bank for local forms development would require an item bank with specifications for each item indicating the standard(s) measured, psychometric properties, and item type (selected response, constructed response, etc.) and tracking capabilities to determine how and where each item is used. Depending on the security of the item bank, such usage statistics might inform item development activities, or be used to verify all items in the bank were receiving exposure. Individuals responsible for building the forms would need a graphical user interface that would facilitate form development. An electronic training module on form development and electronic support, such as a help document or an electronic psychometric avatar, should also be part of the system. To support the ability to aggregate and compare data, some criteria for form design should be fixed, such as reliability or information, test length, and range of item difficulty.

To the extent possible, the interim assessments should take advantage of automated scoring. This will increase the robustness of comparisons and of the aggregation of data. However, participation in distributed scoring and in group scoring of local assessments can be beneficial to teachers for developing their understanding of the content, instructional practices, assessment practices, and judgments of student mastery (Stanley, MacCann, Gardner, Reynolds, & Wild, 2009). At a minimum, the technology requirements mimic those described previously for scoring the summative system. Additional requirements include on-demand electronic scorer trainings, scoring support documentation, and secure methods for uploading and transferring examples of student work or teacher documentation, and an interface for entering scores. Ideally, the technology would be supported by a comprehensive team including scoring trainers and facilitators able to work with states, LEAs, and schools.

Question

4. For the technology “platform” vision you have proposed, provide estimates of the associated development and ongoing maintenance costs, including your calculations and assumptions behind them.

Response

For cost estimates of the development and maintenance of our proposed technology platform, see **Appendix C**.

Project Management

Question

1. Provide estimates of the development, maintenance, and administration costs of the assessment system you propose, and your calculations and assumptions behind them.

Response

Currently state expenditures under the Elementary and Secondary Education Act (ESEA) account for nearly \$1.4 billion annually for activities associated with development, maintenance, and administration for state assessment programs. We have provided cost estimates of the annual costs for a common assessment system for ELA and mathematics at grades 3–8 and costs for end-of-course assessments in high school in **Appendix C**. The Appendix includes assumptions and calculations, which form the basis for this estimate. A couple of the overarching assumptions in this model include the use of a single mode of delivery (online assessment) and the opportunity for teachers to be involved in the scoring of the extended response items.

Question

2. Describe the range of development and implementation timelines for your proposed assessment system, from the most aggressive to more conservative, and describe the actions that would be required to achieve each option.

Response

Proposed Timelines for Assessment System

The Department has encouraged states to adopt common standards by August 2010. Once standards are final and assessment standards are developed, the earliest that field testing could begin for the new summative assessments in spring 2012. An aggressive timeline would include only a single year of field testing followed by full census operational testing in spring 2013. However, from an opportunity to learn standpoint, a second year of field testing would benefit schools and provide two years from the finalization of standards to the implementation of the new summative assessments—fully allowing states and schools to update their curriculum, instruction, and technology. Therefore a more realistic or conservative timeline would include field testing in 2012 and 2013 with operational testing beginning in the spring of 2014.

Introducing the end-of-course assessments may have particular opportunities to learn issues. Because of these, it makes sense to transition these assessments over a period of several years. For example, initial end-of-course assessments in Algebra I, English I, Biology, and World History could be implemented in spring 2014. Geometry, Algebra II, English II, and Chemistry, and US History could be added in spring 2015, and the remaining end-of-course assessments could be implemented in 2016.

Once the new summative assessment system becomes operational there are a number of refinements that could be phased-in over the next five years. These include the following options:

- Moving away from purely summative, end-of-year assessments to on demand assessments that are available online throughout the school year
- Implementing adaptive testing approaches to deliver the assessments
- Strengthening the integration of curriculum, instruction, assessment results, and other data to improve student learning and strengthen teacher capacity
- Tracking and analyzing evidence around college and workplace indicators in the form of a feedback loop to LEAs and to states
- Continuing to expand the use of technology to provide innovative ways to capture student performance, score, report, and integrate data to improve decision-making at all levels of public education

The ability to implement these enhancements will require the states and LEAs to develop comprehensive roll-out plans for aligning curriculum, instruction, and professional development with the revised common standards. In addition, plans will be needed for using data to improve teaching and learning, and technology will need to be sufficient to support online testing.

Question

3. How would you recommend organizing a consortium to achieve success in developing and implementing the proposed assessment system? What role(s) do you recommend for third parties (e.g., conveners, project managers, assessment developers/partners, intermediaries)? What would you recommend that a consortium demonstrate to show that it has the capacity to implement the proposed plan?

Response

There will likely be multiple consortia formed to support common standards. The consortia will vary in terms of development and implementation timeline, use of paper and/or online scoring, approach to end-of-domain or end-of-course testing in high school, use of end-of-year or on-demand testing, and other factors. States may possibly belong to more than one consortium if separate consortia are formed, for example, separating grades 3–8 from high school. There are a number of assessment consortia already in place that states may model their consortiums after (e.g., New England Common Assessments Program [NECAP], American Diploma Project [ADP]). Because the ADP consortium has grown from 9 to 15 states and has increased to include a second end-of-course exam since its inception in 2007, it is worth considering building on this model in the future.

Key elements of the ADP consortium model are described below, in addition it may be useful to review a June 2009 panel presentation at the CCSSO National Student Assessment Conference: [Lessons Learned and the Road Ahead for the American Diploma Project Assessment Consortium](#). (Report located under the heading “Presentations—Educational Assessment Solutions.”)

The ADP Network now includes 35 states dedicated to making sure that every high school graduate is prepared for college or careers. Together, Network states are responsible for educating nearly 85 percent of all US public school students. With increasingly common end of high school expectations among the

states, state education leaders increasingly recognize that collaborative efforts to develop assessments make good policy and economic sense. To that end, 15 Network states in collaboration with Achieve and Pearson formed the ADP Assessment Consortium to develop and adopt rigorous, common Algebra I and II End-of-Course Exams. The ADP consortium includes the following:

- **State Coordination and Direction Team**, which provides the governance structure for the consortium and:
 - Includes assessment directors or other high-ranking policy-making officials from each of the 15 member states in the consortium
 - Oversees production and implementation of the Algebra I and II program
 - Ensures that legal and policy needs of each state are addressed during team deliberations and decision making

The assessment directors each can vote once in decision matters. Contractual matters require unanimous agreement from the states. Noncontractual matters are decided by majority vote, although consensus is preferred. Ohio serves as the lead state for the consortium and contracts directly with the assessment vendor; the other states sign memorandums of agreements to participate.

- **Non-profit advisor/convener/director.** Achieve is a bipartisan, non-profit organization that helps states raise academic standards, improve assessments, and strengthen accountability. Under the umbrella of the ADP, Achieve serves as an advisor/convener/director to the states and manages the assessment vendor. Achieve's responsibilities are to make sure that the goals and objectives of the consortium are implemented with fidelity and to manage the day-to-day oversight of the program quality. New initiatives that the consortium wants to pursue, contract amendments, and the review of documents and test forms are all managed by Achieve. Achieve's project lead works closely with the assessment vendor and the assessment director from the lead state to ensure on-time, quality delivery. Achieve is not a voting member of the consortium and is paid through private funding and the initial memorandum of understanding with the participating states. Funds directly from the testing contract do not flow through Achieve.
- **Vendors.** Pearson is serving as the assessment development and delivery partner for both the ADP Algebra I and II End-of-Course examination programs. Under the contract, Pearson invested in the full cost of development for the exams and retains ownership of the items. States in the consortium purchase exams for each test administration and the pricing is structured to provide a volume-based discount at the consortium-level. Pearson is responsible for item and test development, test administration, scoring, reporting and validity research. A subcontractor develops multiple choice items for the exams. Pearson has a dedicated team of program, test development, research, and technology staff assigned to the program, and Pearson's program manager is the primary point of contact for the states and Achieve.

In addition to its role as the assessment contractor for the Algebra I and II End-of-Course Exams, Pearson is investing in and supporting a research and development agenda to advance the shared goal of strengthening mathematics education specifically as well as the broader vision of moving toward a system of common, rigorous standards and high-quality assessments that help to ensure all students graduate ready to compete and succeed in 21st century global economy.

- **Research Alliance.** Pearson and Achieve assembled an advisory panel of technical experts to gather guidance and recommendations for the research agenda supporting the development and

implementation of the Algebra I and II exams. During the first year of the program, the Research Alliance provided guidance and feedback regarding the validity evidence collected to inform standard setting, which included a college readiness component for Algebra II. The Research Alliance includes up to 15 members, who are experts in the fields of assessment and higher education mathematics. The Research Alliance members are consultants and attend a limited number of meetings during the year. The meetings are lead by Achieve and Pearson and the states are invited to attend the meetings.

- **Teachers.** A large number of Algebra I and II teachers from the participating states have participated in item reviews, data reviews, rangefinding, and standard setting. The teachers are recommended to Pearson by their state assessment directors or content leads.
- **Higher Education.** In addition to higher education involvement through the Research Alliance, a large number of higher education mathematics professionals have been involved in the program from its inception participating in the same types of reviews as the secondary teachers. The higher education participants are often selected by the states and Achieve.

As states consider forming consortia to develop the RTTT common assessments, the ADP Assessment Consortium model offers a good starting point with a few additional considerations:

- Due to size and complexity (e.g., RTTT notice seeks consortia with at least half of the states participating, and will include multiple grades and subjects) the system will likely need to consider including multiple vendors and likely need a prime contractor.
- A more formal and detailed governance structure and mechanism for decision-making, resolving conflicts, and for providing system oversight will likely be required. In addition, the “lead state” model, in which all contracting is done through one state, may be more than any one state is willing or able to support.
- Participating states will need to adopt and implement more consistent policies across states and demonstrate more comparable levels of commitment and participation (e.g., full census testing at grades 3–8 and participation in the end-of-course assessments in high school) to make the consortium successful.
- Steering committees will be required to verify adequate progress in key areas such as technology and innovation, integration of curriculum and assessment, professional development, and use of data for decision making.
- A volume-based pricing model may be replaced with a fixed pricing model if testing is required and not optional—providing cost savings to the states.
- Given the importance placed on college and career readiness, a stronger role may be required for post-secondary and workplace stakeholders.
- As part of their proposals, state consortia should likely include a comprehensive, formal research plan to support the design, development and implementation of new assessments including validity research and program evaluation with both formative and summative research.

References

- Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher*, 33(1), 4-14.
- Achieve (2007). *Aligning High School Graduation Requirements with the Real World: A Road Map for States*. Policy Brief: December 2007.
- ACT (2005). *Crisis at the Core Preparing All Students for College and Work*. Retrieved Nov. 23 from http://www.act.org/research/policymakers/pdf/crisis_report.pdf
- AERA/APA/NCME. (1999). *Standards for educational and psychological testing*. Washington, D.C.: Author.
- Anderson, J. R. (1995). *Learning and Memory*. New York: Wiley.
- Bechard, S. & Godin, K., 2007. Finding the real assessment gaps: A process for states to identify gaps in their assessment systems. New England Compact.
- Berger, K. S. (2007). *The Developing Person through the Life Span* (6th Edition), New York: Worth Publishers.
- Bowles, M., & Stansfield, C. W. (2008). A Practical Guide to Standards-Based Assessment in the Native Language. NLA—LEP Partnership.
- Browder, D. M., Flowers, C., & Wakeman, S. Y. (2008). Facilitating participation in assessments and the general curriculum: Level of symbolic communication classification for students with significant cognitive disabilities. *Assessment in Education: Principles, Policy, and Practice*, 15(2), 137-151.
- Burk, M. (1999). *Computerized test accommodations: a new approach for inclusion and success for students with disabilities*. Washington, D.C.: A.U. Software, Inc.
- Burling, K. & Susbury, S. 2009. Cognitive Interviews Applied to Test and Item Design and Development for AA-MAS (2%). Paper presented at the Council for Chief State School Officers National Conference on Student Assessment. Los Angeles, CA.
- Center on Education Policy, 2009. *State Test Score Trends Through 2007-08, Part 4* Has Progress Been Made in Raising Achievement for Students with Disabilities?
- Conley, David, 2007. *Redefining College Readiness*. Report prepared for the Bill & Melinda Gates Foundation. Eugene, OR: Educational Policy Improvement Center.
- Darling-Hammond, Linda and McCloskey, Laura, 2008. *Assessments for Learning around the World: What would it mean to be "internationally competitive?"* Phi Delta Kappan, Volume 90, Number 4.
- Dolan, R. P., Burling, K. S., Harms, M., Beck, R., Hanna, E., Jude, J., et al. (2006). *Universal Design for Computer-Based Testing Guidelines*. Retrieved May 4, 2009, from <http://www.pearsonedmeasurement.com/cast/index.html>
- Dolan, R. P., & Hall, T. E. (2001). Universal Design for Learning: Implications for large-scale assessment. *IDA Perspectives*, 27(4), 22-25.
- Flowers, C., Wakeman, S., Browder, D., & Karvonen, M. (2007). *An alignment protocol for alternate assessments based on alternate achievement standards*. Charlotte, NC: University of North Carolina at Charlotte, National Alternate Assessment Center.
- Flowers, C., Wakeman, S. Y., Browder, D., & Karvonen, M. (2009). Links for Academic Learning: A conceptual model for investigating alignment of alternate assessment systems based on alternate achievement standards. *Educational Measurement: Issues and Practices*, 28(1), 25-37.
- Fuchs, D., & Fuchs, L.S. (2006). Introduction to response to intervention: What, why, and how valid is it. *Reading Research Quarterly*, 41, 92-99.
- Gong, B.; 2007. Considerations in Designing a "2% Assessment" (AA-MAS) Presentation at USED Alternate Assessment Conference. National Center for the Improvement of Educational Assessment.
- Hargrove, L., Godin, D. and Dodd, B. (2008). *College Outcomes Comparisons by AP® and Non-AP High School Experiences*. College Board: New York.
- Hitchcock, C., Meyer, A., Rose, D., & Jackson, R. (2002). *Technical brief: Access, participation, and progress in the general curriculum*. Peabody, MA: National Center on Accessing the General Curriculum. Retrieved May 20, 2002, from <http://www.cast.org/ncac/index.cfm?i=2830>.

- IMS Global Learning Consortium Inc. (2002). *IMS Question & Test Interoperability Specification*. Retrieved July 20, 2004, from <http://www.imsproject.org/question/index.html>
- Ketterlin-Geller, L. R. (2005). Knowing What All Students Know: Procedures for Developing Universal Design for Assessment. *Journal of Technology, Learning, and Assessment*, 4(2), 1-23.
- Kleinert, H., Browder, D., & Towles-Reeves, E. (2009). Models of cognition for students with significant cognitive disabilities: Implications for assessment. *Review of Educational Research*.
- Koretz, D., McCaffrey, C., & Hamilton, L. (2001). *Toward a Framework for Validating Gains Under High-Stakes Conditions*. Los Angeles: Center for Study of Evaluation, University of California. CSE Technical Report 551.
- Kupermintz, H. (2002). *Teacher Effects As a Measure of Teacher Effectiveness: Construct Validity Considerations in TVAAS (Tennessee Value-Added Assessment System)*. Los Angeles: Center for Study of Evaluation, University of California. CSE Technical Report 563.
- McCaffrey, D., Lockwood, J., Koretz, D. and Hamilton, L. 2003. *Evaluating Value-Added Models for Teacher Accountability*. Santa Monica, CA: RAND Corporation.
- National Commission on the High School Senior Year. *The Lost Opportunity of Senior Year: Finding a Better Way*. Summary of Findings, 2001. Washington, DC: Education Commission of the States.
- National Research Council, 1999. *How people learn: Bridging research and practice*. Washington, DC: National Academies Press.
- O'Bannon, B., Puckett, K. and Rakes, G. (2006). Using Technology to Support Visual Learning Strategies. *Computers in Education*, 23 (1/2) p. 125-137.
- Sanders, W, Saxton, A., & Horn, B. (1997). The Tennessee Value-Added Assessment System: A quantitative outcomes-based approach to education assessment. In J. Millman (Ed.), *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluational Measure?* Thousand Oaks, CA: Corwin Press, Inc., 137-162.
- Solano-Flores, G., Trumbull, E. & Nelson-Barber, S. (2002). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing*, 2(2), 107-130.
- Stanley, G., MacCann, R.; Gardner, J.; Reynolds, L., and Wild, I. (2009). *Review of teacher assessment: Evidence of what works best and issues for development*. Oxford University: Centre for Educational Assessment.
- Stansfield, C.W. (2003). Test translation and adaptation in public education in the USA. *Language Testing*, 20(2), 189-207.
- Stansfield, C.W. & Bowles, M. (2006). Study 2: Test translation and state assessment policies for English language learners. In C. Rivera & E. Collum (Eds.), *State assessment policy and practice for English language learners: A national perspective* (pp. 1-173). Mahwah, NJ: Lawrence Erlbaum Associates.
- Texas Education Agency (2008). *An Evaluation of Districts' Readiness for Online Testing* (Document No. GE09 212 01). Austin, TX; Texas Education Agency, 2008
- Thompson, S., Blount, A., & Thurlow, M. (2002). *A summary of research on the effects of test accommodations: 1999 through 2001* (NCEO Technical Report 34 No. NCEO Technical Report 34). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large-scale assessments* (No. NCEO Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Education Outcomes.
- Towles-Reeves, E., Kearns, J., Kleinert, H., & Kleinert, J. (2009). An analysis of the learning characteristics of students taking alternate assessments based on alternate achievement standards. *Journal of Special Education*, 15(2), 137-151.
- Webb, N. L. (1997). Criteria for alignment of expectations and assessments in mathematics and science education. Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6. Madison: University of Wisconsin, Wisconsin Center for Education Research.
- Wise, L., & Alt, M. (2006). Assessing vertical alignment. Washington, DC: Council of Chief State School Officers.

Appendix A

Definitions

The assessment recommendations from Pearson in this response draw on the definitions provided by the Department as outlined on pages 7-11 in the Race to the Top Application for Initial Funding (<http://www.ed.gov/programs/racetothetop/application.doc>). To help facilitate clarity of meaning for our recommendations, we are providing definitions for a few additional terms:

Construct

An element of knowledge, skill, and/or ability that a test is designed to measure.

Innovative Items

Technology-based test items that use new media and user interfaces to expand the types of stimuli (e.g., videos, animated illustrations) and response modes (e.g., drag-and-drop, interactive simulations) that can effectively be used during assessment. Use of innovative items shows promise for increasing the depth of knowledge and skills to which we can validly assess students, as well as to decrease the impact of construct-irrelevant factors, such as reading ability in non-ELA subject areas, on student performance. (Also see **Performance-Task Items**.)

Performance-Task Items

Test items in which students must respond in more complex ways than for typical selection items such as multiple choice. Performance-task items typically require students to construct responses, with varying degrees of constraint and supports, and as such tap higher-order knowledge and skills, such as problem solving skills and scientific inquiry ability.

Reliability

The degree to which test scores are consistent across replications and thus are dependable for basing inferences about student knowledge and skills.

Validity

The degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests.

Vertical Scale

A method for allowing comparisons of student performance across grades. Scores from separate tests that are vertically scaled can be contrasted to make inferences about student growth in a given subject area.

Appendix B

Description of Performance-Based Assessment Items

The following table summarizes performance-based item types—those that require a more complex student response than multiple choice—and indicates which of the following scoring options are appropriate. These scoring options are defined as follows:

- **Machine scoring.** Simple scoring rubrics are applied through fixed rules automatically by computer, as is currently done for multiple choice items.
- **Automated scoring.** Adaptive algorithms that require human-generated training sets are applied through dynamic rules automatically by computer, as is the case for automated essay evaluation.
- **Human scoring.** Complex scoring rubrics require trained teachers or other qualified scorers.

Performance-Based Item Types and Scoring Methods	
Student Response/Item Type	Scoring Method
Constrained Response <ul style="list-style-type: none"> • Drag-and-drop one or more elements • Select one or more elements • Mark one or more locations (“Hot spots”) 	Machine
Constructed Response <ul style="list-style-type: none"> • Written text (e.g., essay, short answer) • Graphing • Equation/formula construction 	Human readers and/or automated scoring
Simulations <ul style="list-style-type: none"> • Immersive, interactive problems • Multi-step problems • Outcome based responses 	Machine, human readers, and/or automated scoring

Scoring of Performance-Based Items. Performance-based items are grouped in the categories above and require different scoring methods appropriate for the different types of performance-based items.

Sample Items

Samples of each item type are provided in the following pages. Additional sample items, including items that demonstrate interactive functionality, are available on request from Pearson.

Drag and Drop Response

The drag and drop capability allows a student to interactively match responses to a concept. For instance, in a middle school science test, a student can describe the characteristics of a pintail duck by clicking and dragging icons to complete the chart. This feature allows for a deeper assessment of the student's understanding by incorporating grouping, ordering, etc, into the response. These items can be machine scored.

The screenshot shows a digital assessment interface. At the top, there is a toolbar with various tools: Pointer, Eraser, Calculator, Notepad, Highlighter, Straightedge, Eliminator, Exhibits, and Exit Test. The main content area is yellow and contains the following text:

The pintail duck is very common in the United States. An adult pintail duck and a young pintail duck are shown below. Young ducks have some characteristics that are the same as the adult and some that are different.

Five characteristics of pintail ducks are listed. Put all the characteristics that are the same in the box labeled Same. Put all the characteristics that are different in the box labeled Different.

Click on each characteristic you want to select. Then click where you want to put the characteristic.

Characteristics of Pintail Ducks

- Kind of feathers
- Shape of bill
- Webbed feet

Below the text is a table with two columns: "Same" and "Different". The "Same" column has one row with the text "Body size". The "Different" column has one row with the text "Number of legs".

At the bottom of the interface, there is a navigation bar with question numbers 8, 9, 10, 11, 12, and 13. Question 10 is highlighted. To the right of the question numbers, it says "Question 10 of 13" and "Section Review". There are also buttons for "Flag Question for Review", "Previous", and "Next".

Drag and Drop Sample. In this item, the student can describe the characteristics of a pintail duck by clicking and dragging icons to complete the chart.

Select One or More Elements

Pearson's online testing system, TestNav, for example, collects and stores multiple response clicks, which supports the use of items with more than one correct response. These items can be machine scored.

Click the Back button to go back.

After 10 weeks, the student measures the mass of each plant and averages the results for each group. She notices that the more light her pepper plants received, the more mass they had.

Plot 3 points that show a trend for her results. Click on the graph in 3 places to plot the points.

Pepper Growth Experiment

Hours of Light per Day	Mean Mass of Plant
1	2
2	4
3	4

8 9 10 11 12 13

Question 10 of 13

Section Review

Flag Question for Review

Previous Next

WWLastName, WWFirstName | Gr 3 | Section 2

Select One or More Elements Sample. In this example, a student clicks multiple points on a graph to respond to the question.

Mark One or More Locations or “Hot Spots”

Items with hot spots may provide a more interactive testing experience. For example, in a botany test, student knowledge of the functions of various parts of plants is assessed when the student selects various portions of an illustration containing hot spots. These items can be machine scored.

The screenshot displays a digital assessment interface. At the top, a toolbar contains icons for Pointer, Eraser, Calculator, Notepad, Highlighter, Straightedge, Eliminator, Exhibits, and Exit Test. Below the toolbar, a yellow banner reads "Click the Back button to go back." The main content area features two columns of text. The left column says "Choose the part of the plant that takes in the most minerals." The right column says "Click on the diagram to put a '+' on the part of the plant that takes in the most minerals." In the center is a diagram of a tomato plant with a white box labeled "Fruit" pointing to the tomatoes and a white plus sign "+" on the roots. At the bottom, a navigation bar shows question numbers 8, 9, 10 (highlighted), 11, 12, and 13. It also includes "Question 10 of 13", "Section Review", "Flag Question for Review", "Previous", and "Next" buttons. The bottom left corner shows "WWLastName, WWFirstName | Gr 3 | Section 2".

Hot Spots Sample. In this example, a student clicks on the plant diagram to answer the question.

Constructed Response

Constructed response item types may include the following responses:

- Written text (e.g., essay, short answer)
- Graphing
- Equation/formula construction

Depending on the content, constructed response items may allow for a more in-depth assessment of students' abilities. For example, students can demonstrate their work in a math equation by showing the steps taken to arrive at an answer instead of simply selecting a response in a multiple-choice item. This provides more opportunities for students to demonstrate their understanding of various concepts.

Typed responses to constructed response items can be routed automatically for human scoring through an online scoring system. These responses can also be routed for automated scoring, or both human and automated scoring can be used in combination. Pearson uses these capabilities to score formative assessments and portions of the Maryland grade 5 and 8 science assessments.

Below is an example of a constructed response **essay** for a science assessment.

The screenshot shows a web browser window titled "r - 2007 MSA Grade 8 Science". The main content area has a yellow background and contains the following text:

A Sea Wall Just One Molecule High

"There was a large pond, very rough with wind. I dropped a little oil on the water. Though not more than a teaspoonful, it produced an instant calm, [making the water] as smooth as a looking glass."

An incredible experiment, but even more so because of who performed

Compare the properties of oil to the properties of the elements in oil. In your comparison, be sure to include

- the properties of oil
- the properties of the elements in oil
- the motion of the molecules in oil, carbon, and hydrogen

Type your answer in the space provided.

At the bottom of the interface, there are buttons for "Back", "Reset", "Review", and "Next", along with the text "Question 22 of 64".

Essay Sample. In this constructed response format, students respond to the question by writing an essay.

Below is an example of a **short answer** constructed response science item.

r - Grade 5 Science Item Sampler - Pintail ducks

Media

A scientist counted pintail ducks over time. The graph below shows data for a population of pintail ducks in Minnesota. The labels are not included on the axes.

In the boxes, type correct labels for each of the axes.

Y Axis Label

X Axis Label

Year	Population
1992	300
1993	600
1994	150
1995	80
1996	150
1997	80
1998	80
1999	150
2000	150
2001	400

Back Reset Go to... Question 3 of 4 Review Next

Short Answer Sample. In this constructed response format students type short answers in the spaces provided to answer the question.

Response to the Race to the Top Assessment Program Request for Input

r - Feb 2008 ADP Algebra II Student Tutorial

Practice using the tools you were shown. Try adjusting the origin, labeling the axes, and changing the scales.

x Scale: apply y Scale: apply

Back Reset Go to... Next

Graphing Sample. In this practice item, students can use a variety of tools to plot, graph, and label mathematical functions and concepts.

Simulations

Simulation items may include interactive test items, which allow for simulated delivery and response. Multiple steps may be involved, which produce different outcomes, even if the students' response is incorrect to one or more of the steps.

r - Grade 5 Science Item Sampler - Ice cream investigation

The class is going to find out what the salt does to the ice. They will collect water from 2 melting ice cubes. Students put 2 identical ice cubes into 2 funnels. Salt is sprinkled on 1 of the ice cubes. Both ice cubes melt for 5 minutes. The water collects in 2 beakers.

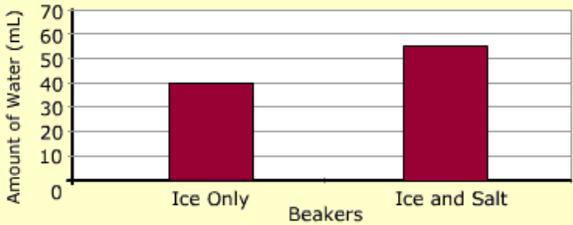
Media



The next day.

Read the water level in the beakers. Make a graph of this data.

Click on a point above each bar where the top of the bar should be.



Beaker	Amount of Water (mL)
Ice Only	40
Ice and Salt	55

Back Reset Go to... Question 3 of 4 Review Next

Simulation Sample. In this example, students can simulate an experiment using the video and graphing tools provided to answer the question.

Appendix C

Race to the Top Assessment Cost Notes and Assumptions

General Assumptions:

- Assume assessment grant awarded in Q4 2010
- Item development will begin in Q1 2011
- Students will be assessed in both Math and ELA in grades 3-8 and in high school through end-of-course tests (EOCT)
- Item types will include multiple-choice and performance-based items with 50 percent of the total score coming from performance-based items. The use of innovative technologies will be leveraged for both item types (e.g., audio streaming, video streaming, drag and drop, simulations, etc.)

Full Census Testing Volumes:

	Total	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	EOCT
50 states & DC	25,835,584	3,627,004	3,585,447	3,601,419	3,659,959	3,715,404	3,764,879	3,881,472
20 states	10,504,540	1,473,201	1,462,112	1,475,925	1,494,976	1,519,513	1,524,047	1,554,766

Research and Development Assumptions:

- Stand-alone field-testing will occur in a four week window in May of 2012 and 2013
- All field-testing will be online, except for Braille and other required accommodations for students with disabilities
- Total test development costs include 2 end-of-course exams, one in ELA and one in a Math course beyond Algebra I; in addition, an estimate is provided for development of end-of-course exams in additional subjects on a per-test basis
- The May 2012 field test will only require 2,500 responses per form
- The May 2013 field test will be administered as a full census field test to the 20 or 50 states participating. Only a portion of the performance items will be scored during this administration
- Research dollars have been included to support the development and ongoing research for the exams

Operational Testing Assumptions:

- Operational testing will begin in the school year of 2013/2014

Response to the Race to the Top Assessment Program Request for Input

- Standard Setting will occur in spring 2014 for all grades and subjects. Reporting will be slightly delayed for this administration to account for this activity
- Annual item release (equivalent of at least one form per year per grade and subject minus linking or embedded field test items)
- The technology platform and the integrated student information system will support test delivery, administration and score reporting for all grades across K-12
- Test Development:
 - There will be 1 form per week per subject and grade if 20 states are testing and 2 forms per week per subject and grade if 50 states are testing
- Administration:
 - While testing will eventually be available on-demand throughout the school year, costing for this effort was for a single 4-week testing window in May
 - Accommodations/accessibility will be built into the items starting with the earliest stages of item development
 - All testing, except for Braille, will be conducted through a secure online delivery platform
 - The annual volumes will match those in the table above for each scenario (20 or 50 states participating)—with each student taking an ELA and Math exam in each grade
 - Costs include translation of Math into 10 languages
 - Costs includes summative testing only at each grade
 - Costs does not include translation or audio options for ELL students
- Scoring:
 - Pearson will score all the field test items
 - Trained and qualified teachers will score all operational performance-based items
 - All grades and subjects are 100% 2nd scored in operational scoring
 - Teachers will be compensated for their effort
- Online Score Reporting:
 - Reporting is entirely online and real-time for Individual Students, School, District, State Reporting
 - Parent Portal website for viewing PDFs of Individual Student Reports
 - On demand reporting allowing for dynamic viewing of student results for teacher, school, district and state
 - Analytic capabilities for further analysis
- Integrated Student Data Information System (includes a comprehensive data portal):
 - Allows educators to mine the assessment data

Response to the Race to the Top Assessment Program Request for Input

- Connects the assessment data to the data contained in the school's student information systems facilitates not only reporting on student growth, but the identification of effective teachers, principles, and techniques
- Facilitates interoperability of data systems and integration of information within and across states in the consortium

Additional Considerations for Consortium States (Not included in total cost):

- Vendor(s)' and states' meeting travel expenses based on our experience with consortium work with ADP
- Statewide technology readiness survey
- Site technology certification to facilitate the transition for schools and districts to online testing

Cost Estimates
Race to the Top Assessment - Grades 3-8 and End-of-Course Test

Research and Development	Year 1 - 2011		Year 2 - 2012		Year 3 - 2013		Year 4 - 2014		Four Year Total	
	Low Range (20 States)	High Range (50 States)								
Test Development*	\$ 13,000,000	\$ 14,000,000	\$ 31,500,000	\$ 47,500,000	\$ 105,000,000	\$ 166,500,000	\$ 17,500,000	\$ 32,500,000	\$ 167,000,000	\$ 260,500,000
Research	\$ 1,500,000	\$ 5,000,000	\$ 1,500,000	\$ 5,000,000	\$ 1,500,000	\$ 5,000,000	\$ 1,500,000	\$ 5,000,000	\$ 6,000,000	\$ 20,000,000
Range Totals	\$ 14,500,000	\$ 19,000,000	\$ 33,000,000	\$ 52,500,000	\$ 106,500,000	\$ 171,500,000	\$ 19,000,000	\$ 37,500,000	\$ 173,000,000	\$ 280,500,000

*ELA and Math tests in each grade for 3-8 as well 2 End-of-Course Tests in High School**; also includes full census field-testing in all grades in 2013

**On average research and development of an online End-of-Course Test with high proportion of performance-based items costs \$5M to \$10M

Student Testing Population (This information was used to populate the table below)	Low Range (20 States)	10,504,540	High Range (50 States)	25,835,584
---	--------------------------	------------	---------------------------	------------

Operational Testing Starting in School Year 2013/2014***	Per Student Cost		Total Cost	
	Low Range (50 States)	High Range (20 States)	Low Range (20 States)	High Range (50 States)
Technology Platform and Secure Online Test Administration	\$ 2.00	\$ 2.75	\$ 28,887,485	\$ 51,671,168
Performance Scoring (Including trained & qualified teachers: 50% of score are performance items)	\$ 37.00	\$ 41.00	\$ 430,686,140	\$ 955,916,608
Integrated Student Data Information System (Including online score reporting as well as a comprehensive data portal)	\$ 1.00	\$ 1.25	\$ 13,130,675	\$ 25,835,584
Range Totals	\$ 40.00	\$ 45.00	\$ 472,704,300	\$ 1,033,423,360

***Costs assume that we would have one integrated online platform for test delivery and score reporting for all of K-12.

Additional Considerations for Consortium States
 Vendor(s)' and States' Meeting and Travel Expenses \$500 - \$650 per participant per day (including lodging)
 Statewide Technology Readiness Survey (per state) \$750,000 - \$1,250,000 (based on number of schools and districts in the state)
 Site Technology Certification \$2,500 - \$5,000 per site





December 2, 2009

The Honorable Arne Duncan
U.S. Department of Education
400 Maryland Ave. S.W., Room 3E108
Washington DC 20202

ATTN: Race to the Top Assessment Program–
Public Input Meetings

Dear Secretary Duncan:

I am writing on behalf of the more than 1.4 million members of the American Federation of Teachers (AFT) to provide our comments on the Race to the Top Assessment Program. The AFT has appreciated the opportunity to participate in the department's recent meetings to provide input on the proposed application guidelines for the next generation of assessments. We also are ready to help in the development process of these assessments, which must be aligned with curricula and reflect a common set of K-12 standards.

The AFT supports the Department of Education's efforts to develop and implement innovative common assessments that will allow students, including English language learners and students with disabilities, to demonstrate their mastery of skills and knowledge at each tested grade level.

The AFT also supports the department's efforts to provide funding for partnerships that develop common assessments based on a common set of standards. However, the standards and assessments alone will not be enough. These are only two pieces, the bookends, of a much more complex, comprehensive system that must include the tools to get it done, specifically: content-rich, sequenced curriculum; standards-based guides for teachers that provide essential background knowledge; model lesson plans that new teachers can teach from and more experienced teachers can draw from as they see fit; pre-service teacher education and in-service professional development that prepare teachers to teach the specific content for which they are responsible; time for teachers to analyze data and collaborate on instructional planning; textbooks that, because they are based on clear standards of reasonable length, are

American Federation
of Teachers, AFL-CIO

AFT Teachers
AFT PSRP
AFT Higher Education
AFT Public Employees
AFT Healthcare

555 New Jersey Ave. N.W.
Washington, DC 20001
202/879-4400
www.aft.org

Randi Weingarten
PRESIDENT

Antonia Cortese
SECRETARY-TREASURER

Loretta Johnson
EXECUTIVE VICE PRESIDENT

VICE PRESIDENTS

Shelvy Y. Abrams
Mary J. Armstrong
Barbara Bowen
Linda Bridges
Kenneth Brynien
Elsie P. Burkhalter
Stacey Caruso-Sharpe
Kathy Chavez
Lee Cutler
Edward Doherty
Kathleen M. Donahue
Thomas A. Dooher
Eric Feaver
Andy Ford
Ed Geppert, Jr.
David Gray
Judy Hale
David Hecker
Richard Iannuzzi
Jerry T. Jordan
Dennis Kelly
Ted Kirsch
Francine Lawrence
Alan Lubin
Louis Malfaro
Merlene Martin
Michael Mulgrew
Maria Neira
Ruby Newbold
Candice Owley
Sharon M. Palmer
Marcia B. Reback
Laura K. Rico
Pat Santeramo
Sandra Schroeder
Phillip H. Smith
Marilyn Stewart
Ann Twomey
Adam Urbanski

slim and focused; teaching and learning conditions where teachers can teach and students can learn; and an accountability system that ensures that these conditions are provided. Without these important components, standards and assessments will have little or no impact on student achievement. Of course, the most important tool is the creation and support of collaborative environments, where the adult educators can really work together to help all kids succeed. We ask that the Department of Education require state applicants to the Race to the Top Assessment Program to incorporate these components in their *standards-based systems* prior to administering assessments and enforcing any form of consequences.

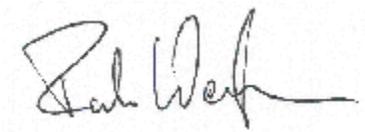
We advise you to require applicants to develop and implement strong *assessment systems* that will provide all students, including students with disabilities and English language learners, an equal opportunity to demonstrate what they know and are able to do. This system should include:

1. The AFT's "smart testing" criteria, which incorporate aligned standards, curricula, assessments and professional development; tests that do not duplicate across education system levels; user-friendly test results; accountability for results; transparency; and appropriate inclusion of English language learners and students with disabilities.
2. Teacher involvement in the development, administration and scoring of assessments as some of our international counterparts now do. This involvement would be supported by aligned high-quality professional development and adequate time for collaboration and data analysis.
3. High school assessments that are fair and provide students multiple ways to demonstrate their knowledge and skills, thereby eliminating the chance of unfairly denying diplomas to students who cannot successfully do so through multiple choice exams.
4. Appropriate content and language-proficiency assessments for English language learners and accommodations that help these students overcome the linguistic barriers that prevent them from demonstrating knowledge of academic content and skills tested.

5. Assessments that incorporate technology, are universally designed, performance-based and embedded in curriculum will increase not only the rates of accessibility and participation of students with disabilities, but also will increase the chance of success for all students.

The AFT hopes that the Department of Education will carefully consider these comments. We look forward to working with the department to implement a program that helps create sustainable change that will improve teaching and learning in our schools.

Sincerely,

A handwritten signature in black ink, appearing to read "Randi Weingarten". The signature is fluid and cursive, with a long horizontal stroke extending to the right.

Randi Weingarten
President, American Federation of Teachers

Recommendations to the U.S. Department of Education's Race to the Top Assessments Program from the American Federation of Teachers

General Assessment Input

The American Federation of Teachers (AFT) has begun advocating for common state standards since 1983 with then AFT president Albert Shanker's response to the landmark report "A Nation at Risk." Today, we continue to advocate for these standards because we believe we must prepare students to succeed in the highly mobile, instantly connected world in which we live. Students must be able to study, work and live in states other than the ones in which they were educated, if they so choose or if circumstances demand it. In the current system, however, individual states develop their own standards and assessments; as a nation, we have failed to develop a system that is fair to all students, teachers and schools regardless of their ZIP codes.

As we stated in testimony before the House Committee on Education and Labor this past April, imagine the outrage if, during the Super Bowl, one football team had to move the ball the full 10 yards for a first down while the other team only had to go seven. Imagine if this scenario were sanctioned by the National Football League. Such a system would be unfair and preposterous. Yet, this is what we currently do in our education system. A report by the Fordham Institute earlier this year concluded, "Schools that make AYP in one state fail to make AYP in another. Those that are considered failures in one part of the country are deemed to be doing fine in another. Although schools are being told that they need to improve student achievement in order to make AYP under the law, the truth is that many would fare better if they were just allowed to move

across the state line.” This type of conclusion highlights the need for common expectations and measures of achievement across state lines.

The AFT has been at the forefront of the standards-based movement because we see the need to ensure that our students are learning what they need to know to compete in a global economy, and the need to address the intolerable and seemingly persistent achievement gap between advantaged and disadvantaged students. In the process, however, we have learned that a conversation about common rigorous standards is too quickly followed by a conversation about a common summative assessment. We believe in accountability, but we caution that standards and assessments are only the bookends of a truly comprehensive standards-based education system. Without the support of aligned curriculum, professional development, time for instructional planning and data analysis, adequate teaching and learning environments, and time for teachers to collaborate, then the bookends have nothing to hold together.

Smart Testing

When used correctly, assessments provide useful feedback about student learning and can guide the system to ensure that schools, teachers and staff get the information they need to help all students meet academic expectations. The AFT has advocated, and continues to advocate for, exams that test what teachers are expected to teach and students are expected to learn. This is what the AFT refers to as “smart testing.” This form of testing is concerned with what is tested and why, whether the testing instruments are up to the task, and how test results are used. It assesses the effectiveness of the curriculum, informs professional development, and provides information to improve teaching and

learning. Smart testing starts with strong, grade-specific state content standards, and includes a number of interrelated pieces:

- Well-developed grade-by-grade curricula;
- Assessments aligned to content standards;
- An efficient, valid and reliable testing system that does not duplicate testing across education system level;
- Appropriate inclusion of English language learners (ELLs) and students with disabilities in testing programs;
- Timely provision of user-friendly testing results for teachers and students;
- Supportive professional development, including coverage of what the content standards are and how they relate to state curricula and assessments, how to teach to the content standards, and how to use testing data to inform instruction;
- Accountability for results; and,
- Transparency of the system.

Some important pieces of the smart testing criteria have been clearly violated or neglected under the current system. Standards are often so broad and ambitious, even in places that have grade-by-grade curricula, the expectations are unrealistic and overwhelming. In focus-group interviews conducted by the AFT, we have heard from many teachers who are currently required to follow guides that pace the curriculum throughout the school year. However, because these guides aim to touch all the required standards, they cover so much material that teachers are concerned about not having time to take advantage of teachable moments for fear of falling behind on the pacing guides. Teachers also mention having to make difficult choices such as taking an extra day or two or three to reteach

material that students have not mastered, at the expense of falling behind on the pacing guide knowing that at year's end, they will be rushed or simply not able to cover all of the required material. Further, some teachers have explained that the pacing guides do not always cover all the standards, and teachers are left on their own to figure out how to cover the additional material in an already overwhelming schedule.

Under NCLB, states are mandated to administer summative assessments once a school year. However, some states and many districts have developed additional interim and/or benchmark assessments resulting in multiple layers of testing at the classroom level. During focus-group discussions, teachers have calculated that up to 25 percent of the school year can be consumed by the summative, interim and benchmark assessments alone. These additional assessments often aim to emulate the summative assessment, so that students are tested and retested on similar material. In other cases, these assessments do not align to the summative assessment, so teachers spend the school year administering assessments and receiving data that do not align or inform progress toward higher achievement on the summative assessment currently used to evaluate schools. This practice does not make the best use of the already scarce instructional time.

Assessments must be aligned to the standards, and those overseeing the development of these assessments must be required to provide public evidence that demonstrates this alignment. This documentation includes such things as item specifications, test specifications, test blueprints, test development reports, or assessments frameworks. This documentation must readily be available to teachers, parents and the general public. In a study conducted by the AFT in

2006 in which we examined the alignment between state standards and state tests, we found that only 11 of 50 states met our criteria for alignment.

Teacher Involvement

A much better approach would be to provide teachers with professional development that trains them on appropriate methods to incorporate formative assessment into their instruction. Teachers do this already when they see a student struggling with an assignment and provide additional help, when they identify a pattern of error and reteach the material to that particular student. But, a standard approach to formative assessment can be provided to teachers through high-quality professional development. Teachers must be at the forefront of assessment literacy. To do this, they must be provided adequate time to engage in meaningful and challenging professional development that focuses on assessment development and assessment literacy. The AFT offers two such courses:

- *Making Data Work for You* is a course to help educators become savvy consumers of data. Developed jointly by the American Federation of Teachers, the New York State United Teachers, the Rhode Island Federation of Teachers and Health Professionals, the Toledo Federation of Teachers and the United Federation of Teachers (in New York City), the course is designed to provide participants with the language, knowledge, and tools to make informed changes—individually and collaboratively in teams—to improve schools, inform and adjust instruction, and advance student learning.

- *Making Classroom Assessments Work for You* helps participants acquire the knowledge and tools they need to understand the role of assessments to improve instruction and student achievement, and to advance learning individually and collectively. The course helps participants better understand how to organize and use standards to help guide instruction and assessment; select, develop and use quality classroom assessments; and plan instruction based on what students need to know and be able to do to meet standards.

The concept of teacher involvement in the assessment process is not groundbreaking. Other nations already have seen its value and have successfully facilitated teacher involvement on a large scale. Singapore uses school-based assessments that include open-ended essays and extended research projects. These assessments are administered and scored by teachers. Teachers are provided high-quality training on how to score these assessments as well as time to score them. In Queensland, Australia, student assessment is based on a partnership between the Queensland Studies Authority and the schools. In this system, teachers develop and score assessments under state guidance. Victoria, Australia, uses a dual approach in which assessments are developed by a central body and teachers use central and school-based tasks to assess student learning. In Alberta, Canada, teachers are involved in developing or piloting assessments, or serving on scoring panels. This level of teacher involvement requires significant amounts of time and resources. However, the investment is well worthwhile: Teachers in these countries report that the training required to accomplish this work is among the best quality of professional development they receive, they also say that they take the strategies of test development and scoring back to their own classrooms for everyday use.

By involving teachers in the scoring process, these regions also have been able to administer assessments that can better measure higher skills and levels of comprehension. These assessments include project-based assessments, open-essay assessments and oral language assessments in multiple languages as is the case in Singapore. Alberta has gone so far as to prohibit the use of multiple-choice questions. By involving teachers in the assessment process, we could improve the quality of our assessment system and potentially eliminate overtesting by providing teachers the training and tools to incorporate effective formative assessments, therefore eliminating the perceived need for benchmarks and interim assessments.

Developing a System that Works

If standards and assessments are to be helpful in improving teaching and learning, they cannot be adopted in a vacuum. For a standards-based system to achieve its goals, which include helping inform instruction, it must consist of:

- Standards that are detailed and explicit and build on knowledge and skills previously acquired as students move through the education system. They must be rooted firmly in subject-matter content and specific enough to lead to a knowledge-rich curriculum that can be mastered during the school year. These standards must pay attention to both content and skills, and must be grade by grade for K-8 and by course at the high school level.
- Curriculum that provides teachers with a detailed road map for helping students reach the standards. The curriculum must focus on the content and concepts to be mastered grade by grade, and include instructional resources, textbooks, instructional strategies, performance indicators, and unit and lesson plans.

- Assessments that provide information on how well the system and/or students are doing and indicate where changes in instructional strategies and resources are necessary if we are to improve learning for all children. Assessments must be aligned to the standards and curriculum, and must be valid, reliable and used for the purposes for which they were designed.
- Accountability , in which all parties are held responsible for providing the supports for student achievement. This includes assisting students who are having difficulty meeting the standards, providing professional development for teachers, and implementing standards for strong teaching and learning environments, as well as having school policies that encourage students to take learning seriously by providing rewards and consequences based, in part, on state assessment results.
- Professional development that is aligned to all other components of the system and helps teachers and other instructional staff deliver the content, differentiate instruction and adjust delivery based on data analysis and best practices, as well as on multiple sources of information about student learning.
- Time for collaboration and data analysis. The system must provide common planning time as well as individual planning time for teachers and instructional staff. This time is essential for educators to share and model lessons; review student achievement data; and discuss how to adapt instruction, planning and assessments to meet the needs of their students.

Both the development and implementation of such a system must be informed by teachers' collective experience and must be supported by teaching and learning conditions that foster student achievement.

Those overseeing the development and implementation of a new system must be required to demonstrate, through transparency, how they intend to develop and implement all the components of a truly comprehensive standards-based system. Transparency would “demystify” how (or if) the pieces connect to function as a unified system. A transparent system is not necessarily aligned, but only with transparency can we determine if the standards, tests and other components of the system all are aligned. A transparent system must provide information to parents, students, teachers and the public about the development, purpose and use of all its components.

Using Student Assessment Data To Evaluate Teachers

The AFT believes there is a place for measures of student learning in a teacher’s evaluation. However, standardized assessments should not be the single or predominant factor in teacher evaluation systems.ⁱ Evaluating individual teachers using their students’ standardized test scores is of serious concern because current testing instruments are limited in their ability to capture the full range of learning, and because of the instability of value-added measures.ⁱⁱ Standardized student achievement tests have not been validated for evaluating teachers. In other words, they were never designed to do so, and using them for this purpose is simply invalid.

Research shows us that even the best value-added models provide measures of student learning that vary enormously from year to year, especially for individual teachers (versus whole school), and even more so for teachers in small classes and in small schools.ⁱⁱⁱ Although test scores may play a role, student achievement should include evidence of growth in knowledge and skills based on

multiple measures. Just as no single measure can evaluate teacher performance, no single measure can or should account for student learning. Some examples of the multiple sources that can provide evidence of student learning include:

- Student performances, group work or presentations scored using a rubric;
- Writing samples;
- Student progress toward targeted learning objectives;
- Portfolios;
- Grades;
- IEP goals and objectives;
- language proficiency goals for English language learners; and
- Student “capstone” projects (e.g., graduation, end-of-course research or thesis paper).

A more meaningful approach to assess student growth would be to collect evidence of learning by examining student work, but this would require states and districts to invest resources in the development of a standardized approach to analyzing student work.

Further, the standard of proof (e.g., regarding accuracy, validity, reliability) when using student achievement data to evaluate teachers will differ depending on the decisions being made. For example, the potential consequences of a teacher’s evaluation vary greatly when that information is used as a basis for determining if the teacher needs targeted professional development versus whether that teacher should be granted tenure. Consequently, the standard of reliability and validity imposed on high-stakes compensation and tenure decisions must be different from and, arguably, higher than what would be necessary when designing targeted professional development programs.

High School Assessments

At the AFT, we believe that in the field of public education it is our job to prepare students for work, college and life. When students complete high school, they must have the knowledge and skills to achieve success in work and college and to lead successful lives and find fulfillment wherever they choose to go or circumstances take them. However, life after high school varies from student to student. In reviewing the end-of-high school standards being developed by the Common Core States Standards Initiative, our teachers have expressed concern over the development of a system that has a single set of expectations for all students. After all, even if the goal stands that we are preparing students for career and college readiness, the question remains: What career and what college? The entrance requirements of an Ivy League school are not the same as those of a local four-year university, which are not the same as those of a community college or career technical training school.

In envisioning a new version of standards-based accountability, it is crucial that we envision a system that is ambitious, yet realistic and fair. There currently are 26 states that deny students a high school diploma on the basis of an exit exam. Many of these states offer alternative paths for students who cannot achieve passing scores. These paths may lead to alternative certificates, such as a certificate of high school completion or a special education diploma—documents that are not always accepted as high school diplomas by postsecondary education institutions—a practice that may leave some students in limbo.

If the goal is to prepare students for work, college and life, then we must use assessments that accurately measure the knowledge and skills needed to succeed

in work, college and life. We recommend learning from our international counterparts and use project-based assessments and open-essay assessments. Rhode Island has been using student portfolios at the high school level. The Education Department should examine that system and others like it to learn from their successes and flaws.

We must envision a system that grants all students multiple forms of documenting and demonstrating their true levels of knowledge and skills. We also must acknowledge that not all students will take the same path after high school, and it is our job to provide all of them with the knowledge and skills they will need for whatever paths they choose, including technical career training, which the AFT has supported for many years.

Input on Assessment of English Language Learners

We must address the growing challenges—from inadequate assessment practices to lack of instructional resources to exorbitant dropout rates—faced by English language learners and the educators who teach them every day. What's most disturbing is that the achievement gap between ELLs and other groups has not dramatically narrowed in decades.

Improving instruction and closing the achievement gap for ELLs largely will depend on the development and proper implementation of high-quality assessments that are aligned to standards, curriculum and instruction as well as to English language proficiency standards. We need to be able to measure both English language proficiency and knowledge of academic content, so that students receive sound instructional attention and educators have accessible data they can refer to throughout the school year.

The Race to the Top grants will be crucial to school reform efforts that include the development of improved assessments for ELLs. Improvements are greatly needed given that current testing practices—which assess ELLs’ content knowledge in English— are often not fair, valid, reliable or appropriate, and make it difficult to distinguish between lack of linguistic abilities in English and learning disabilities or educational progress.

The following are critical elements that states must consider to make sure their assessments provide a valid and reliable measure of what English language learners know and are able to do.

Accommodations

Accommodations for ELLs are necessary to allow these students to participate meaningfully in assessments. These accommodations involve changes to testing procedures, testing materials or the testing environment. Effective accommodations for ELLs address their unique linguistic and cultural-background needs without compromising the test construct. And scores from accommodated tests should be sufficiently similar in scale to the test scores of students who did not take the test with accommodations so that the scores can be compared.

State policies must offer accommodations that help ELLs overcome the linguistic barriers that prevent them from demonstrating the knowledge of academic content and skills tested. Without adequate accommodations, ELL test scores cannot accurately reflect what students know and can do.

Poor assessment practices that do not make use of appropriate testing accommodations for ELLs often result in the misidentification of students, schools and school systems.

Because accommodations were originally developed for students with disabilities, many states have not distinguished between accommodations for ELLs and students with disabilities. In part, this is because the amount of research on accommodations for students with disabilities far outweighs the studies conducted on accommodations for ELLs. Although more ELL-specific research is needed to know which accommodations are consistently the most effective for ELLs, the following accommodations^{iv} generally have been shown to be promising and have positive effects:

- Onscreen or same-page pop-up English language dictionaries/glossaries;
- English dictionaries (if commercial dictionaries are allowed, they should not include definitions or examples that include the answers for particular test items);
- Glossaries (word-by-word bilingual glossaries are more effective than dictionaries because standard dictionary definitions are very difficult for ELLs to understand);
- Side-by-side dual language (Spanish-English) tests;
- Translated (Spanish) assessments for all core content areas except English language arts (especially for students at lower English language proficiency levels and for students who received Spanish instruction in the content assessed);
- A“plain English” version of the test (especially for students at intermediate levels of English language proficiency instructed in English). Plain English text is language that has been modified in its syntax, grammar and vocabulary to avoid ambiguity, colloquialisms or multiple meanings. Although

plain English assessments offer a way to eliminate language, graphics or cultural references that are not directly related to what is being assessed, it should not be applied to authentic literary passages or quotations. (Plain English is also referred to as modified English, simplified English or plain language.)

- Extended time (more effective in combination with a dictionary or glossary).

While an accommodation cannot alter the construct being assessed or provide extra assistance in answering the question, the accommodation must make the content accessible to the student. For example, if a student is asked to calculate the average speed for two trains, it would be appropriate to provide a glossary with basic definitions of the main words—"train," "average" and "speed"—but not have a definition that explains how to calculate an average or the formula for speed. When the words "train," "average" and "speed" are generally defined, it helps an ELL access the meaning of the test item without revealing the answer; if the glossary goes beyond basic explanations, however, it might compromise the validity of the test.

English Language Proficiency, Content Knowledge, and Alignment to Standards and Assessments

Performance on a test given in English will depend largely on the student's level of English language proficiency, as well as on his or her prior formal schooling, age, language of instruction, and type of specialized program the student may be enrolled in. All of these factors must be considered when selecting an assessment and accommodations. As indicated earlier in the type of accommodations listed, some accommodations are more useful than others based

on the level of English language proficiency,^v so this must be taken into account.

The impact that interrupted formal schooling has on achievement is a challenging concern (particularly at the secondary school level) that cannot be overlooked. Students who have missed substantial periods of time in school can be far behind educationally. There is a marked difference between a student who has missed no more than a year of school and a student who has missed much more. If all students with interrupted formal schooling are put in the same category, then an inaccurate picture of achievement will emerge, and the test outcomes will not be clear. In addition to socioeconomic level, ethnic background and other factors, test outcomes (for English language proficiency and content) also should be disaggregated by level of interrupted formal schooling.^{vi}

The following actions are needed to help states improve their assessment practices and the ways in which they test students for English language proficiency and content knowledge:

- Statewide implementation of English language proficiency assessments that are aligned to English language proficiency standards;
- Implementation of uniform, valid and reliable standardized tests of English language proficiency (such as the English language proficiency assessments developed by the WIDA–World-Class Instructional Design and Assessment–consortium of states. These particular assessments are research-based and aligned to English language proficiency standards that have been adopted by the states in the consortium);^{vii}
- Ensuring that English language arts assessments are not used to measure English language proficiency;

- Ensuring that content assessments are matched to a student's level of English language proficiency;^{viii}
- Ensuring that content assessments used for accountability purposes are also a valid, reliable and fair way to assess ELLs;
- Ensuring that English language proficiency standards are aligned with state academic content standards;
- Evaluating the current process involved in developing English language proficiency standards and assessments, and making sure that the process is informed by research and best practices; and
- Evaluating the current process involved in developing and implementing the two types of assessments that ELLs take—English language proficiency and content assessments—and making sure the staff members who are responsible for administering the exams have the preparation and resources to do it effectively.

In the early stages of language acquisition, research indicates that ELLs encode and decode text in English at a lower pace than text in their native language. Further, processing a second language requires very complex memory recall processes, which may be compromised when an assessment is not matched to the student's level of English language proficiency.

Further, if content tests that are not matched to a student's level of English proficiency are used in high-stakes decisions, the results of ELLs who have not reached full proficiency will not be valid. Their scores would be at least as much a product of their language level as of their content knowledge. The toll that a rigorous exam can take on ELLs who have not had enough time to learn the language can have far-reaching consequences.

Content Assessments in Native Languages

Language of instruction or academic language knowledge in the native language are factors in whether ELLs can benefit from a test given in English or their native language. Choosing to administer bilingual or native-language assessments as an accommodation for ELLs is not an easy matter that simply depends on checking a box to indicate the student's native language. Native-language tests do not "appropriately assess" content knowledge if students do not have the requisite academic language proficiency in the native language and/or if they have not been instructed on the core subject in the native language.

The decision to administer a test in the native language must take into account the students' oral proficiency and literacy in their native languages, as well as the language in which they have received core content instruction. Getting an accurate picture of the particular differences within the ELL population is a daunting, yet essential, task as test administrators and school-based decision-makers are faced with selecting appropriate accommodations for individual students.

The following actions are needed to help states improve their native-language assessment practices:

- Develop assessment or survey tools to gauge academic native-language proficiency prior to administering a native-language test; and
- Develop a plan to address the assessment needs of students who speak a language for which there is no test or linguistic accommodation.

Other Issues To Consider

Given the great challenges that currently exist, states will need to make significant investments to institute improved assessment practices that are particular to the needs of ELLs. In addition to the previously mentioned recommendations, states will need to do (or make sure that school districts do) the following:

- Evaluate the validity and reliability of current English language proficiency and content exams used for accountability purposes and to diagnose learning gaps;
- Issue research-based guidelines for the appropriate design and development of content assessments for ELLs; and
- Caution that implementing accommodations does not mean diluting content instruction or not holding ELLs to the same high academic standards as all other students.
- Train and support the staff involved in making state- and district-level assessment decisions about ELLs for accountability purposes;
- Ensure that all teachers who have ELLs in their classrooms are knowledgeable about ELL-specific assessment issues and accommodations, and make sure that teachers have the support and resources to establish frequent communication between all the teachers and staff who are in charge of instructing ELLs;
- Ensure that teachers receive ongoing, job-embedded professional development on assessment issues so that they can improve and better tailor the design of their own classroom-based formative assessments, as well as assess student work and other special project-based collaborative work that form part of a student's academic performance portfolio;

- Ensure that commercially developed assessments and curricula are research-based, have been normed on ELLs, and have a demonstrated track record of effectiveness with ELLs;
- Ensure that students who are ELLs and also have disabilities identified on an IEP also be offered whatever additional special education accommodations are permitted by the state. These accommodations should be selected at the local level by qualified personnel who know the students;
- Begin to phase out paper-and-pencil tests in favor of computerized assessments that are tailored to students' specific needs and skill levels; and
- Secure the wherewithal to carry out this work.

There is much that remains unanswered as to what works. We advise that this process and the development of assessments be informed by the work of researchers who are fully aware of the issues that need to be addressed and are at the forefront of research on ELLs and quality assessments.

Input on Assessment of Student with Disabilities

The American Federation of Teachers applauds the direction that Education Secretary Arne Duncan is taking to facilitate the development and implementation of common high-quality assessments that will improve accessibility and participation of students with disabilities.

Although assessment systems are designed to yield information about individual student progress, instructional effectiveness and alignment of curriculum standards, many systems struggle or fail to provide access to systems in multiple formats;

and many do not secure the very information the system was designed to capture.

AFT believes that modern assessment systems should accurately reflect what students with disabilities have gained from their experiences in school, and how society will benefit from its investments in the futures of these students. Prior to the 1997 reauthorization of the Individuals with Disabilities Education Act, participation of students with disabilities in statewide assessments was minimal, with extensive state-to-state variation. A convergence of two pieces of federal legislation, the No Child Left Behind Act of 2001 and the Individuals with Disabilities Education improvement Act of 2004, increased accessibility and participation requirements of students with disabilities in local and statewide assessment systems. However, the struggle on how to appropriately measure student skills and content knowledge against the backdrop of a complex, global demand for 21st-century skill sets that will ensure success for college preparation or career readiness remains. This has also been a struggle for general education students and English language learners.

To this end, the AFT strongly believes that innovatively designed assessment systems which incorporate the use of technology across content areas have the potential to increase the rates of accessibility and participation of students with disabilities as well as increase the chance of success for general education students and English language learners. We believe such systems should pursue measurements of student outcomes that are:

- Universally designed;
- Performance-based; and
- Embedded in curriculum.

Universally Designed Assessments

Universally designed assessments, like universally designed instruction, come from the universal design theories developed in architecture. In architecture, this approach helps to avoid costly modifications by designing structures with all potential users in mind from the very beginning. In assessments, this approach would allow greater participation of the widest range of students and would provide each student comparable opportunities to demonstrate achievement of the standards being tested.

Performance-based Assessments

Underlying tenets of performance-based assessment are that teachers should have access to information that can provide ways to improve achievement, demonstrate exactly what a student does or does not know or understand, relate learning experiences to instruction and combine assessment with teaching.

Embedded in Curriculum

Assessments embedded in curriculum allow teachers to acquire real-time snapshots of student mastery. It occurs simultaneously with learning such as projects, portfolios and "exhibitions." This supports increased access to curriculum that is relevant and meaningful to students with disabilities.

The AFT looks forward to working with Secretary Duncan to identify programs, develop technology, and collaborate on providing technical assistance and professional development to practitioners who will facilitate access to assessment systems for students with disabilities.

Technology & Innovation Input

Advancements in technology would lend support to development of universally designed assessments that reflect better measurements of student learning, teacher effectiveness and alignment of curriculum. For example, computer-based testing has the potential to increase efficiency and utilization of student assessment systems while decreasing the cost of distributing, delivering, scoring and returning test results. Computer-based adaptive testing also has showed promise in providing immediate feedback to students and teachers. However, the technology platform must be flexible enough to tailor the test-taking experience to meet the needs of individual students, particularly the needs of students with disabilities who require accommodations. Not only will such a technology require a high degree of flexibility and customization, teacher development on how to use the technology is paramount. Technological infrastructures must be acquired and sustained to meet the ever changing needs of diverse student populations.

It would not be fair to spend an entire year using paper and pen, then expect students to be ready for a computer-based assessment. Technology must be infused in the curriculum throughout the school year, and to do this will require bringing our school buildings up to date. As one teacher explained, in her 100-year-old building, she has one power outlet in the classroom. In today's college and working world, we all are surrounded with technology. If we truly intend to prepare our children for this type of college and working world, we must provide them with at least a somewhat similar K-12 learning environment.

ⁱ Relying solely on standardized test scores as the measure of student learning is problematic. According to the National Comprehensive Center for Teacher Quality, defining

teacher effectiveness as a teacher's ability to improve student gains on standardized achievement tests should be avoided because:

- Teachers are not exclusively responsible for students' learning;
- Consensus should drive research, not measurement innovations;
- Test scores are limited in the information they can provide; and
- Learning is more than average achievement gains.

For more information, see Little, O., Goe, L. & Bell, C. (2009). *A practical guide to evaluating teacher effectiveness*. Washington, DC: National Comprehensive Center for Teacher Quality.

ⁱⁱ Value-added measures aren't ready for prime time. Subjective evaluations (such as those done by principals and/or peers) and value-added measures that attempt to identify which teachers are effective can produce results that are very different. (For a discussion of this, see Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2008). *Can you recognize an effective teacher when you recruit one?* Working Paper #14485. Cambridge, MA: National Bureau of Economic Research.) Further, these measures don't tell us anything about *why* teachers vary in effectiveness making it impossible to predict which teachers will be most effective (Goe, L., Bell, C. & Little, O., 2009).

ⁱⁱⁱ See Aaronson, D., Barrow, L. & Sander, W. (2003). *Teachers and student achievement in Chicago public high schools*. Technical report, Federal Reserve Bank of Chicago; Ballou, D. (2005). Value-added assessment: Lessons from Tennessee. In R. Lissetz (Ed), *Value Added Models in Education: Theory and Applications*. Maple Grove, MN: JAM Press; Bos, M. D. McCaffrey, T. Sass, H. Doran, D. Harris, J. Lockwood (2006). *An empirical investigation of the value-added effects of Florida*. Unpublished manuscript submitted to U.S. Department of Education, Institute for Education Sciences; and Goldhaber, D. & Hansen, M. (2008). *Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions*. National Center for Analysis of Longitudinal Data in Education Research (CALDER).

^{iv} In general, the bilingual accommodations (side-by-side dual language tests, translated test, etc.) are only useful for students instructed in the language of the translated assessment.

^v English language proficiency standards are different from English language arts (ELA), or reading standards. English language proficiency standards involve knowledge about *language* knowledge and skills rather than *content* knowledge and skills. English language-proficiency standards should define the knowledge and skills that students need to attain English language proficiency and to acquire academic content knowledge in English.

Therefore, English language-proficiency standards should define, in addition to the language skills and knowledge specific to the needs of ELLs, the academic language necessary for all students to access content in all the other academic content areas.

^{vi} States should consider disaggregating scores for students based on a two-tier system. Tier one would include students who have missed six months to two years of schooling, and tier two would include students who have missed more than two years.

^{vii} If it is not feasible for a state to administer the same test, then the test must, at least, be equated to others being used so that uniform cut scores can be determined for levels of proficiency and reclassification.

^{viii} To the extent feasible that the English language proficiency tests are valid and reliable

December 2, 2009

Subject: Race to the Top Assessment Program

The National Center for Learning Disabilities (NCLD) is a not-for-profit organization founded in 1977 that works to ensure that the nation's 15 million children, adolescents and adults with learning disabilities (LD) have every opportunity to succeed in school, work and life. We work with a national network of more than 40,000 parents, teachers and individuals with LD. Our 32-year commitment to children and adults with LD is based on the guiding principle that federal policies should reflect what research tells us. From research we know that:

- Learning disabilities are neurologically based
- They do not go away
- They affect some 5% of the population
- They require early and accurate identification and effective intervention if students with LD are to succeed in school and life
- 2.6 million students are diagnosed with learning disabilities and receive special education services in our schools, representing 44% of students with disabilities nationwide
- 60% of students with disabilities spend 80% or more of their day in the general classroom
- The majority of students identified with LD have their primary deficit in the area of reading.

As the Individuals with Disabilities Education Act (IDEA) definition of specific learning disabilities stipulates, these students have neurological differences that are *not primarily the result of mental retardation, emotional disturbance, or of environmental, cultural or economic disadvantage*. Additionally, IDEA eligibility determination criteria requires that a student should not be determined to be a child with a specific learning disability if the determinant factor is lack of instruction in reading or math or limited English proficiency. These definitional and qualifying criteria establish students with LD as competent to participate in general education curricula and achieve at a proficient level or higher when provided with high quality instruction by trained professionals as well as appropriate accommodations.

NCLD supports the accountability components of the current ESEA, particularly the expanded assessment and accountability provisions it contains. To that end, we have produced several reports designed to inform parents, educators, policymakers and other stakeholders of the positive impact of these accountability provisions for students with disabilities. Two of these reports are titled *Rewards and Roadblocks* and *Challenging Change*. Additionally, NCLD produced a detailed report examining the current situation regarding testing accommodations for students with disabilities. All are available on our website -- www.ld.org.

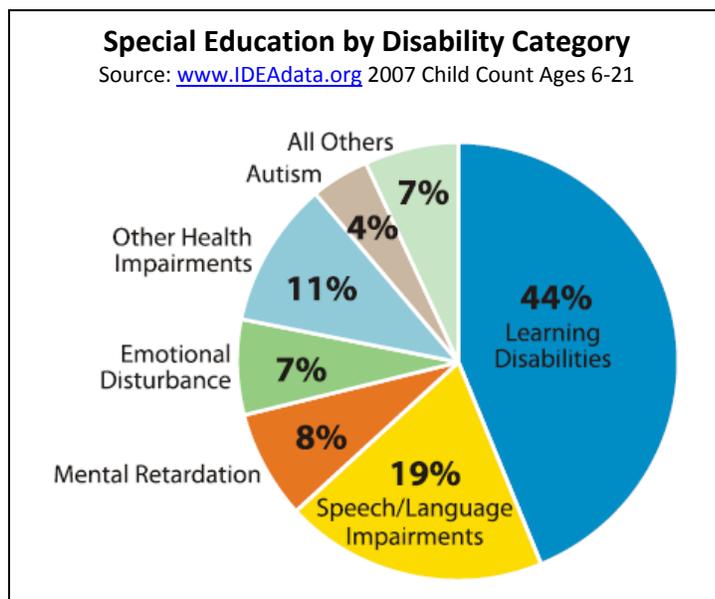
Ensuring that students with LD can participate in large-scale assessments that produce valid and reliable results is a top priority for NCLD. The U.S. Department of Education's initiative to provide funding to consortia of States to develop common, high-quality assessments aligned with a common set of K-12 standards provides an unprecedented opportunity to create equity among diverse learners, including students with disabilities. The next generation of summative assessments must not nibble around the edges of innovation. They must, given our knowledge and expertise and the flexibility provided by technology, facilitate the full and equal participation of all learners.

NCLD appreciates the opportunity to provide comments regarding the design and development of the potential competitions and the notice inviting applications (RFA) that the Department plans to issue by March 2010 as part of the Race to the Top (RTTT) Fund made available by the American Recovery and Reinvestment Act of 2009.

In its Federal Register notice, the Department posed the following question regarding the assessment of students with disabilities:

Taking into account the diversity of students with disabilities who take the assessments, provide recommendations for the development and administration of assessments for each content area that are valid and reliable, and that enable students to demonstrate their knowledge and skills in core academic areas. Innovative assessment designs and uses of technology have the potential to be inclusive of more students. How would you propose we take this into account?

The challenge perceived to be involved in the assessment of students with disabilities (IDEA-eligible students, in this case) is easily brought into focus by the fact that most IDEA-eligible students should participate in the same general assessment taken by all students. The IDEA-eligible designation has been mistakenly perceived as the most salient characteristic of such students. For the approximately 1.6 million students (eligible for special education under the specific learning disabilities category of IDEA and representing 44% of all IDEA-eligible school-age students-see box) who participate in NCLB assessments annually (grades 3-8, 10 or 11) this perception is far from accurate. In fact, most of these students spend most of their instructional day in general education classes with little if any special education supports or services. Most receive some form of special education through a “resource” model of service delivery – a small amount of time a few days per week spend with a special educator in hopes of providing remediation for academic deficits, mainly reading.



Our recent study, *The State of Learning Disabilities 2009*, indicates that most students with LD are not receiving intensive, individualized services adequate to remediate their academic deficits, despite the IDEA definition of “**specially designed instruction**” (“*adapting, as appropriate to the child’s needs, the content, methodology, or delivery of instruction to address the unique needs of the child that result from the child’s disability and to ensure access of the child to the general education curriculum, so that the child can meet the educational standards within the jurisdiction of the public agency that apply to all children*”IDEA Federal Regulations, 2006) . Too often students are accommodated instead of remediated – leading to an over reliance on instructional and testing accommodations. However, this lack of adequate instructional intervention has been highlighted by the accountability provisions of NCLB and has, for many if not all, students with LD, resulted in new attention, increased efforts and improved achievement (see, for example, *Challenging Change: How Schools and Districts Are Improving the Performance of Special Education Students*, NCLD 2007).

NCLD offers the following comments regarding innovative assessment designs and uses of technology to be considered in the assessment grant program design:

- **Require assessments to be designed within innovative test delivery models, particularly computer-based, online delivery systems.** Some advantages of online assessment include:
 - immediate score reporting so test results can guide instruction
 - decreased administrative burdens on school personnel
 - increased security of testing materials, and
 - more flexibility in test scheduling.

- **Require assessment design to incorporate universal design for learning (UDL) principles.** NCLD believes the true solution is to design assessment systems differently from the start, creating them from the outset to be accurate for the widest range of students, including those with disabilities. Universal Design for Learning (UDL) provides the foundation for research-based guidelines for creating flexible and valid on-line, computer-based assessments (see *Universal Design for Computer-Based Testing Guidelines* Pearson Educational Measurement & CAST, June, 2009; <http://www.pearsonedmeasurement.com/cast/index.html>) building upon prior physical and sensory access-oriented Universal Design for Assessment work (Thompson, Johnstone, & Thurlow, 2002).

A UDL approach also offers guidance for enhancing student engagement and persistence. Flexibility in recruiting attention, sustaining effort and supporting self-regulation are all highly individualized and nearly impossible to address without employing the inherent transformability, discrimination and data collection of digital media. The proponents of computer adaptive testing often point to the “automatic” difficulty adjustments of that approach as enhancing student engagement by decreasing the challenge presented to them. This is the same rationale used to support the simplification of the curriculum for struggling students, identical to the “out of level” testing that results in moving students with disabilities further away from the mainstream curriculum. Universal Design for Learning seeks to maintain high achievement standards for all students through the use of customized scaffolds and supports that reinforce the importance of maintaining grade-level expectations for all learners.

While UDL was originally conceived for students with disabilities, NCLD believes it is critical to recognize that UDL can benefit all students. UDL offers a way to design assessments that will accommodate flexible goals and needs for a variety of learners. By presenting material through several means, assessments that are based on UDL allow several types of learners to access the material and demonstrate their knowledge. A UDL approach will eliminate the need for many test accommodations required in traditional testing situations.

- **Require assessments that embed individual student accommodations and allow student control over the test environment.** Researchers have developed systems of online testing environments that provide accommodations that adjust to individual student preferences on demand (such as those developed by Nimble Assessment Systems) as well as online accommodation decision-making tools (such as STELLA developed by Rebecca Kopriva and colleagues at the University of Wisconsin) that increase test validity. Research shows that accommodations delivered within a computer-based testing environment increase the consistency and integrity of accommodations and result in improved utilization by the student. Students should be provided with an optimal testing environment that allows maximum student engagement and persistence.

- **Require states to accept only research-based testing accommodations considered as non-standard.** By “non-standard accommodations” we mean accommodations that influence the target skill, or measured construct, as opposed to standard accommodations that influence an access skill

or non-measured construct. Any accommodation that influences the target skill or the skill measured by the test must be supported by rigorous research evidence. NCLD's report on State Testing Accommodations Policies highlighted the fact that many states are implementing test accommodation guidelines that are not defensible through research. While universally designed tests delivered within online testing environments are sure to eliminate the need for many test accommodations required in traditional tests, some accommodations will continue to be needed by certain students. Common assessments based on a common set of standards can provide for the development of a common set of test accommodations across states. The standardization of test accommodations across states will dramatically improve both the validity and comparability of test results, making test data more useful to educators, parents and policymakers.

- **Require that any “adaptive testing” be aligned with grade-level standards.** While online testing environments hold great promise, they also offer opportunity to lower student expectations through “adaptive” approaches that adjust item difficulty based on student responses. Such approaches are not appropriate for summative assessments used for system accountability. While computer adaptive testing might be useful for formative assessment, its use in summative assessment would surely lead to decreased challenge for some students and a lowering of academic expectations for those students. The current ESEA testing requirements do not allow for “out-of-level” testing. This standard has resulted in the demise of this heretofore-widespread practice for students with disabilities. Today, schools are being held accountable for the performance of students with disabilities on general assessments with only limited exceptions. This positive advancement has resulted in improved access to the general curriculum, expanded learning opportunities and heightened expectations for millions of students. Therefore, any computer adaptive testing developed under this assessment program initiative for use as a summative assessment must be aligned to grade-level academic and performance standards. Exceptions for any subgroup of students - such as students with disabilities and English language learners - should not be permitted within any assessment framework proposed under this program.
- **Require empirical analyses of test items including the study of interactions between specific items and specific student populations.** Items should be analyzed to ensure that they do not disadvantage certain populations of students in their format and/or linguistic complexity. Research studies, such as cognitive labs, should be designed to investigate the interaction between students and test items. Interactions will differ within one broadly defined population of students (for example students with LD); therefore reviewing items in the absence of their specific interactions with students is insufficient. For assessments to provide useful results, all learners and their specific needs must be included in test development procedures, the field-testing of items, and post-hoc analyses of item by student interactions.
- **Require evaluation of test items that ensures total elimination of construct irrelevant, extraneous information.** Work conducted in conjunction with the development of more accessible assessments as well as alternate assessments for students with disabilities has shown that general assessment test items frequently contain irrelevant information that disproportionately impacts students with disabilities, as well as other groups such as English language learners. Recent reviews of large samples of test items used in four statewide general achievement tests have indicated that less than 5% of the items met the criteria to be “maximally accessible for nearly all test-takers” developed by one team of researchers (Elliott, Rodriguez, Roach, & Kettler, 2009). Test item design must give greater attention to the relevancy and necessity of information.
- **Do not require or fund the development of Alternate Assessments Based on Modified Achievement Standards (AA-MAS) through this assessment grant program.** This assessment

option, currently authorized by ESEA regulations promulgated in April 2007 in response to pressure from states, is not supported by empirical evidence. Although there are IDEA-eligible students who are not achieving grade-level proficiency (just as there are many students without disabilities who are not proficient at grade-level), a policy that allows a significant percentage of them to be assessed on other than their enrolled grade-level academic achievement standards remains unjustifiable. Unfortunately, this federal policy is based on research studies on reading intervention that involves cohorts of few if any IDEA-eligible students. Yet this research was used, for political convenience, to justify a policy that is compromising the academic expectations for millions of students.

Through the work of several USED-funded grant programs (both Enhanced Assessment Grants and General Supervision Enhancement Grants) we now know that students with disabilities perform across the proficiency range on state assessments and do not fall consistently at the low end of the proficiency scale. Thus, the group of students who might be considered “persistently low performers” contains students of all demographic, racial and ethnic characteristics. We also know from field-testing of test items developed for “modified achievement standards” that both IDEA-eligible students and non-IDEA students respond equally well. This finding indicates that many students benefit from many of the techniques being used to create the AA-MAS-techniques that are drawn from UD, cognitive load theory, and just plain good test design.

Rather than lowering expectations for a substantial group of IDEA-eligible students based on faulty assumptions and irrelevant research, efforts should be made to develop assessments more accessible for all students. While NCLD recommends that development of the AA-MAS not be funded under the Department RTTT program, it is critical to learn from the work done by several states and state consortia under the Enhanced Assessment Grants and General Supervision Enhancement Grants program. To this end, it is critical that the Department analyze and synthesize this work and make it available to RTTT assessment grantees.

- **Require any growth models to include all students.** The Department has stipulated that one of the general requirements of the assessment systems to be designed under the RTTT grant program is measurement of individual student growth. While supportive of the concept of growth and the possible addition of a growth component to an accountability system, NCLD recognizes the difficulties of using growth models for certain populations, including IDEA-eligible students. NCLD was pleased with the original guidance issued by the Department regarding growth model pilots. However, those guidelines were not upheld in the approval process. The Department has, in fact, approved growth model pilots that do not include all students with disabilities. **Earlier this year, NCLD submitted a comprehensive set of recommendations to the Department and the Congress regarding growth models and IDEA students.**
- **Require all measures of college or career readiness to include all students.** The Department has stipulated that one of the general requirements of the assessment systems must be measurement of whether each individual student is on track toward college or career readiness by the time of high school completion. NCLD welcomes this new focus on post-school outcomes for all students. Over the past two reauthorizations of the IDEA, improvements have been achieved that seek to improve post-school outcomes for students with disabilities. However, these improvements have been slow to materialize. In fact, the Department, through its Office of Special Education Programs, has informed States that certain performance aspects of the State Performance Plan, such as the percentage of IDEA students graduating with a regular diploma, do not need to be considered when reviewing and rating LEA performance on IDEA implementation. In response to questions about such guidance, Department officials have responded that goals such as graduation with a regular diploma

are not goals of IDEA, but rather only “*aspirational*”. Therefore, it must be made clear that the college/career readiness expectations of any assessment program apply equally to all students. Allowing certain populations of students, because of services they receive through other programs such as special education to be excluded from these expectations, could be considered a discriminatory action.

NCLD looks forward to working with the Department as it refines the components for the upcoming RTTT assessment competition. Please do not hesitate to call on us for assistance.

Sincerely,



James H. Wendorf
Executive Director

NATIONAL CENTER FOR LEARNING DISABILITIES, INC.

PUBLIC POLICY OFFICE

12523 Summer Place ▪ Oak Hill, VA. 20171 ▪ PH 703-476-4894 ▪ Email: LKaloi@nclد.org

NCLD HEADQUARTERS

381 Park Avenue South ▪ Suite 1401 ▪ New York, NY 10016 ▪ TEL 212.545.7510 ▪ FAX 212.545.9665

* * * *

NCLD works to increase opportunities and improve outcomes for children and adults with learning disabilities (LD) by providing accurate information to the public, developing and disseminating innovative educational programs, and advocating for more effective policies and legislation to help individuals with LD.

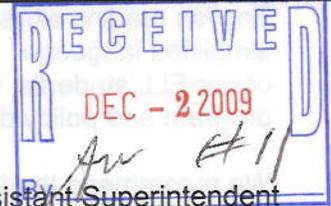


SUPERINTENDENT OF PUBLIC INSTRUCTION

Randy I. Dorn Old Capitol Building · PO BOX 47200 · Olympia, WA 98504-7200 · <http://www.k12.wa.us>

December 2, 2009

Assessment of English Language Learners - Denver, CO



For the record, my name is Joe Willhoft. I am the Assistant Superintendent for Assessment and Student Information in the Washington Office of Superintendent of Public Instruction. I want to thank the Department and its staff for organizing these meetings and for the efficient coordination of the various logistics involved to have an organized set of meetings.

I have a few comments to make regarding the development and administration of assessments for English language learners (ELLs). Washington feels there is a compelling national interest that is served by attending to the accurate and valid assessment of the academic progress of our ELL students. Our commitment to this issue is demonstrated by Washington being the lead state on an IES-sponsored Enhanced Assessment Grant designed to develop validity framework states can use on their English proficiency assessments.

Washington has students in grades K-12 speaking 188 separate languages. Those 88,000 students represent nearly nine percent of the total K-12 enrollment for the state. Approximately one out of every eleven students in our State is an ELL student. Developing 188 separate assessments is unreasonably cost prohibitive. Nevertheless, we have developed approaches we think show promise for the nation.

We currently translate our mathematics and science assessments into the six most-frequently spoken languages. This provides a language-specific accommodation for about 83% of our ELLs. Our mode of test translation is relatively unique. To our knowledge, it is used by only two other states. Our assessment operations contractor provides audio CDs and DVDs which translate the items and directions for students. Students use and provide their responses in English in a standard test booklet. They have the option and flexibility of listening to all or selected items and directions through headphones from a CD player or computer. We find that many of our ELL students prefer to take the standard form of the test, but like the opportunity to have a particular test question, or perhaps the stem for a set of items read aloud to them in their native language. We have found this approach to be very cost effective as separate booklets in multiple languages do not have to be developed, proofed, printed, and distributed. Additionally, the logistics of accommodating a CD are less burdensome than ordering, distributing, and tracking physical booklets in multiple languages. Finally, our ELL students receive much if not most of their instruction in English. Technical content-based vocabulary is not taught to them in their native language, it is taught in English. A physically translated booklet may actually disadvantage ELLs when they encounter technical vocabulary they have learned in English.

As successful as our audio translations are, we still can't accommodate all our ELL students – what about the other 17% who speak a language that is not one of the “top six”? For these students, and for all students, we provide an English-only glossary for our math and science tests that provide very simplified descriptions of non-assessed terms. For example, our mathematics glossary might include an explanation of “school carnival” (a term that might appear in the stem of a problem-solving item), but would not include an explanation of “perimeter”, as perimeter is an assessed term, that students are supposed to know. Making the simplified glossary available to all provides an accommodation for students who lag in their



English reading skills, but does not overly crowd the text in the test itself with an abundance of simplified language. These two approaches have been used in Washington to meet the needs of our ELL students, with the aim of obtaining accurate test scores that can lead to valid program and policy decisions for our schools and districts.

We recommend the Department include specific requirements for the accommodation of ELL students, including but expanded beyond Spanish-speaking students. We further recommend that any across-state efforts include an accommodation plan that can be responsive to the variety of languages found in different states.

Thank you for your time and attention.

Assessment for learning ! AND ! for accountability"

Mark Wilson"
UC, Berkeley"

Presented at the first seminar of the K-12 Assessment and
Performance Management Center at"

ETS, Princeton, NJ"
on October 29-30, 2009. !

Linking summative to formative..."

•! Summative Assessment : A Definition"

An assessment activity is summative insofar as it is being used to provide a summary of what a student knows, understands or can do, and not to help by providing feedback to modify the teaching and learning activities in which the student is engaged.!

Outline"

- ! Linking summative to formative: The wish of every large-scale testing program"
- ! Reversing the “benchmark perspective” on the relationship between large-scale and classroom assessments"
- ! The role of learning progressions and learning performances"
- ! Implementing this new logic for assessment: ! The BEAR Assessment System"
- ! Relating progress variables to learning progressions "
- ! Conclusion and Prospects"

Linking summative to formative..."

•! Formative Assessment: A Definition:"

*An assessment activity is formative if it can help learning by providing information to be used as **feedback**, by teachers, and by their students, in assessing themselves and each other, to modify the teaching and learning activities in which they are engaged."*

The wish of every large-scale testing program"

- ! *To have the results of the large-scale tests be useful “diagnostically” to teachers in the classroom!*
- ! Asked for by State Testing Directors, promised by testing companies ..."

The wish of every large-scale testing program"

- ! A common “solution:” "
 - ! give raw scores for subscales"
 - ! avoid appearance of having to report uncertainty (std errors)"
- ! Another “solution:” "
 - ! make little copies of the state test"
 - ! administer regularly throughout year"
 - ! “microsummative” tests"

Reversing the “microsummative perspective” on the relationship between large-scale and classroom assessments"

i.e., this is the “microsummative perspective” on the relationship between large-scale and classroom assessments:!



E.g., Consequences"

- !“ When it becomes so whittled down to specifics, that’s when it kills... as a teacher I don’t mind marking, its when you are marking in a very narrow way, where you are not allowed to make assumptions, that deadens...Especially with the KS3 tests. You prepare them in a very specific way and you boost them and you give them strategies and you programme them to do things in a certain way. And that is not the way I would naturally teach.” (Kate)"

" " " " " "(from Harrison et al.)"

Reversing the “microsummative perspective” on the relationship between ! large-scale and classroom assessments!



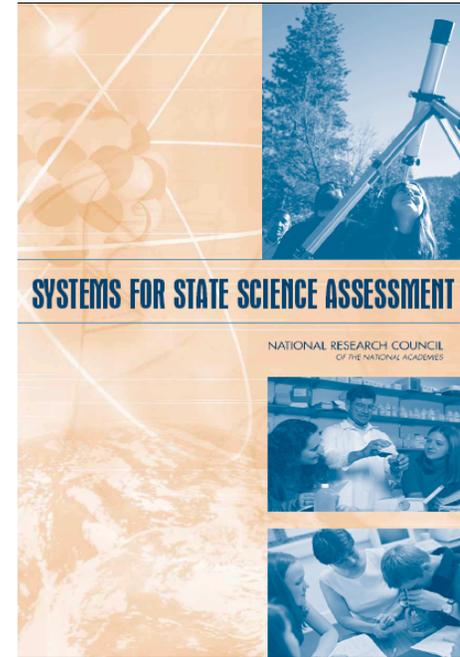
E.g., Consequences"

- ! recent Teachers Network survey"
 - !http://teachersnetwork.org/tnli/survey_highlights.htm"
- ! NCLB testing:
 - !“somewhat useful” 37%, "not at all" helpful 42%
 - !encourages rote drill 40%
 - !eliminate curriculum material not tested 44%
 - !encourages them to improve their teaching effectiveness 3%
 - !an effective way to assess the quality of schools 1%
 - !“strongly agree” that NCLB with its Adequate Yearly Progress (AYP) goals has contributed to teacher burnout 69%"

How do we make this happen?
Thinking about alternatives"

- ! “A mile wide and an inch deep”"
 - ! now-classic criticism of US curricula in Mathematics and Science"
- ! Need to find a more efficient way to use item information than by testing *every* standard with multiple items"
- ! Need for standards to be interpretable by educators, policy-makers, etc."
- ! Need to enable long-term view of student growth"

The role of learning progressions and learning performances!



Learning Performances"

- ! *Learning performances*: a way of elaborating on content standards by specifying what students should be able to when they achieve a standard ("Standards for State Science Systems," NRC, 2005)"
 - ! E.g., students should be able to describe phenomena, use models to explain patterns in data, construct scientific explanations, or test hypotheses"
 - ! Reiser (2002), Perkins (1998) "

Learning performance example"

- ! Benchmark (AAAS, 1993):"
 - ! [The student will understand that] *Individual organisms with certain traits are more likely than others to survive and have offspring*"
- ! LP expansion (Reiser et al, 2003):"
 - ! Students *identify and represent mathematically* the variation on a trait in a population."
 - ! Students *hypothesize* the function a trait may serve and *explain* how some variations of the trait are advantageous in the environment."
 - ! Students *predict, supported with evidence*, how the variation on the trait will affect the likelihood that individuals in the population will survive an environmental stress."
 - ! PLUS sample items, responses, etc."

How to build Learning Progressions and Learning Performances?"

- ! One approach: progress variables"
- ! Detailed assessments based on a cognitive model"
 - !NRC's Assessment Triangle"
- ! Progress variables operate at the *formative* level"
 - !Useful in classroom, as part of instruction, etc."
- ! Learning Progression can be seen as a *bundle* of progress variables"
 - !Operate mainly at the summative level"
- ! One approach to building progress variables..."

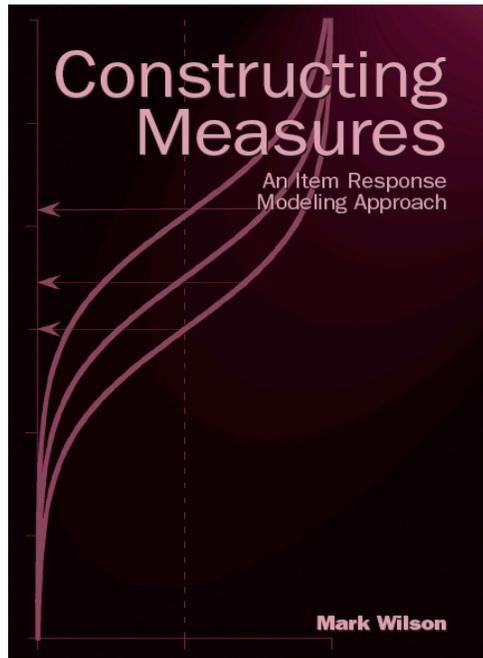
Progress Variables, ctd."

- ! Borrow interpretative and psychometric strength from easier and more difficult items, so that we don't need as many as does the "benchmark approach"."
- ! Progress variables are a principal component of the BEAR Assessment System (Wilson, 2005; Wilson & Sloane, 2000)..."

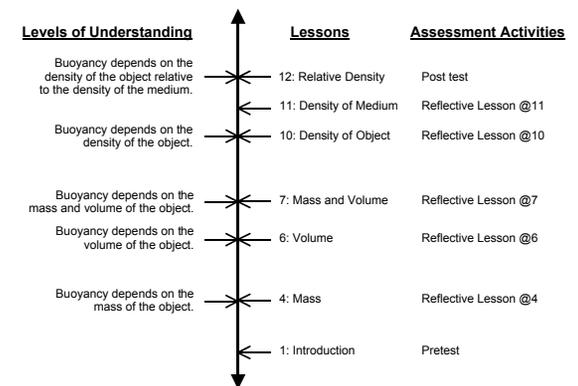
Progress Variables"

- ! *Progress variable*: Assessment expression of a simple and ordered part of a learning progression"
- ! Aim is to combine what we know about "
 - ! (i) learning development, "
 - ! (ii) how items get more complex, and "
 - ! (iii) the patterns of item difficulty "
- ! to make the interpretation of the results more *efficient* and *useful*!

Implementing this new logic for
assessment: !
The BEAR Assessment System!



! —!assessment system should be based on a developmental perspective of student learning ! ! ! !



!

–!there must be a match between what is taught and what is assessed

!

–!a set of principles that allows one to observe the students under a set of standard conditions that span the intended range of the item contexts

!

–!that teachers must be the managers of the system, and hence must have the tools to use it efficiently and use the assessment data effectively and appropriately

!

–!Categories of student responses must make sense to teachers "

Please answer the following question. Write as much information as you need to explain your answer. Use evidence, examples and what you have learned to support your explanations.

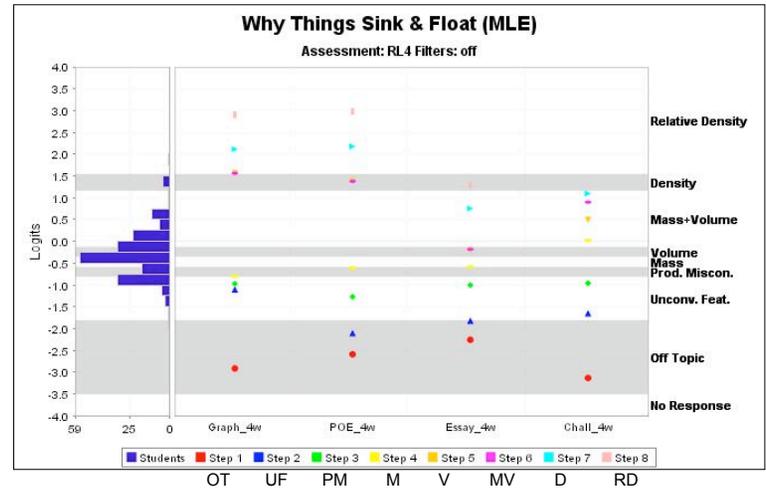
Why do things sink and float?

Level	What the Student Knows	
RD	Relative Density	
D	Density	
MV	Mass and Volume	
M	V	Mass Volume
PM	Productive Misconception	
UF	Unconventional Feature	
OT	Off Target	
NR	No Response	

!
!

-! reliability and validity evidence, evidence for fairness

-! multidimensional item response models, to provide links over time both longitudinally within cohorts and across cohorts



Evaluate a student's locations over time



What might a whole learning progression look like?"

Image of a Learning Progression'

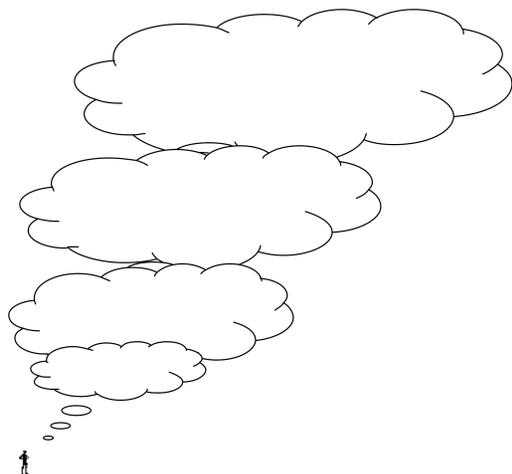
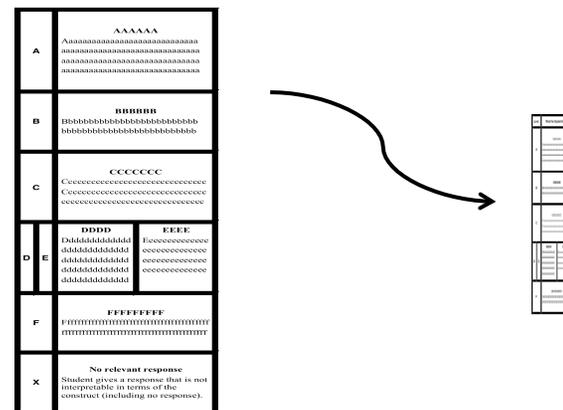


Image of a Construct Map'



Example of a Learning Progression:! Living by Chemistry"

- ! Collaboration with professor in Chemistry Dept. at UC Berkeley: Angy Stacey"
- ! Assessment system for Chemistry from high school through grad school"
- ! Based on concept of 3 "main ideas":
–!Matter, Change, Energy"



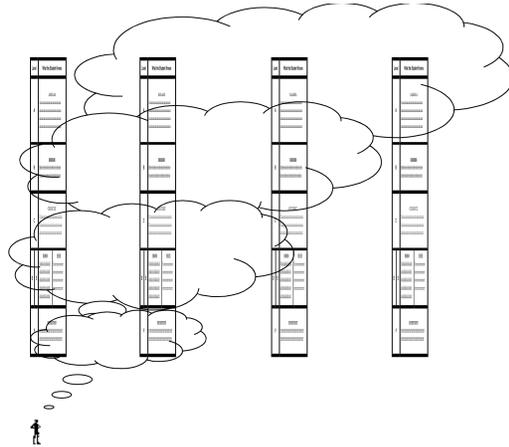
ChemQuery"

Criterion referenced assessments, tracking student learning

ChemQuery Construct"

Student levels of understanding	Matter	Change
III. Formulation	number mole  mass	particulate macro  conservation
II. Recognition	Atomic symbols, octet rule	Chemical equations, conservation of mass (atoms/stuff/grams)
I. Notions	Solid, liquid, gas	Stuff happens

How this might look:!
the levels of the learning progression are!
levels of several construct maps"



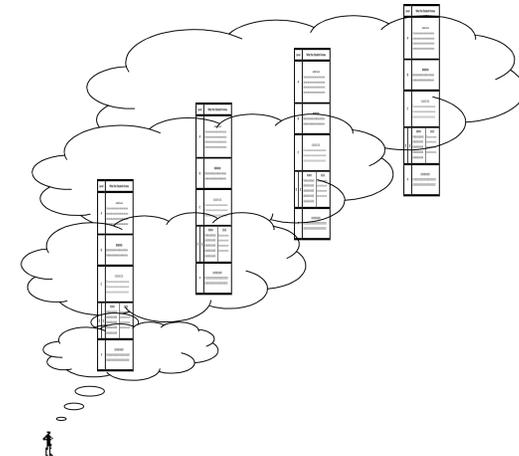
Second example of learning progression:!
Desired Results Developmental Profile (DRDP)"

- ! Collaboration with California State Department of Education"
–!Child Development Division"
- ! Observational instrument for children from birth to kindergarten."
- ! Observers are expected to be their care-givers, and teachers"

Empathy"

I/T		PS		SA	
Developing Ideas	Developing	Integrating	Understanding Someone Else	Considering the Needs of My Community	
Offers comfort to someone showing distress	Offers simple assistance when he or she thinks it is needed- even if not really needed	Uses words or actions to demonstrate concern for what others are feeling	Shows awareness of feelings of others with appropriate words or actions	Shows understanding of feelings and experiences through words or actions for people who live in his or her community (may not know them)	
Discovering Ideas	Exploring	Building	Focusing on Me	Considering Other Perspectives	
Shows concern for others' feelings	Shows awareness when others are unhappy or upset	Accurately labels own feelings, as well as those of others	Demonstrates awareness of own feelings	Shows how someone else might feel in a certain (hypothetical) situation	
Acting with Purpose					
Changes behavior based on others' expressions of emotions					
Expanding Responses					
Shows awareness of others					
Responding with Reflexes					
Responds to others with reflexes					

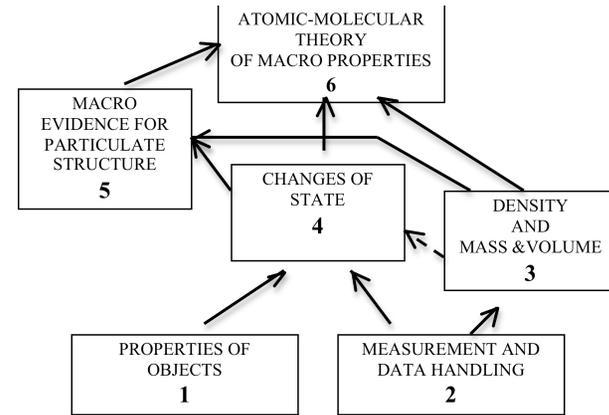
Here the levels are staggered..."



A more complicated relationship"

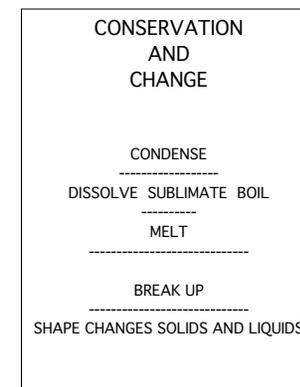
- ! Example drawn from recent work by Black & Wilson, based on literature review by Smith et al (2005)"
- ! Smith, C., Wiser, M., Anderson, C.W., Krajcik, J., and Coppola, B. (2004). *Implications of research on children's learning for assessment: matter and atomic molecular theory*. Commissioned paper prepared for the National Research Council's Committee on Test Design for K-12 Science Achievement, Washington, DC. (http://www7.nationalacademies.org/bota/Test_Design_K-12_Science.html)
- ! Elementary and middle school perspectives on the Atomic-Molecular model"
- ! An image of a learning progression..."

Molecular Theory of Matter"



What is in those boxes?"

- ! A peek at the one in the middle..."
- " Conservation and Change"

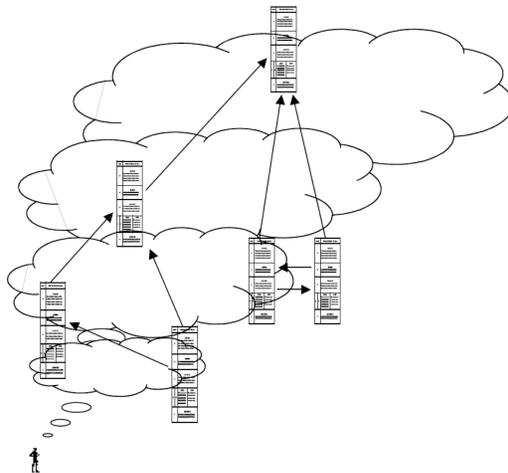


What are the entries in these boxes?"

- ! "Progress variables"!
- ! A peek at the one at the bottom: "
-! SHAPE CHANGES SOLIDS AND LIQUIDS ... "

Level	Description of child understanding
A. Macro changes in shape or size	
A3 Amount Conserving	When an object is cut up or changes shape, it is still the same material, and (altogether) it has the same amount of "stuff" (weighs as much) as it used to.
A2 Material Conserving	When an object is cut up or changes shape, it is still the same material, but (altogether) it has more/less "stuff" (weighs less/more) than it used to.
A1 Magical Thinking	When an object is cut up into or changes shape, it is not still the same material.
B. Micro changes in shape or size	
B4 Micro: Amount Conserving	When an object is cut up into very small pieces, it is still the same material, and (altogether) it has the same amount of "stuff" (weighs as much) as it used to.
B3 Micro: Material Conserving2	When an object is cut up into very small pieces, it is still the same material, but (altogether) it has less "stuff" (weighs less) than it used to.
B2 Micro: Material Conserving1	When an object is cut up into very small pieces, it is still the same material, and (altogether) it has no "stuff" (weighs nothing).
B1 Micro: Magical Thinking	When an object is cut up into very small pieces, it is not still the same material.

A possible diagram'

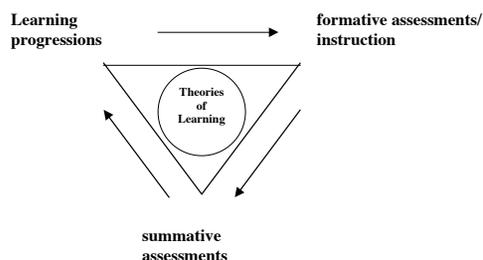


Conclusion:! Thinking about alternatives"

- ! Need for standards to be interpretable by educators, policy-makers, etc."
-! Learning performances"
- ! "A mile wide and an inch deep"
-! Learning progressions as a way to build depth"
- ! Need to find a more efficient way to use item information than by testing every standard with lots of items"
-! Learning progressions can be more efficient"
- ! Need to enable long-term view of student growth ""
-! Learning progressions as a way to enable longer-term thinking"

Conclusion: !

Learning Progression as a Core for Both Instruction and Assessment"



Prospects"

- ! NSF recent *rfps*:"
 - ! Instructional Materials Development (IMD) program includes four components... "
 - ! **(1) Learning Progressions** -- supports the creation of instructional frameworks centered on learning progressions in science and technology education and the development of associated teacher resources and models for professional development."
- ! CPRE-sponsored forums on learning progressions (in both maths and science)"
- ! AERA annual meeting seminars and an NSF-funded meeting next summer ("LeaPS") "

References"

- ! American Association for the Advancement of Science (1993). *Benchmarks for Science Literacy*. New York: Oxford University Press.
- ! Catley, K., Reiser, B., and Lehrer, R. (2005). *Tracing a prospective learning progression for developing understanding of evolution*. Commissioned paper prepared for the National Research Council's Committee on Test Design for K-12 Science Achievement, Washington, DC. (http://www7.nationalacademies.org/bota/Test_Design_K-12_Science.html)
- ! Cowie, B., Moreland, J. & Jones, A.T. (2007). *Bridging the Formative-Summative Divide in Primary Classrooms*. Paper presented at the AERA annual meeting (Chicago).
- ! Harrison, C., Black, P.J., Hodgen, J., Marshall, B. & Serret, N. (2007). *Strengthening Teacher Assessment Practices*. Paper presented at the AERA annual meeting (Chicago).
- ! Reiser, B.J., Krajcik, J., Moje, E., and Marx, R. (2003). *Design strategies for developing science instructional materials*. Paper presented at the National Association for Research in Science Teaching Annual Meeting, March, Philadelphia, PA.
- ! Smith, C., Wiser, M., Anderson, C.W., Krajcik, J., and Coppola, B. (2004). *Implications of research on children's learning for assessment: matter and atomic molecular theory*. Commissioned paper prepared for the National Research Council's Committee on Test Design for K-12 Science Achievement, Washington, DC. (http://www7.nationalacademies.org/bota/Test_Design_K-12_Science.html)
- ! Wilson, M. (2005). *Constructing measures: An item-response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates (<https://www.erlbaum.com/shop/tek9.asp?pg=products&specific=0-8058-4785-5>)
- ! Wilson, M. (2008, March). *Measuring progressions*. In A. C. Alonzo & A. W. Gotwals (Chairs), *Diverse perspectives on the development, assessment, and validation of learning progressions in science*. Symposium conducted at the annual meeting of the American Educational Research Association, New York. Retrieved March 30, 2008, from <http://myweb.uiowa.edu/alonzo/aera2008.html>
- ! Wilson, M. & Bertenthal, M. (Eds.). (2005). *Systems for state science assessment*. Report of the Committee on Test Design for K-12 Science Achievement. Washington, D.C.: National Academy Press. (<http://books.nap.edu/catalog/11312.html>)
- ! Wilson, M., and Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education*, 12(2), 181–208. Available at: http://www.leaonline.com/doi/pdfplus/10.1207/S15324818AME1302_4



December 2, 2009

The Honorable Arne Duncan
Secretary of Education
Attn: Race to the Top Assessment Program - Public Input
U.S. Department of Education
400 Maryland Avenue, SW, room 3E108
Washington, DC 20202.

Dear Secretary Duncan:

We were pleased to see “Innovations for Improving Early Learning Outcomes” in the recently published notice of final priorities for Race to the Top. While we maintain that a state’s commitment to research-based pre-kindergarten education should be a competitive priority, we thank you for taking seriously the comments of the early education community and look forward to working with you on this important issue.

We are writing to comment on another aspect of the Race to the Top competition – the Race to the Top Assessment Program. As pre-k is a fundamental component of our nation’s education system, appropriate standards and assessments should begin at pre-k and be aligned with kindergarten readiness. It is short-sighted for the Department to exclude developmentally appropriate assessments of school-readiness in a competition intended to help states develop “a next generation of assessments.”

We commend you for helping schools “get out of the catch up business” by securing a high-quality early education for all children. Yet, the discussion of measuring school readiness, a key indicator of children’s chance for school achievement, is missing from the public meetings regarding Race to the Top Assessment Program. High-quality pre-k is the most rigorously researched option for school reform, and we believe the Race to the Top Fund can better achieve its goals by incorporating a greater emphasis on this proven education strategy.

We recommend that the Race to the Top Assessment Program include early education as a clear priority by incorporating developmentally appropriate assessments for pre-k. An excerpt from the public comment submitted in August by The David and Lucile Packard Foundation, Buffet Early Childhood Fund, W.K. Kellogg Foundation and The Pew Charitable Trusts provides specific language to fill the gap that currently exists in the competition.

“Standards and Assessments: Voluntary, national standards should begin at pre-k, be grounded in child development principles, and be aligned with national assessments of kindergarten readiness and third grade performance.

“Over 45 states have come together to focus on voluntary, national standards that are benchmarked to high international standards. These standards, thus far, have been defined as K-12. Many states, however, have already developed early learning standards for pre-k that articulate up through third grade. To begin voluntary national standards as late as kindergarten risks abrogating key principles of early childhood development and learning, which span the critical years between the ages of 3 and 8. If the standards only begin at age 5, children’s developmental trajectory and the requisite skills and experiences that all children need before kindergarten to become proficient readers and learners by third grade will be ignored. We urge that the call for voluntary standards begin at pre-k and that the standards are grounded in child development principles. Not doing so risks, at a later time, a standards gap, varying across the nation’s 50 states, between what’s expected in pre-k and what’s expected in kindergarten.

“Like standards, assessments are critical to improved educational outcomes for children and improved state and national results. Given how differently individual children develop in the early years, the field is, quite rightly, concerned about driving individual assessment down into the younger years. Nonetheless, valid, reliable, and developmentally appropriate progress monitoring and assessments of children in the early years should be a critical piece of reform. In particular, the Race to the Top should call for the development of assessments aligned to national, voluntary standards for kindergarten readiness and for third grade reading and math. These measures are key predictors of long-term success. With such assessments and the data that can be used for improvement strategies and achieving quality, the Race to the Top will maximize its chances for long-term success.”

If states are to develop high-quality assessments linked to common K-12 standards, it is imperative that valid, reliable, and developmentally appropriate assessments aligned to standards for kindergarten readiness are a part of the assessment program. We urge you to incorporate the call for assessments aligned to voluntary standards beginning at pre-k in the notice of application for Race to the Top Assessment Program. Please do not hesitate to contact us with any questions. We are happy to work with your staff and to connect them with experts on the technical aspects of early childhood assessment.

Sincerely,



Marci Young, Director
Pre-K Now



Kathy Patterson, Senior Officer
Pew Center on the States

Cc: Thelma Meléndez de Santa Ana, Assistant Secretary for Elementary and Secondary Education

The Pew Center on the States identifies and advances state policy solutions. Pre-K Now, a campaign of the Pew Center on the States, collaborates with advocates and policy makers to lead a movement toward high-quality, voluntary pre-kindergarten for all three and four year olds.

TESTIMONY OF GERALD L. ZAHORCHAK
SECRETARY OF EDUCATION
COMMONWEALTH OF PENNSYLVANIA

December 2, 2009

Having just completed a nearly two-year debate on the role high-quality, standard assessments can play in advancing education reform, the Commonwealth of Pennsylvania is well-positioned to participate in the public input process for the U.S. Department of Education's (USDE) Race to the Top assessment program. We look forward to employing national assessments—both as a way to equalize educational opportunity across 500 diverse schools districts in the commonwealth and as a tool for drawing our state's academic targets even closer to the standards that indicate readiness for the knowledge economy.

In my discussions with policymakers, educators and other stakeholders, the following tenets emerge as priority areas for any national assessment system:

- A comprehensive approach: Assessment systems developed and administered in isolation will not raise student achievement. USDE is to be applauded for its willingness to design assessments within the context of a national common core of academic standards and historic levels of support for strengthening instructional practice. This approach will be well-received in Pennsylvania, where the deployment of new assessments is coupled with supports including a voluntary model curriculum, instructional diagnostic tools, and innovative professional development.
- Multiple measures: Decisions about the academic progress of a district, school or individual student should never be based on a single score or snapshot of data. Indeed, our state's recently-enacted system of high school graduation requirements provides students with several rigorous pathways to a high school diploma, including a proficiency determination informed by both state assessments and local course grades. Pennsylvania believes that state- and locally-developed measures should continue to play important roles in guiding education policy and practice.
- High-quality test construction: To develop an assessment system that can drive genuine improvement in the quality of teaching and learning, USDE must place a priority on the use of test items that require students to analyze complex issues, solve problems, and write persuasively.
- Greater balance between instruction and assessment: USDE should draw on the expertise of the nation's leading psychometricians to craft assessments that can achieve a variety of goals—and therefore reduce testing time. In Pennsylvania, our emerging end-of-course exams are designed to replace existing final exams, satisfy graduation requirements, and serve as the high

school-level assessments for AYP purposes. We would strongly support a similar federal approach that maximizes use of testing time while ensuring reliable and valid results.

The Pennsylvania Department of Education (PDE) argues strongly for an end-of-course model of assessment in the secondary level grades as a way to advance all of these priorities. We feel the end-of-course model—which situates assessment close to the point of instruction—improves integration among teaching, curriculum and measurement and provides both educators and students with valuable, timely feedback.

In addition, the subject-specific nature of end of course assessments allows for deeper, more meaningful measurement of complex problems and tasks. Finally, as mentioned above, end-of-course assessments can replace existing, locally-developed finals, which will yield improvement in the quality of educational measurement at scale without increasing testing time. (And while it is important to acknowledge the difficulty of attaching multiple purposes to a single assessment, the use of end-of-course exams may actually *reduce* testing time.)

PDE’s efforts to institute stronger, more consistent state-level graduation requirements were met with significant opposition from local school boards and special interests. What helped us carry the day was strong support from educators for the end-of-course assessment model that serves as the foundation for the new requirements:

As the superintendent of our state’s largest public school system, an urban district where most of our students are minorities, I [strongly support] the proposed strengthened high school graduation requirements... These tests will serve as an excellent way to measure whether students have met our statewide graduation requirements, are ready to succeed in college or the workplace and will enhance the equity of the academic experience for all public school students across the state.

– Arlene C. Ackerman, superintendent, Philadelphia School District,

End-of-course exams “provide progress monitoring where it counts... in the classroom, where teachers [can] assess instructional effectiveness and student achievement in real time, as students complete their coursework and prepare to move to the next level.”

– Lawrence Korchnak, Superintendent, Baldwin-Whitehall School District

“I believe [end-of-course exams] will provide a more accurate indication of student achievement than the currently-administered [comprehensive assessments]. In addition, these exams will allow districts to evaluate the...rigor of their academic programs in critical content areas.”

– C. Port Williams, Assistant Superintendent, Huntingdon Area School District

This support will ensure that common end-of-course assessments are successfully implemented and that results are used appropriately to inform instructional practice.

Further, I believe a common system of end-of-course exams can support thoughtful cross-state and cross-sector comparisons that can inform stakeholders and provide clear and reliable signals about a student's trajectory toward college- and career-ready skill levels. It is a simple truth that scores from a common Algebra II or Biology assessment have more saliency than results from a comprehensive assessment delivered in relative isolation from a student's course taking. In its 2008 report, *State High School Exit Exams: Moving Toward End-of-Course Exams*, the Center for Education Policy noted that stakeholders see real potential for using end-of-course exams to improve alignment of expectations among K-12, postsecondary institutions, and employers.

PDE recognizes that an end-of-course model of assessment is less appropriate for the elementary grades, and we therefore urge USDE to also support the development of grade-level assessments in reading and math through eighth grade. We believe this hybrid approach will ensure greater consistency in academic expectations at the elementary and middle levels, while allowing for important flexibility that responds to more individualized patterns of course-taking in the high school grades. It should also be noted that this system allows middle school students the opportunity to take end of course exams prior to high school as appropriate for accelerated coursework.

As the efforts to develop the common assessment proposal move ahead, we hope that USDE will continue to gather input from the states and other stakeholders. While Pennsylvania welcomes the opportunity for states to create well-aligned and fiscally-efficient assessments, it is imperative for the USDE to ensure that the formation of state assessment consortia further develop and advance all states' capacity to provide rigorous, high-quality assessments and avoid the possible unintended consequence of "watering down" quality that could occur as states search for common ground or compromises within their efforts to collaborate.

Thank you for this opportunity to provide comment on USDE's common assessment initiative and the potential for this reform to accelerate our state-level efforts.

December 2, 2009

The Honorable Arne Duncan
Secretary
U.S. Department of Education
400 Maryland Avenue SW
Washington D. C. 20202

Re: Race to the Top Assessment Program

Dear Secretary Duncan:

The Council for Exceptional Children – the largest professional organization of teachers, administrators, higher education faculty, researchers and others concerned with the education of children with disabilities, gifts and talents or both – appreciates the opportunity to provide input on the Race to the Top Assessment Program. This issue is of great importance to CEC's 40,000 members who serve on the front lines of educating our nation's 10 million children and youth with disabilities and/or gifts and talents.

The Elementary and Secondary Education Act, as amended by the No Child Left Behind Act, has revolutionized how students with disabilities participate in our national accountability system. NCLB's requirement to disaggregate subgroup data has increased transparency and enabled the public to have better information regarding student performance, especially for those populations who have traditionally been overlooked, such as students with disabilities. This policy change reinforces the need to have high expectations for students with disabilities. CEC urges the Department to reinforce that the creation of new assessments and assessment systems must uphold high expectations for students with disabilities by building on what we have learned about students with disabilities and acknowledging that students with disabilities are general education students first.

While NCLB has increased transparency of the performance of students with disabilities, it has done little to support the education of high achieving students. In fact, by some indications, NCLB has been detrimental to students who are high achieving because the emphasis on reaching proficiency has overshadowed addressing the academic needs of students performing above proficiency.

As the Administration and Congress contemplate reforms to NCLB, CEC supports serious consideration of how the assessment and accountability systems can better compliment each other to ensure that all students receive a challenging, enriched, educational experience that fosters growth and promotes career and workforce preparedness.

CEC believes our accountability system is only as strong as the assessment on which it is based. Therefore, CEC has advocated for revamping current assessments, which take a one-

size-fits all approach to learning. As recent report by the Government Accountability Office pointed out, “cost and time pressures have influenced state decisions about assessment type – such as multiple choice or open/constructed response – and content. States most often chose multiple choice items because they can be scored inexpensively within tight time frames resulting from the NCLBA requirement to release results before the next school year.”¹ Clearly, there is a need for the federal government to support states in constructing high-quality, useful assessments that are driven by the desire to allow students to demonstrate their knowledge and skill rather than cost and time pressures. CEC commends the Administration for investing in the ‘next generation of assessments’, as it has stated, through funding provided by the American Recovery and Reinvestment Act of 2009.

CEC urges the Department to focus on the following areas:

- ◆ Creating assessments that are accessible to diverse learners, including students with disabilities and/or gifts and talents;
- ◆ Creating better Alternate Assessments based on Alternate Achievement Standards (AA-AAS) and Alternate Assessments based on Modified Achievement Standards (AA-MAS); and
- ◆ Creating assessments that provide meaningful feedback to educators and families

Additionally, CEC encourages the Department to consider other initiatives that are integral to having an effective assessment system, such as identifying professional development that is necessary to ensure educators and families understand and can effectively utilize assessments; exploring efforts to scale-up and disseminate promising practices in assessment; and to identify additional opportunities for research.

CEC hopes the Department will take this opportunity to truly move assessments forward by funding future-focused, technology enhanced, accessible assessments to enable our nation’s students to demonstrate their knowledge and skills.

If our comments raise any questions or concerns, please feel free to contact Deborah Ziegler, Associate Executive Director at debz@cec.sped.org or 703-264-9406 or Kim Hymes, Director of Policy and Advocacy at kimh@cec.sped.org or 703-264-9441.

Very Truly Yours,



Deborah A. Ziegler, Ed.D.
Associate Executive Director, Policy and Advocacy Services

¹ Government Accountability Office (September 2009). *No Child Left Behind Act: Enhancements in the Department of Education’s Review Process Could Improve State Academic Assessments*. GAO-09-911.

Creating Assessments that are Accessible to Diverse Learners

CEC urges the Department to fund the creation of assessments and assessment systems with the needs of diverse learners in mind. Current assessments were not created to address the diverse learning needs of students, especially students with disabilities and/or gifts and talents.

As a result, attempts have been made to retrofit assessments with the use of accommodations and other strategies to broaden accessibility for students with a wide range of disabilities. Instead of this piecemeal approach, CEC recommends that the Department fund grants that consider the needs of diverse learners – including, but not limited to, students with disabilities and/or gifts and talents – from the beginning.

Additionally, most current assessments were not designed to accurately reflect the knowledge and skills of students who are gifted and talented because they impose an achievement ‘ceiling’, limiting a student to demonstrating only mastery of grade level content. For some students who are gifted and talented, assessments measuring only grade level content are limiting and prevents educators and parents from receiving accurate data about the capabilities and performance of students performing at above grade level. Such data would allow educators to make modifications to curriculum, instruction, and teacher training necessary to provide appropriate programs and services for our most advanced students.

Specifically, CEC urges the Department to fund grants that create assessments which:

- ◆ Are norm referenced for students with disabilities and/or gifts and talents;
- ◆ Are formative and summative in nature in an effort to provide educators with useful feedback;
- ◆ Take into account accommodations and modifications;
- ◆ Utilize the principles of Universal Design for Learning; and
- ◆ Integrate the use of technology, such as computer adaptive testing

First, assessments created by the RTTT Assessment Program must be norm referenced for students with disabilities and/or gifts and talents during the development of the assessment. Norm Referenced Tests are designed to illuminate achievement differences between and among students across the achievement continuum. This technique provides teachers with very useful information. Specifically, it can help them understand how to group students for instruction based on similar ability levels in certain areas.

Similarly, students with disabilities need accommodations and modifications to deal with their unique learning challenges. Unfortunately, when tests are not designed from the beginning with these in mind, needed accommodations and modifications may be prohibited because they are thought to invalidate the test. This excludes a large population of students from the accountability system, and undermines transparency. CEC believes that by incorporating norm referenced tests and considering accommodations and modifications which may be needed, from the start, we can prevent many of the challenges that we

currently face and more effectively include students with disabilities and/or gifts and talents in the assessment system.

Additionally, CEC encourages the Department to support grants that utilize the principles of Universal Design for Learning, which consist of:

- ◆ Providing multiple means of representation (examples include providing options for how information is perceived and comprehended and how language and symbols are used)
- ◆ Providing multiple means of action and express (examples include providing options for physical action, expressive skills and fluency, and executive function)
- ◆ Providing multiple means of engagement (examples include providing options for recruiting interest, sustaining effort, and self regulation)

While UDL was originally conceived of for students with disabilities, CEC believes it is critical to recognize that UDL can benefit all students. UDL offers a way to design assessments that will accommodate flexible goals and needs for a variety of learners. By presenting material through several means, assessments that are based on UDL allow several types of learners to access the material and demonstrate their knowledge. As NCLB has taught us, one-size-fits-all initiatives are often unsuccessful. UDL offers an antidote to this enforced conformity, which allows a single assessment to address multiple learning needs and provide a better picture of student's abilities. An assessment can only be considered an accurate picture of a student's knowledge and skills if it is designed to allow a student to most effectively demonstrate what they know. Funding grants which incorporate principles of UDL is essential to help reveal a more accurate picture of how all students perform.

Therefore, as the Department moves forward in considering what elements grantees should include in their application, CEC urges the Department to include UDL and utilize the Center for Applied Special Technology (CAST) and the National UDL Taskforce, as valuable resources.

Lastly, CEC urges the Department to fund grants that integrate technology into assessments and assessment systems, such as NimbleTools and computer adaptive assessments. Any such use of technology should incorporate necessary accommodations (i.e. text enlargement, text to speech, etc.) to ensure accessibility to a broad range of learners. Effectively utilizing technology in the classroom and in assessments holds great promise for Secretary Duncan's 'assessment of the future' vision. CEC hopes grants will be funded that support creation, implementation, and professional development for computer adaptive assessments.

Creating better Alternate Assessments based on Alternate Achievement Standards (AA-AAS) and Alternate Assessments based on Modified Achievement Standards (AA-MAS)

CEC believes that students with disabilities must be fully included in assessment systems, and the overwhelming majority of students with disabilities can and should participate in the general assessment. Furthermore, CEC believes that it is critical to maintain the highest expectations for students with disabilities in both assessment and accountability systems,

which are closely intertwined. However, CEC recognizes that for some students with disabilities, the general assessment may not appropriately allow them to demonstrate their knowledge and skill. Instead, these students should have an assessment that best enables them to demonstrate what they know.

For certain students with disabilities, current federal policy allows states to use an alternate assessment based on alternate achievement standards (AA-AAS) and an alternate assessment based on modified achievement standards (AA-MAS). While this policy has been in place for some time, the consistency and availability of these assessments varies widely between states. In fact, the Government Accountability Office, in its report titled *No Child Left Behind Act: Enhancements in the Department of Education's Review Process Could Improve State Academic Assessments*,² identified significant challenges that states report in implementing alternate assessments, such as ensuring validity and reliability for a diverse range of disabilities, increased direct costs of developing and administering such assessments, professional development for teachers, and lack of research about the development of alternate assessments.

A recent study by the National Center for Special Education Research³ within the Institute Of Education Sciences, illustrates these inconsistencies. It concluded that many states approach the AA-AAS differently. Some states use a portfolio or body of evidence to constitute the entire assessment. Others use techniques such as a rating scale/checklist, performance task/events, or multiple choice/constructed response assessments. The inconsistent approach to these assessments across states creates varying standards and expectations and fails to provide the information we need to accurately evaluate the knowledge and skills of students.

Additionally, states, the education, and disability communities are very uncertain about the development of an alternate assessment based on modified achievement standards (AA-MAS). Questions like, who should participate in such an assessment, and what will its impact be on the accountability system remain. Currently, the Department has approved only Texas's AA-MAS, and denied proposals put forth by many states. This uncertainty is also highlighted in a recent white paper commissioned by the New York Comprehensive Center in collaboration with the New York State Education Department titled, *Considerations for the Alternate Assessment based on Modified Achievement Standards: Understanding the Eligible Population and Applying that Knowledge to their Instruction and Assessment*⁴. Key issues addressed in this report summarize the concerns in the field and include: identifying and understanding students who may participate in an AA-MAS, challenges of conceptualizing what low achievers know and how to assess their competence, designing a modified assessment including technical considerations and practical applications.

² Government Accountability Office (September 2009). *No Child Left Behind Act: Enhancements in the Department of Education's Review Process Could Improve State Academic Assessments*. GAO-09-911.

³ Cameto, R., Knokey, A.-M., Nagle, K., Sanford, C., Blackorby, J., Sinclair, B., and Riley, D. (2009). State Profiles on Alternate Assessments Based on Alternate Achievement Standards. A Report From the National Study on Alternate Assessments (NCSER 2009-3013). Menlo Park, CA: SRI International.

⁴ <http://www.cehd.umn.edu/NCEO/AAMAS/AAMASwhitePaper.pdf>

Therefore, in recognition of the inconsistent policies and uncertainty surrounding the AA-MAS, CEC believes the Department should use the RTTT Assessment Program to support research and pilot programs in the development and implementation of assessments so that all students are fully included into the assessment system in a meaningful way.

Creating Assessments that Provide Meaningful Feedback to Educators & Families

As the Department considers its grant proposal, CEC encourages the Department to place a strong emphasis on the importance of creating assessments that yield meaningful information for educators and families. Assessments should be tools that help inform instruction, identify areas of strength and weakness, and help inform decision making. However, assessments can only be effective if they are presented in a way that enables a student to accurately demonstrate their knowledge and skill. Educators need meaningful professional development to help them understand how to use assessment data to inform and drive instruction. Parents need to understand what complex scores show about how their child is learning, and educators must be able to describe results and help parents interpret this complex data meaningfully.

To this end, CEC encourages the Department to fund grants that included professional development and training. Considering how assessments can provide meaningful feedback to educators and parents from the first stage of assessment creation, will help ensure their success.

Conclusion

CEC appreciates this opportunity to provide feedback as the Department moves forward in funding grants through the RTTT Assessment Program. All students will benefit from assessments that allow them to effectively demonstrate their knowledge and skill. Our ability to have a true understanding of how our students are performing depends on having accurate assessments from which to evaluate them by.

Testimony for US Dept of Education

The Role of Public Service Media in a National Assessment System

I am presenting testimony today as both a former Superintendent of Instruction for 10 years in the State of Ohio and the Senior Vice President for Education for the Corporation for Public Broadcasting (CPB), where I serve as chief education policy advisor and consultant to the public service media system. It is my pleasure to provide comments to the U.S. Department of Education regarding the proposed Race to the Top (RTTT) assessment initiative. The Corporation for Public Broadcasting is a private, non-profit corporation that was created by Congress in 1967. It promotes universal access to public telecommunications services (television, radio, and on-line) by supporting over 1100 radio and television stations across America. CPB has a long and well documented record of funding for diverse and innovative educational programming that is second to none. However, beyond programming, public service media helps teachers, caregivers, parents, and communities educate children. CPB is a strong ally in raising the academic bar and closing achievement gaps for all students, particularly the underrepresented and underserved.

Because I have been a policy leader in both public education and now public service media, I understand how our publicly funded television and radio stations can enhance a national system of student assessment. A national assessment system can provide for the better integration of curriculum, instruction, assessment, and educator development, and public service media can provide the digital content and technological know-how to assist with this innovative and digitally-based system.

As state superintendent, I wanted a coherent, comprehensive assessment system that assured that all students had the opportunity to learn. In Ohio, we saw a future assessment system that was built upon clear and succinct academic content standards that incorporate 21st century skills such as problem

solving, innovation, and collaborative learning. We envisioned performance tasks embedded in mini-curricular units that would be crafted to allow teachers to individualize learning opportunities for students through individualized student plans. We began to craft a focused, professional development system for superintendents, principals, teachers and parents to better understand and participate in the development of this new assessment process. We set up a system that provided differentiated reports to superintendents, principals, teachers, students and parents to report districts', schools' and students' strengths and weaknesses and improve professional practice for educators and give reliable information to parents to support their children's learning.

As State Superintendent of Public Instruction, I saw how funding constraints limited Ohio's opportunity to develop formative and summative assessment systems that could use multiple measures such as portfolios and performance based assessments. Through a grant from the Gates and Hewlett Foundations, Ohio is now working with Stanford University and 27 school sites to develop performance assessment tasks, with strong statistical validity and reliability systems modeled after the moderations panels found in Queensland, Australia, Finland, and other higher performing countries.

To show the value of public service media on a national assessment system, I would like to address:

- Technology and Innovation in Assessment
- Project Management and National Consortia

Public service media has rich and trusted digital content such as video and audio programming, online games, simulations, podcasts and other digital learning objects that allow for multiple representations of the same concepts essential for assessing and teaching diverse learners. These resources, much of which is in the public domain, motivate and engage the audience. Some of this content has been subject to rigorous evaluations that demonstrate its efficacy for enhancing the learning outcomes of poor and underserved children. Our public media system is in the process of aligning this content with

academic standards through the PBS Digital Learning Library and CPB's American Archive program. In addition, we can customize digital learning objects for assessment projects. These resources can be used for performance tasks, performance portfolios, constructed responses and essays. Our content can blend both academic and technical studies and test subject matter competency, and can create tasks that stress habits of mind for collaboration, design, invention, and entrepreneurship – skills essential for success in the 21st century.

The public broadcasting system has a long history of educator development. Systems like PBS TeacherLine, ThinkPort (Maryland Public Television), E-Learning for Educators (a collaboration of southern and mid-western stations as well as Delaware and New Hampshire) and Teachers' Domain (WGBH, Boston) are but a few examples. Our system is and can be even more helpful in facilitating teacher and administrator training in assessment literacy. Specifically, our system can provide online training on developing items for formative and summative assessments, scoring of performance tasks, the interpretation and use of results for all types of assessments, and the creation of curricular materials that can be shared electronically within and across districts, schools, and states. Our system can also provide important information to parents and community leaders. Our stations are community based and are experienced in convening stakeholders around a host of educational issues and facilitating both professional and social networking.

Public broadcasting is now experimenting with new digital media, such as I-pods, cell phones, mobile TV, and other handheld devices that can be of service to test developers as they continuously adapt to the ever-changing technology. Our local stations have experience using adaptive technologies with individuals with special needs. A leader in this area is WGBH in Boston. We also have content in Spanish. Public service media can be a valuable partner in multiple consortia that will provide differentiated assessment models for special populations and be an active participant in the

development, design, research and evaluation of a national assessment system. We have a close working relationship with the Council of Chief State School Officers, the Council of Great City Schools, and the Partnership for 21st Century Skills. We are a knowledgeable and cost-effective public partner in a national consortia that holds the promise to improve instruction for all students and holds everyone accountable for results.