Issues Concerning the Development of Assessment Systems

Public comment by Gerald M. Eads II, Ph.D.

Race to the Top Assessment Meeting

Sponsored by the United States Department of Education

Atlanta, Georgia

November 17, 2009

Gerald Eads currently conducts research for the Professional Standards Commission, Georgia's teacher certification agency. He has served as head of testing for the Virginia Department of Education, research faculty for the State Data & Research Center at the Georgia Institute of Technology, and assistant for policy research to the director at the Georgia Office of Educational Accountability. His doctoral training was in experimental and educational psychology and measurement. The opinions offered herein are entirely those of the author and do not represent the position of any agency.

1. Validity

The RT3 assessment executive summary notes that "the framework would focus on the design and quality of assessment systems and not accountability policies" yet it is the accountability policies which drive validity – and validity is at the very core of the issue of quality. The testing system you are considering must be validated for *each* of your purposes: instructional improvement, measuring school, principal, and teacher effectiveness, and predicting college "readiness." It would behoove us, for example, to require state consortia to demonstrate that the testing system differentiates among teachers on meaningful dimensions. Just because "scores go up" does *not* mean in and of itself that anything of value changes – higher graduation rates, increased college success, lower unemployment rates, employment persistence, etc. Unless we can *demonstrate* externally referenced value, a testing system is nothing more than the proverbial boat - - that is to say, a hole in the water into which we pour money. Developing an assessment system without considering policy intent and implication is little different from building the atom bomb and refusing to address the consequences. Your Framework *begins* with policy validity issues, not the least of which is "Individual student achievement as measured against standards that build toward college and career readiness by the time of high school completion" – we cannot know that the tests are valid for that purpose unless we undertake the requisite longitudinal work to determine the relationship of the test scores to the desired outcomes. I urge you to require such study.

2. Test Design

"Standards" are nothing more than minimum competencies. It matters not whether the "standards" are set at the 10th, 30th, or 90th percentile levels – they are *still* just Newspeak for minimum competencies. The executive summary talks of both standards *and* growth – the latter requiring full-range testing – in the same breath. If "standards" become as they are now nothing more than different levels of test performance, then it will be hard to avoid setting performance levels for different goals – college, tech school, direct job placement – on the same tests, similar to what some European countries do. It's called tracking. We rarely consider the consequences of so-called "high standards" – dropout rates, narrowed curriculum, and so on, and thus we rarely try to address those problems at the policy level, but leave the classroom teacher to struggle with the aftermath of our short-sightedness. We have yet to establish that in fact high-stakes tests of any nature actually improve education. I dearly hope you require states to study the long-term impact of testing policy. If scores go up, and teachers get hired and teachers get fired, and *nothing else changes* – then the millions spent on testing will be for naught.

3. **Testing Time**

We must, as noted in the USED summary, study the impact of testing time on instruction, particularly so if we are to consider multiple components during the school year. I've looked at the testing calendars for several districts in this state – we begin serious disruption of instruction in March, but it starts earlier. It is not only the time that students spend in actual testing but the time when some students are sent to testing and teachers must adapt their instruction for the remaining class so that tested students are not punished for their absence – or, at least, everyone is shortchanged equally. Every hour a child is being tested is an hour that child is not being taught. I am not aware of any research that attempts to determine the impact of lost instructional time due to testing. I would hope that you require attention to test administration design and efficiency such that the disruption to instruction is kept to a minimum.

4. **Societal Considerations**

We focus our testing on "standards" in basic areas. The USED summary suggests we begin development in language arts and mathematics. It's been adequately demonstrate that testing in only a few high stakes areas narrows instruction. Today's accountability movement seems to focus only on the development of skills to produce "good workers" with virtually no consideration for other purposes of public education, for example, *to produce good citizens*. What is the impact on the society of reductions in the arts, language and social sciences? We pay lip service to a desire to test "critical thinking" yet we seem determined to remove from schooling those things worth thinking about. Perhaps it would be useful to undertake study of the relationship between test scores and indicators of positive and negative societal outcomes such as incidence of voting and arrest. Teachers in public schools serving upper socioeconomic strata spend little time preparing students for 10th and 20th percentile tests, but if teachers in poor schools must spend all their time getting their students to pass minimum competencies in English and math, how does that impact instructional time spent in other areas, and do those changes, if any, have impact on such things as citizenship behaviors? I do indeed understand that we must "start somewhere," and that there are reasonable arguments for beginning this effort with reading and math, but let us make sure that Mr. Obama's initiative does not do to our public schools what someone else did to Afghanistan, getting things off to a reasonable start and then walking away.

Submit As Written Input From Expert

11·17·09
4:15

Designing Student Assessments in Response to the Race to the Top Initiative

Thomas H. Fisher, EdD
Fisher Education Consulting, LLC

Marriott Hotel, Atlanta Airport
November 17, 2009

I. Introduction

I am pleased to be with you today to discuss how states can respond to the requirements of the Race to the Top initiatives authorized under the American Recovery and Reinvestment Act of 2009. This program seeks to encourage states to improve their statewide student assessment programs with the aim of better preparing students for success in postsecondary education and in the 21st century workforce. This is a significant undertaking.

There is no single way to accomplish this objective. Each proposed new solution or approach will have both intended and unintended consequences that will impact your state's educational programs. You will need to proceed with due diligence.

II. Practical Lessons Learned

I have been in the large-scale assessment business for 40 years at both the local and state levels and have learned a few lessons along the way. Based on these experiences, I want to offer a few ideas that may help you in your planning.

1. Statewide assessment programs are designed to fulfill specific policy directives and purposes. One does not develop a test and then seek purposes for it. You first determine what kind of decisions need to be made about students, schools, and systems and then design a testing system that will produce scores to facilitate those decisions. You should not make statements that assign additional purposes unless you have conducted research needed to validate such new purposes. For example, building a high school test that allows valid assertions about probable success in the first semester of college does not imply that the same test will predict who will actually graduate from college or who will become a good employee.

2. Developing a statewide assessment program where one previously did not exist is actually fairly simple compared to redirecting an existing program. The latter

requires much more effort and planning. Events must be carefully sequenced so there will be no break in services or in the data produced by the tests. This is particularly important if you presently have a high-stakes testing program and want it to be replaced with a new, more challenging test also high stakes for the students. Any student currently working toward an existing high-stakes diploma must be permitted to continue until the diploma has been awarded. The state, therefore, would have to operate both types of assessment programs for four or five years. There obviously will be cost implications.

3. If the proposed common standards assessment program will have high stakes at the student level, it must adhere to the legal principles established through prior litigation such as in Florida's Debra P. v. Turlington and the G. I. Forum case in Texas. Students have a property interest in receiving a high school diploma, and this right cannot be ignored. At the very least, this means that students must understand when they enter 9th grade what the academic expectations will be, and the state must be able to demonstrate that all students were provided the opportunity to learn the required content. Additionally, in some states, test security is an issue since a student is provided the opportunity to see the test he or she failed. This means that there must be alternate forms available for each re-test administration.

4. If the assessment program is tied to high stakes only at the school and district level, there are no personal legal rights to honor, but there certainly are practical and political considerations. Educators surely must believe that the standards are fair and equitable for all students. Other interested parties will want assurances that the standards are sufficiently challenging to accomplish the goals of RttT. This is a difficult balance. No one will be happy if, under a new testing program, all schools earn failing grades. Or, if all schools achieve the highest possible evaluations, it would not lend credibility to the assertion that more challenging expectations really are being implemented.

5. It is appropriate for us to have high expectations for students, but some students and their parents may not believe that post high school education is necessary. The local culture may very well believe that the possession of a college education means the young man or woman will move away from home, never to return. This breaks up the continuity of the nuclear, multi-generation family structure. I believe this dilemma can be solved by authorizing differentiated high school diplomas. These diplomas would include a minimum skills level of performance, a higher career or technical diploma, and a college or university proficiency diploma. Students would move toward whichever one they desire and should be allowed to change their selection at any time if they wish.

6. There are no absolute proficiency standards or cut-scores to be applied to statewide assessment tests. We have developed very good approaches to setting such criteria, and they are available to be applied to the next generation of tests. There always is a question as to whether a given test is too hard, too easy, or just right. Some states have had difficulty in increasing the rigor of their current tests. In fact, within the last few weeks, there have been news stories from three different states describing how attempts to increase academic standards have not been successful. A possible solution to

this situation is to adopt cut-scores along with a requirement that they be reviewed every two years with the intention of increasing the rigor so as to maintain momentum.

7. Do not assume that classroom teachers understand principles of psychometrics or good test design. Teachers are not required to take courses in testing and measurement during their undergraduate teacher preparation training. There is no reason to assume that all teachers know how to write good test questions for use in their own classrooms, much less write items for a statewide assessment. On the other hand, they can be taught how to do these tasks, and we have many examples of talented teachers who can produce good test items or can serve as graders of students' responses to constructed response items. If it is desirable to include more sophisticated stimulus material and assessment strategies in the next generation of assessment and classroom instruction, it will require significant pre-service and in-service training.

III. Ideas for the Assessments of the Future

I now turn my attention to suggestions for your consideration as you think about a next generation assessment program. For purposes of these remarks, I will assume the adoption of a new assessment program for grades 3-8 in reading, language arts, and mathematics in accordance with RttT. I also will assume that states will provide tests at the secondary level for the purpose of measuring progress toward college-ready graduation standards. These tests may or may not have high stakes associated with them. Remember that at this time about half of the states use high stakes high school graduation tests.

1. At the high school level, each state must make a choice between using traditional tests of reading, writing, and mathematics or switching to specific end-of-course tests. The general tests are much easier to design and implement. By contrast, using end-of-course tests, usually given via computers, creates difficulties with scheduling of test administration late in the term, field-testing new items, instantaneous grading, selection of equated passing scores, and re-test opportunities. The determination of passing scores for each new version of an EOC test is particularly problematic.

2. There is insufficient linkage in most states between high school expectations and college entrance expectations. Student course grades in high school may not have much validity and can be artificially inflated. (I note that my local high school had ~~over a dozen~~ eleven valedictorians in 2009. A school in California boasted of 59 valedictorians. As USA Today commented, "Being first in your class ain't what it used to be.") Earning a passing grade may not mean the student actually has sufficiently mastered the content. Moreover, across the high schools within a single district, the teachers of, for example, Algebra II, probably have different grading standards and content expectations. This problem is compounded by variations in course content and grading standards across <u>all</u> high schools in the state. I recommend that each district superintendent be required to certify that all of the mid-term and final examinations in specific ELA and mathematics courses be coordinated so that one high school does not have easier standards than another. Implementation of state EOC tests will guarantee consistent educational quality

only if students actually are required to pass with no alternative pathways to earning the course credit. I also recommend that if your state really wants to prepare students for success in the postsecondary world, the development of such tests and setting of the passing scores should involve college and university teachers, analysis of the current content expectations for college freshman English and mathematics courses, and long-term tracking of student success.

3. The RttT literature envisions high-quality assessments and an overall system that includes formative assessments, interim assessments, and coordinated instructional improvement systems, all structured to permit fast turnaround of test results. Student growth over time will be monitored and linked to the performance of individual teachers. I summarized this in two sentences, but the implications are enormous.

Let us first imagine that there will be a summative instrument administered via computers, scored promptly, with results available in a matter of hours. A system of this nature will require a large and varied item bank, a sophisticated computer system for administration, rules for retesting students who fail, and up-to-date equipment in all schools. It is doubtful that such a system could include any items that are scored by human readers.

Now, consider the matter of interim tests. In one scenario, there may be one administered at mid-year, subject to the constraints just mentioned. Or, there might be two interim tests, one each at the middle of each semester. Again, if immediate turnaround of results is demanded, constructed response items probably could not be included. If we assume that the purpose of the test is to determine if the student is "on track" for the end-of-the-year test, it will be necessary to validate the degree to which a given score can predict a level of performance on the final assessment at the end of the year.

The Department of Education envisions the development and implementation of a series of formative assessment activities to be administered in conjunction with the instructional program throughout the school year. These activities could be computer administered, but they also could be paper/pencil tasks or small group projects graded by classroom teachers. The immediate question to be asked is how will teachers be involved in the creation, administration, and scoring of these assessments? Will this require additional work time for which they must be compensated? What will their training needs be? And, what will be the relationship of these activities to whatever content is included in the ordinary offerings of textbook publishers?

4. The continued administration of summative, interim, and formative assessments will require an enormous number of test items of all types, especially if some items will be annually released to the public. It is not clear that these items can all be gathered from existing sources, so new and continued item development will be required. To support such efforts, I would create, within each participating state, assessment development and training centers, associated either with universities or large school districts. These centers can be used to review, edit, and validate new test items, write

new instructional materials, and train teachers, principals, and instructional leaders as needed.

5. There is no question that modern computer and web technology enables the administration of instructional programs as well as assessment exercises. But, it is not clear that all instructional objectives can be measured via computers (e.g., working in groups). Further, there always is a question about the adequacy of bandwidth and the availability of a sufficient number of computers within individual schools so that all students can be tested within authorized windows. If computers are not available or if some students cannot take the tests in this way, alternate testing procedures must be created. Technology continues to improve, even as we meet today. I noticed last week that Microsoft will be offering software that will permit several students to link their keyboards and monitors to a single classroom computer in a type of mini-network. This should expand the number of work stations without having to purchase more computers.

6. The assessment system being envisioned by RttT requires thorough and prompt reporting of results. Depending on the way in which the computer tests are scored, the results could be quickly available. Results aggregated at the school, district, and school levels will require additional time. The district and state results should be readily available to the public for viewing or the downloading of research data files. Currently, not all states have web sites that offer such services.

7. I have said nothing about the continued use of nationally normed achievement tests. Many parents will want to know how well their child performed compared to national norms. If RttT leads to a substantial number of states using the same summative testing apparatus, de facto norms will be available and can be used to describe a student's performance relative to other students, although this distribution may not be "national." Otherwise, to provide national normed data, it will be necessary to either separately administer such a test or to link the RttT tests to the scales of an available norm-referenced test. Again, there are cost implications.

IV. A Few Managerial Matters

1. States should track all high school students after they graduate and move toward employment, college enrollment, and graduation with a postsecondary degree. If you want to see an example of how this can be done, I recommend that you look at the Florida FETPIP operation. You can find it on the web.

2. Initial development of the new challenging assessment program will be supported by RttT funding. It is not clear what the long-term fiscal impact will be on states that implement the new academic standards and assessments. I recommend that you include budget analyses and expenditure forecasts so you will be adequately prepared. New assessments that feature constructed response items will be costly and, indeed, some states presently are minimizing or eliminating such questions because of the costs and time required for grading.

3. The RttT program is encouraging states to develop consortia to develop the needed assessment instruments and procedures. This can be a means of spreading the costs and gaining economies of scale. It will be necessary for you to determine whether or not there are any laws preventing participation in such consortiums, negotiating ownership of the resulting products, building a management team, deciding how to parcel out the work activities, and agreeing on many details of the work. Maintenance of a tight work schedule will be absolutely required. Consideration must be given as to whether the consortium will be used just for initial development and administration or if it will be an on-going commitment for test renewal, administration, scoring, and reporting. Will your state permit such an obligation of funds into future years?

4. RttT allows states to adopt common academic standards, but each state may have up to 15% additional standards, by content area, that may be unique. If each state decides to measure additional content, there could be 50 different analyses of data, item sets, equating processes, and so forth. Alternatively, each state could assess its additional content by administering separate tests with results not reported on the same scale as the common national test. National comparison data obviously would not be available for these items.

5. Finally, we must understand that each state already has existing contracts for assessment and school accountability services. Typically, these contracts have initial terms that may range from two to five years with several optional renewals. The contracts were the result of competitive bids in order to get the best possible pricing. To change direction for the existing statewide assessments means the states will have to negotiate contract amendments without the benefit of competition or must cancel the existing contract, issue a new RFP, and select a contractor to operate the newly developed assessment program.

Conclusions

As I began my remarks today, I said that the Race to the Top initiative presents a complex set of challenges. As you move forward, I urge that you review every aspect of the state's approach to curriculum requirements, instructional programs, promotion and graduation requirements, the design of the present testing programs, and the school accountability programs so you will understand the adjustments and changes that must be made to move toward meeting the Race to the Top objectives.

Thank you for your attention. I look forward to our discussions today.

[redacted]

From: Clifford Lazar [LazarCW@ca.rr.com]
Sent: Tuesday, November 17, 2009 1:29 PM
To: Race To The Top Assessment Input
Subject: I would attend Denver if I were a speaker

Here is a summary of my proposal:

Grading Teachers and Pricipals

The solution for measuring and rewarding teachers and principals fairly, is to take the readily available demographic factors into consideration.

Collect the census data for all the census tracts in the school service area. That includes median family income, percapita income, female head of household, race, education, persons per household, crime rate, etc. Last time I looked the data was available on the Internet. The first time I used demographic census data was to forecast the vote in the California Open Housing Initiative.

Each census tract has a centroid, i.e., latitude and longitude. Each school also has a latitude and longitude. For each school select the surrounding census tracts with in the avarage radius for the schools in the district, or if the boundaries are known, that actual census tracts.

Correlate (regression) the census and crime data with reading and math scores, using multiple regression.

For each school in the district, predict the reading and math scores for each school, based on their sociological factors.

If the school performance results exceed the forecast, reward the teachers and principal. If the school is below the forecast, look at the reasons for the poorer performance. If the staff is incompent, weed out the poor performers. It may be that reducing class size will have the biggest impact.

Benefits of Fairness

The benefit of this methodolgy is that it is fair to the teachers and principals. All the counter arguments posed by the teachers' unions are answered and still we have a meaningful evaluation with good and bad performers. A second benefit is that it is immediate. Value-added requires multiple years of results, meanwhile the demographics and the teaching staff are changing.

This technology was used at Atlantic Richfield Oil Company, where I worked, to determine the better gas stations. Poorer ones were shut down.

I would be happy to demonstrate how this works successfully.

I taught this at USC in my course, "Economic Development in the Ghetto, Barrio and Indian Reservation."

Clifford W. Lazar
Lazar Developments

9717 Cisco Street
Los Angeles, CA 90034-1053

tel: 310-838-3885
fax: 310-838-8066

alt. email: LazarCW@ca.rr.com

Speaker # 2
Rcv'd 11.17.09

My name is Steve Lazer and I am Vice President for Assessment Development at Educational Testing Service. I'd like to express my thanks for this opportunity to speak to you.

This is a moment of unparalleled opportunity to improve assessment. Advances in technology, coupled with innovative assessment task design and advanced psychometric and cognitive models, make it possible for us to obtain a richer picture of what students know and can do than ever before. And the United States seems willing to invest in such improvements.

However, with opportunity comes risk. We will not get a chance like this again. Different stakeholders will set diverse priorities for an assessment system. Any assessment is a compromise and cannot do everything. As a result, we will need to make certain key choices.

Given limited time, I'm going to focus on three recommendations. I have other ideas on a broader system, and would be happy to share those on another occasion. My recommendations today pertain to summative and not formative assessments.

1. **Consider the use of intermittent assessments**

It is likely that that the system will still require end-of-year "testing events" at grades 3 through 8. However, these may not be the only elements of the summative system. It is possible to use data collected over the course of the year through one of several models. One example would be to take a long test, including performance tasks, and break it into pieces and give those over the course of the year. A variant is a more general "final" with summative tests (perhaps that count less than the final) given along the way. A different approach is to have students complete standardized projects over the course of the year. If development and scoring were carefully controlled, it might be possible to use these data as part of a summative system. Of course, any of these models assumes some degree of connection with the scope and sequence of instruction. This may be politically or operationally difficult to mandate.

2. **Use the range of exercise types needed to measure the emerging constructs**

It is quite clear that new standards and the new constructs those standards might imply will demand item types well beyond simple 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.

One interesting question related to this is how to score these tasks. Certainly the need to build an ongoing operational system would suggest strongly that we try to make use of technology to score forward-looking and complex tasks. We should do so, although we need to be careful not to limit measurement to fit the limitations of the scoring systems. In other words, we need to realize that it is likely that elements of the new assessments will require human scoring.

If this is true, then we need to use technologies to efficiently distribute and exert quality control on human scoring. Assuming teachers are to be asked to score, we also need to take what we have learned in various programs to ensure that scoring is a valuable professional scoring activity, and not just drudgery.

One closing note about innovative item types: we must not repeat some of the mistakes we made in the 1980s and 1990s. We cannot create a system so expensive as to be operationally untenable. We cannot act as if the technical challenges associated with performance tasks in particular are mere inconveniences: validity, fairness, and reliability really do matter if we want to make decisions about students, teachers, schools, and school systems that are

meaningful, defensible, and equitable. This is not to say that technical concerns should "win out" over all others, but rather that we should carefully consider the implications of various decisions, so we are not caught by surprise and so that we can build research activities around addressing key issues. Finally, we must make sure that complex items live up to their promise to provide more nuanced and richer information, and that we do not engage students in long intricate activities that yield only a single summary score.

3. **Build a technology-delivered system**

Some will argue that since states lack the infrastructure needed to move all testing on to computer, assessments will need to remain paper based. Others will argue that at a minimum we will need parallel paper and computer versions of tests. However, in spite of this I believe that standardized elements of the new assessment of common-standards should be delivered on computer or similar technology.

Emerging standards can only be measured through the use of technology. 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 bring to bear skills, that wouldn't (and couldn't) be invoked on a paper test.

Maintaining parallel paper and computer systems on which results were supposed to be interchangeable would effectively prevent measurement of such skills. Technology also allows for flexible (adaptive) and on-demand testing. Rapid return of scores and seamless data/information interchange is facilitated by technological delivery. Technology allows for provision of a range of accommodations for students with disabilities and English-language learners that might not otherwise exist.

This sort of testing will need to change the way we think a bit. It is likely that we will not test only at fixed times, but rather in broad windows that allow us to make effective use of available computers.

I think adaptive administration is important, because as standards get more rigorous, we need an assessment system that measures those standards but still gives us some meaningful information about what students at lower (and higher) levels of performance know and can do. It also may help us enhance the instructionally actionable information that issues from these systems.

There are several issues that must be resolved in moving to adaptive testing. The first is that there is a tension between adaptivity and use of human-scored questions. This problem can be solved: Adaptation does not need to be

done item-by-item, and staged approaches should allow human scored components to be targeted after initial routing decisions have been made and therefore used more appropriately. Also, an adaptive system might be more effective if it could, at times, make use of "off-grade" content for individuals. This is not currently consistent with NCLB, so Congress would need to facilitate a change in this area. Finally, we need to make sure adaptive systems make sense for all students, including those with special needs.

Thank you again for the opportunity to talk. We have real opportunities. Success will involve creativity, selection of priorities, care, and in some cases patience. We at ETS commend and support ED's efforts to create a next generation assessment system to help measure, and to more effectively model, learning and instruction.

Speaker # 1

Received 11.17.09

Atlanta – General Assessment Meeting
Tuesday, November 17, 2009
10:00 AM – 5:00 PM
Public Speaker Testimony
Roy Truby

Thank you for the opportunity to provide comments on the notice of proposed requirements and approval criteria for the $350 M Race to the Top (RTTT) Assessment Program. I am Roy Truby, Sr. Vice President of State and Federal Programs for Renaissance Learning, Inc.

First, I would like to disclose that RLI has an interest in the general assessments component of the RTTT Assessment Program which includes a newly developed interim assessment solution called The STAR Constellation.

Renaissance Learning's interim assessment programs—STAR Reading, STAR Math, and STAR Early Literacy—are used by over 44,000 schools. The STAR Constellation, adapted from the STAR assessment platform, is poised and ready to incorporate the common core state standards. Our goal is to facilitate a seamless transition to these new standards by providing states with an opportunity to quickly launch an efficient and reliable, multipurpose interim assessment solution that will inform instruction.

My current position is not the reason I am providing testimony. My testimony is based upon my experience in testing policy at the local, state, and national level. For 13 years, I served as the executive director of the National Assessment Governing Board (NAGB) which has policy oversight for the National Assessment of Educational Progress (NAEP). Prior to my position at NAGB, I was a state school superintendent in two states (Idaho and West Virginia). I also served as district superintendent in Greenville County, South Carolina.

To stay within the five minute timeframe, I would like to skim the surface on three major topics: 1) balanced assessment systems, 2) development considerations, and 3) the value of multiple-choice testing.

Balanced Assessment Systems

First, states need a balanced assessment system. Why is one needed? There seems to be tremendous consensus among assessment experts, researchers, and educational policy makers that an end-of-year summative test does not provide timely information to guide instructional decisions and improve student learning[1]. In fact, it is not intended to. Summative tests are primarily for accountability.

A balanced assessment system must go beyond the annual summative test to also include interim and formative components that will help educators accurately assess the learning progress of a large and diverse student population and use assessment information to adjust instruction and improve achievement.

For states to provide a balanced assessment system they will need to provide leadership—not regulation—on implementing interim and formative assessments that will inform instruction and indicate whether students are on track to meet end-of-year learning goals.

So let's follow a logic chain. What do we know? Teachers feel like they are drowning in a sea of mandated tests. Tests that take too long to administer and provide little to no information they can use to make instructional decisions that would benefit their students. If we encourage assessments that are longer to administer and have delayed scoring based on the amount of constructed-response items we're just intensifying the problem. Does that make sense to anyone?

The only realistic response is to use computer-adaptive technology that includes:

- √ Shorter development time

[1]Gong, B. (2009, March 9). *Using balanced assessment systems to improve student learning and school capacity: An introduction.* Paper commissioned by CCSSO and the R&D Consortium sponsored by Renaissance Learning.
Perie, M., Marion, S., Gong, B., & Wurtzel, J. (2007, November). *The role of interim assessments in a comprehensive assessment system: A policy brief.* Retrieved November 10, 2009 from http://www.nciea.org/publications/PolicyBriefFINAL.pdf.

- √ Primarily multiple-choice items
- √ Efficiency administered, lasting only 15-20 minutes
- √ The capacity to generate results instantly

A balanced system is also essential for implementing the kind of tired intervention system supported in IDEA and Title I and being implemented more frequently in schools across the country. Such systems, which are commonly known as Response to Intervention or RTI require data-based instructional decision making and the use of screening and progress monitoring assessments. We know that RTI works for student of all ethnicities and social-economic status, but assessment tools for screening and progress monitoring must be implemented in practical ways for teachers.

Development Considerations

Second, new assessments based on the common core state standards need to be implemented in the very near future, and interim and formative tests should lead the way.

Developing innovative assessments from scratch is unlikely to result in operational assessments ready in the next two or three years. Based on my own experience, I'd say 5 to 10 years is more likely (the more innovative the assessments, the longer the development time).

As Executive Director of NAEP's governing board I had the experience and misfortune of having "exclusive control" over President Clinton's Voluntary National Test. I had never before or since been given exclusive control over anything so perhaps there was a lesson to be learned when congress passed a "cease and desist" order on that program.

But I digress, the point I want to make is despite the large allocation and support by President Clinton the order came almost two years and $10 million into the *beginning* stage of development. We were just getting off the ground after too years. If we expect to have tests aligned to the common core state standards and if today's state accountability tests are to be augmented by balanced systems that include interim and benchmark assessments, there will be a premium on efficiency to preserve the maximum amount of teacher time for instruction. This can only be accomplished through computer-adaptive tests.

We need to get this done within a reasonable time; within the next three years we need to roll out the assessments to inform instruction and improve instruction.

To accomplish this goal, we should keep the way open for acquiring and perhaps expansion of existing assessment methods for assessing the common core state standards, in the interest of minimizing development time, reducing technical risk, and minimizing the duration of reliable interim benchmark assessments of the common core state standards.

Value of Multiple-Choice

Last, multiple-choice items are less expensive, easier to develop, administer, and score and should be the primary means in which the common core state standards are assessed. We recognize that not all standards can be assessed using multiple-choice items; however psychometric research has shown that you can efficiently and accurately predict proficiency on those items using well constructed multiple-choice items[2].

The current nuances surrounding the design of an assessment system aligned to the common core state standards seem to favor minimizing the use of multiple-choice items and instead favor constructed and extended-response formats as well as performance measurements.

My experience at NAEP showed me that constructed-response items consumed 45% of the cost and administration time, yet comprised just 15% of the items.

---

[2]Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Erlbaum.

Atlanta – General Assessment Meeting
Tuesday, November 17, 2009
10:00 AM – 5:00 PM
Public Speaker Testimony
Roy Truby

Minimizing the use of multiple-choice items would be inadvisable for several reasons:

- History repeats itself. Past attempts have failed to supplant the use of tried and true multiple-choice items in major assessments.
- When I served as Executive Director of NAGBE we had great pressure in the early 1990's to replace multiple-choice items with open ended items. We found that performance-based assessment, and even constructed-response and extended-response assessments are difficult and expensive to develop, score, and operate.
- The NAEP Board wanted to be state-of-the-art, but not state of the fad. Unfortunately, education is known for fads and assessments that are overly dependent on constructed-response item is a fad that is making a come-back.
- Constructed and extended-response items are likely to be much more time-consuming in administration and scoring.

In summary, with over 30 years of experience in testing policy at the local, state, and national level, I have never witnessed such an historic opportunity to make substantial changes to state assessment systems. I believe these changes start with integrating a balanced assessment system, led by interim and formative assessment systems with multiple-choice items as the primary testing item.

Thank you again for the opportunity to comment on the $350M RTTT Assessment Program proposed notice.

Sincerely,
Roy Truby
Sr. V.P., State and Federal Programs
Renaissance Learning, Inc.

Speaker # 3

Rcv'd 11.17.09

US. Dept. Education

Input Panel Meeting
Speaker Question Response

My Proposal of Computer Base Testing Assessment System

Joe N. Weems Jr.
Nov. 17, 2009

Joe N. Weems Jr.
Parkside Elementary School
Atlanta Public School

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.

The Assessment system currently used and widely accepted in the educational system includes assessments such as SATs, ACTs, and GREs to declare an individuals competence and readiness for our United Stated educational system. I propose a Computer Based Testing (CBT's) assessment system, which has a formative assessment and teaching component with a format including practice testing, example testing and actual testing.

(a) Describe how this assessment system would address the tensions or tradeoffs in meeting all of the general requirements and required characteristics.

The propose CBT assessment system would be standard based and the questions would be designed to meet each individual standard to ensure the measurability of the students meeting or not meeting the readiness standards.

(b) 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.

Strengths- CBT systems are widely used and accepted by University and other business industries, such as Computer Based Test ACTs, SATs, GREs, GACE and most recently GKIDS.

Limitations- Like any assessment CBTs can be abused and invalidated depending on a variety of variables such as the noise level or an individual's disability.

Question

(2) For each assessment proposed in response to question (1),

describe the--

(a) Optimal design, including- [cir] Type (e.g., norm-referenced, criterion-referenced, adaptive, other);

I would like to see both the norm-referenced and criterion-referenced combined in my propose Computer Based Testing assessment system. If research suggests this design as not optimal then administering two separate tests is an alternative to ensure the reliability and validity of the assessments. I also propose CBTs begins in Pre-K and continues through high school graduation. This is also allows monitoring of the individuals testing results over the duration of their schooling.

(b) Frequency, length, and timing of assessment administrations including a consideration of the value of student, teacher, and administrative time);

Testing frequency of the proposed CBTs will be three times per year: Base-line, Mid Year and End of the year. Ultimately this will require three visits to the schools computer lab by students.

(c) Format, item-type specifications (including the pros and cons of using different types of items for different purposes), and mode of administration;

There will be twenty five questions covering each area: Social Studies, Science, Language Arts and Math, totaling 100 questions. This allows the teacher to identify the student's strengths or weaknesses. The teacher can then identify appropriate practice test to ensure standards are being met.

(d) 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;

There will not be a difference from what is commonly used in the U.S. CBT educational assessment system today. The writing portions on CBTs are scored using the rubric of 1-4 to show competency in writing.

(e) 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.

Students will have sample assessments to take at school as well as at home. Test Banks can be used to assist teachers and parents in teaching the students how to prepare for the CBT.

Question

3. 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?

An LEA may use the funds to provide free staff development for educators and to create publications that support CBTs such as study guides. Funds can also be used to organize regular meetings for top educational assessment designers, and to fund additional research.

(a) 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?

Activities that would advance the transition to and implementation of the consortium's common, college and career ready standards and assessments is to supply teachers with staff development in creating electronic portfolios for students, teachers, and

administrators. Support may also be given for staff development on National Education Technology Standards (NETS).

Question

(4) How can such assessments be administered and scored in the most time efficient and cost-effective ways.

With the proposed use of CBTs students will receive the results immediately after completion via the computer. CBTS are cost effective in that they reduce the cost of paper-pencil test.

Question

(5) 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?

With the proposed CBTs both competencies based and grade-level based can be evaluated to ensure readiness for the next grade level, competency and career ready.

(a) How would your design ensure high expectations for all students?

The design will be researched based allowing each student to have a reliable assessment that is standard base. The students will have the opportunity to watch videos and listen to audio modification, and multi-language versions of the test.

11.18.09 Speaker #7
Received @8:13 Am
(7)

Race to the Top: Assessments for Students with Disabilities
November 18, 2009
Atlanta, GA

I'm Sue Bechard, Director of Inclusive Educational Assessment at Measured Progress, and I appreciate having the opportunity to speak here today. Measured Progress is a non-profit company based in New Hampshire. 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). Consistent with our not-for-profit educational mission, our bottom line is teaching and learning.

From the start, we've worked with states on standards-based assessments that are customized, inclusive, innovative, non-traditional, and geared toward a variety of student populations – general, special education, and English language learners. Our 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 worked at Measured Progress for ten years to develop and implement statewide alternate assessments for our clients and direct numerous federal grants to investigate aspects of assessments for students with disabilities. My prior experience included thirty years of teaching students with disabilities, state department of education work, research, and program evaluation.

I would like to make three recommendations today regarding features that Race to the Top proposals should include for assessments for students with disabilities

My first recommendation 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.
- It is necessary to require that principles of universal design are applied to all components of the system, from the standards and descriptions of expectations of students, to assessment development and test construction, and to instruction. Development plans should describe how the assessments will provide multiple methods of presentation and expression and flexible options for engagement.
- 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.

- Also, a strong evaluation plan should be articulated. We know that assessment impacts curriculum & instruction, and studies to investigate both intended and unintended consequences of the assessments need to be planned.

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 with rigorous professional industry standards and provide real opportunities for all students to demonstrate what they know and can do.

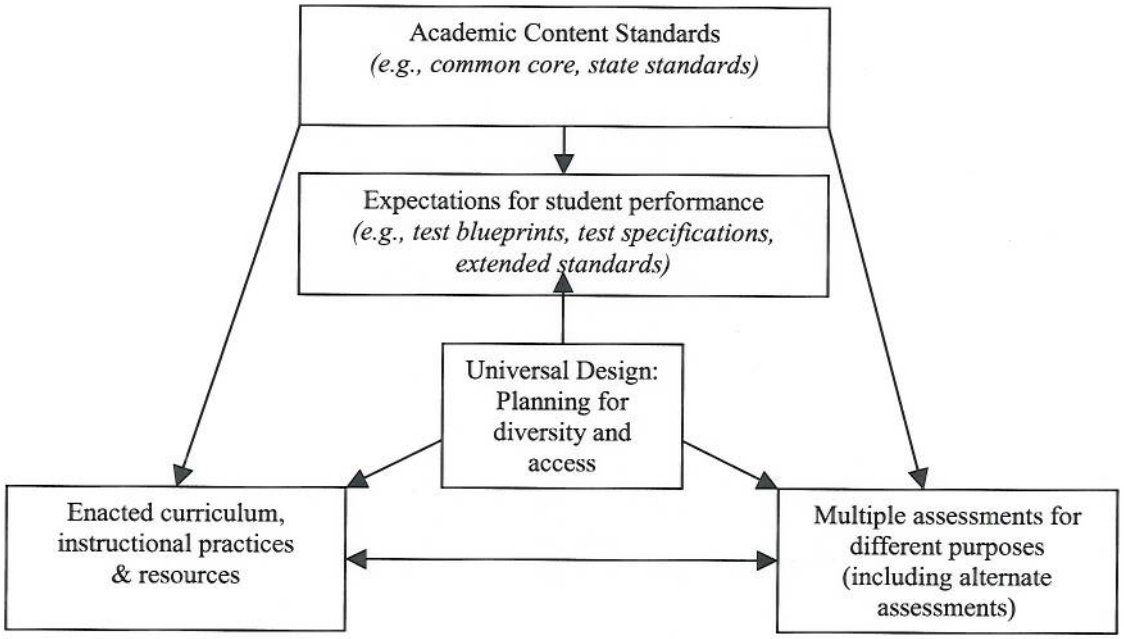
- Race to the Top should provide opportunities to explore more meaningful ways to measure achievement of students with and without disabilities and should allow for the development of assessments that are not burdened by high stakes but are devoted to the finding better ways to allow students to show what they know so teachers can determine better ways to teach. 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 our current research in this area for non-disabled students and there is practically no information regarding students with disabilities.
- Next, proposals should describe precisely what academic content will be assessed. 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 designs and uses of technology**, the promise of technology is not only about 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.
- Using technology for assessment will permit adaptive and scaffolding strategies to be employed, so we can find out more about what each student actually knows and investigate the misconceptions they have when they are not being successful. This will allow us to more effectively understand the learning pathways for exceptional 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.

Thank you.

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.)

Race to the Top Presentation
Atlanta, GA – 11/18/09

Renée Cameto, Ph.D.
SRI International
333 Ravenswood Ave., BS 132
Menlo Park, CA 94025
650-859-6451

Question addressed: 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?

Students with disabilities challenge conventions with respect to the teaching, learning, and assessing of academic content. Assessment designers are challenged to develop systems that adequately and reliably show what these students know and can do. Indeed, it is the sheer variability in this target population, the assumptions about measuring their achievement, and the variability of design implementation procedures that make traditional approaches inapplicable without some reformulation. The application of Evidence Centered Design (ECD) to assessments of students with disabilities directly addresses the most pressing issues by using a replicable design process that can be applied to all content areas and all types of evidence from performance tasks and portfolio activities, to technology-based simulations and animations, to traditional multiple choice item formats. The use of ECD can enhance the technical characteristics of assessments, improve the efficiency with which future assessments can be developed, and documents the myriad of design decisions required when developing a valid assessment of student learning.

We at SRI International have been working with a number of federal, state, and organizational partners on a variety of ECD research and development activities. Currently we are engaged with several consortia of states on projects funded by NSF, IES, and EAG grants. Through nearly a decade of work, we are finding that the ECD framework is an innovative and effective approach to the assessment design and development process for all standards-based assessments, including those for students with disabilities.

SRI has a web-based assessment design system that embodies the principles of ECD and provides a platform for the design of assessments. This web-based system is fully operational and has been successfully applied to assessments in reading, math, science, and the social sciences, to assessments for students in elementary through post secondary school, and to assessments for students in general education and in special education, including students with significant cognitive disabilities.

ECD is a five-step process. The first step is Domain Analysis. In this step, the specific content to be included in the assessment is determined. State content standards and the pending common core standards are examples of domain analyses. In the second step, Domain Modeling, a high level description of the overall components of the assessment is created and documented. The third step is the creation of the Conceptual Assessment Framework. In this step, the knowledge, skills, and abilities (KSAs) to be assessed, the evidence that needs to be collected, and the features of the tasks that will elicit the evidence are specified. We also specify any non-focal KSAs that may be required to respond correctly to an assessment task but are not the target of the assessment (for example, reading skills in a mathematics examination). By identifying these non-focal KSAs, we can reduce the construct irrelevant variance they introduce. The fourth step in the ECD process is called Implementation. In this step, the assessment items or tasks are authored using the specifications created in the conceptual assessment framework just described. In addition, scoring rubrics are created and the scoring process is specified. In the last step, Delivery, the processes for the assessment administration and reporting are created. Here we answer questions such as will the assessment be administered online or with paper and pencil, will it be delivered one-on-one or by classrooms, and for what audience will reports be created.

ECD works synergistically with other innovative approaches to assessment, such as Universal Design for Learning (UDL). UDL principles are incorporated in our ECD process during assessment design and item authoring by considering multiple means of perception, expression, cognition, language and symbol use, executive functioning, and engagement. This can include consideration of augmentative and alternative communication systems.

Creating appropriate and valid assessments for students with disabilities and those with significant cognitive disabilities challenges traditional assessment approaches and assumptions. Solutions to the challenges to serve students with disabilities have frequently resulted in techniques and approaches that have applicability to the broader population. There are a range of new approaches that move us closer to having the assessment systems that students with disabilities deserve and that can improve assessment technologies for all students. ECD is one of these innovative approaches.

Received @ 751 AM
11.18.09

Guidelines for Accessible Item Development

Lizanne DeStefano, University of Illinois

These guidelines identify the major themes and dimensions of construction that should be addressed when developing blocks of NAEP items to create an accessible block. The guidelines are meant to aid item developers in assessing relevant and irrelevant aspects of the item's construct that contribute to the overall difficulty and accessibility of the item. The guidelines offer common strategies for reducing difficulty without compromising content, construct validity or alignment with the NAEP framework.

These guidelines should be applied judiciously. Their application may vary from item to item depending upon the measurement intent of the item. Generally, these guidelines should be followed unless the targeted construct of the item precludes doing so.

Word Choice

Careful word choice is an essential component of quality item construction. Word choice refers to language used within the statement of a problem, as well language used in the alternative answer choices. Careful word choice should be a central consideration during the item modification process.

- **Clarity** - Word choice throughout all items should be unambiguous and concise. It is more important for item wording to be clear than for it to be precise. For example, avoid the ambiguous phrase "about how much" when writing problems that require estimation or rounding.
- **Plain Language** - Plain language, as a writing and editing tool, is designed to clearly convey meaning without altering what an item is intended to measure. All items should use plain language. Even when the intent of the item is for the student to define, recognize, or use mathematics vocabulary correctly, the surrounding text should be in plain language. Plain language should increase access and minimize confusion.
- **Terminology Appropriateness** - Terminology used should be current and relevant to a broad population. Use of outdated technology, terminology, etc. can distract from the content of a problem.
- **ESL Considerations** - Use of commonly accepted and culturally non-specific words, phrases, and terminology is encouraged whenever possible. Be careful of literal interpretations of items. When using words with multiple meanings, make sure the intended meaning is clear. Avoid ambiguous words such as if, could, may, can, etc. Use high-frequency words as much as possible. Avoid the word "not" whenever possible.
- **Parallel Item Construction** - Item wording should provide parallel syntactic construction. Use of the present tense verb is preferred. Wording within and between the statement of a problem and its possible answer choices (including distracters) should be consistent in tense and vocabulary.

question in an item (how, what…) from the rest of the information presented in the item.

- **Answer Spacing** - Provide appropriate space for an answer. Too little or too much space for an answer can falsely suggest an answer of a certain length.
- **Clarity** - Use format to clarify text. Use bullets, space between pieces of text, and boxing of text to emphasize or separate information.
- **Item Separation** - Provide a clear distinction between each item. Some NAEP items provide information (e.g., a graph, chart) before the statement of the problem. In such cases, the item should always begin on a new page in order to provide a clear distinction between problems.

Graphics

Graphics such as pictures, charts, and diagrams are visual images reflecting information.. Graphics can be very effective in supporting text, illustrating mathematical concepts and increasing item access. If used improperly, however, graphics can add substantial confusion and distract test takers from the intent of the item. Graphics should be used judiciously.

- **Clarity** - Visuals should be clear and precise. Adding a visual may clarify the measurement intent of an item.
- **Mathematical Accuracy** - Visuals should utilize standard mathematical notation and formatting.
- **Simplicity** - Visuals should only contain necessary information. Remove unnecessary graphics. Avoid misleading graphics, such as charts with inconsistent scales.
- **Completeness** - Visuals should provide a representation of the important parts of the item. Visuals should mirror and parallel the wording and expectations of the problem.
- **If a visual** within a given item is adding to the unintended difficulty of the item, it should be altered or removed.

Appropriate Use of Context

Contextual information includes problem scenarios, explanations, specific directions, and background text. Using contextual information can place mathematical concepts in more realistic conditions and provide background information that test takers may need. However, the contextual information should not interfere with the mathematics being assessed. It should not be a barrier to a student's ability to demonstrate his or her mathematical knowledge. Contextual information should be included only if the item is intended to assess mathematics in context.

- **Plain Language** - Use plain language as much as possible.

- **Clarify Answer Requirements** - Cue students about the number and type of solution(s) they should provide (e.g., written description, graphical representation, etc.). This is especially important in open response items that could be solved using multiple approaches.
- **Avoid Deceptive Cues** - Do not mislead students to perform inappropriate operations.
- **Provide Definitions when Appropriate** - It may be appropriate to provide a brief definition, example, or illustration of a mathematical concept if doing so does not compromise the objective of the assessment item.
- **Use Cues to Clarify Item Intent** - Remember, the objective and intent of all testing items should be as clear as possible.

Computational Appropriateness

Each item on the NAEP is assigned a mathematical complexity rating (low, moderate, high). The task asked of the student should reflect an appropriate computational level. Generally, it is possible to reduce the computational complexity of an item while preserving its alignment with the NAEP framework.

- **Assess Computational Complexity** - Do not require students to perform calculations that are unnecessarily difficult. Calculations should not distract from the general idea being assessed in any given item.
- **Gauge Time Constraints** - Do not require students to perform calculations that are unnecessarily time consuming. Calculations should not distract from the "flow" of the testing experience. Remember that TIME is a precious resource during the testing experience.
- **Computational Progression** - Do not require students to perform counterintuitive operations.
- **Encourage Mathematical Accuracy** - Do not ask students to estimate or round when exact calculation is appropriate or easier.
- **Calculator Use** - Items should be constructed with calculator use/availability in mind. Computational complexity should be appropriate to the testing context. Remember, the availability of a calculator should not increase the complexity of a problem.

Grade Level Appropriateness

Item modifiers should identify the objective(s) being assessed by each NAEP item as well as the grade level at which it is meant to be assessed. Items in an accessible booklet should be constructed to assess at or below the grade level under consideration. For example, a fourth grade accessible block should not contain items that are constructed to assess a learning objective at an eighth grade level. In most cases, the NAEP framework provides leveled descriptions of each learning objective. Students being assessed using an accessible block should not be asked to perform a task at a level higher than is appropriate for their grade.

Aligning the Accessible Block Assessment with the NAEP Framework

Lizanne DeStefano, University of Illinois

Similar to standard blocks of NAEP assessment items, all accessible blocks should be developed so that they are aligned with the content expectations defined by the 2005 NAEP Mathematics Framework. Unlike standard blocks of NAEP assessment items, there will be less variability in the level of complexity of items in a NAEP accessible block. Drawing upon Webb and others*, five interrelated dimensions are considered in structuring the NAEP assessment so that it is aligned with the NAEP framework:

1. The match between the content of the assessment and the content of the framework: The assessment as a whole should reflect the breadth of knowledge and skills covered by the topics and objectives in the framework.

2. The match between the complexity of mathematical knowledge and skills on the assessment and in the framework: The assessment as a whole should represent the balance of levels of mathematical complexity at each grade level as described in the framework. However, an accessible block is meant to provide important statistical information about students at the lower end of the performance continuum. Therefore, it is appropriate for an accessible block to contain items that assess students' ability to perform tasks associated with Basic and Proficient levels of achievement.

3. The match between the emphasis of the assessment and the emphasis of topics, objectives, and contextual requirements in the framework: The assessment should represent the balance of content and item formats specified in the framework and give appropriate emphasis to the conditions in which students are expected to demonstrate their mathematics achievement, reflecting the use of calculators, manipulatives, and real-world settings.

4. The match between the assessment and how scores are reported and interpreted: The assessment should be developed so that scores will reflect both the framework and the performance described in the NAEP achievement levels.

5. The match between the assessment design and the characteristics of the targeted assessment population: The assessment should give all students tested a reasonable opportunity to demonstrate their knowledge and skills in the topics and objectives covered by the framework (with a special emphasis here being placed on providing students at the lower end of the performance continuum an opportunity to show what they know and are able to do).

- Are all instructions clear?

Recommendations for Common State Assessments
Including Students with Disabilities

November 18, 2009

Good morning, my name is Jeannette Downes and I come before you as a parent, educator and union leader. As a parent I support a child who had IEPs in elementary school, testing accommodations in high school and now, for the first time, learning accommodations in college. As an educator I support 150 teachers and paraprofessionals as a UFT Teacher Center representative providing professional development services. My school serves 500 children with severe autism, cognitive delays, hearing impairments, and emotional disturbances. My role as a parent and educator is to ensure our children have what is needed to succeed in school and life.

I am an active member of Local 2, United Federation of Teachers, New York City. We support 118, 000 active educators and 1.1 million students at 1,600 schools. On behalf of the 1.4 million members of the American Federation of Teachers, I come before you today with some thoughts on how the US Department of Education, Secretary Duncan, and you, Joanne Weiss can better align assessment systems to a common set of K12 standards.

I am encouraged by the active and ongoing dialogue occurring across the nation, however, I am worried about the implementation of such assessment systems if they are not well designed, **well funded**, understood by front-line implementers, and are punitive in nature.

As a practitioner, I would support assessment systems that are innovatively designed and incorporate the use of technology across content areas which have the greatest potential to increase the rates of accessibility and participation of all students including students with disabilities and English language learners. I believe such systems should pursue measurements of student outcomes that are:

- Universally designed
- Performance-based
- Embedded in curriculum

Universally designed assessment systems would include formats that have multiple means of presentation, expression, and engagement. Such assessments would support the types of accommodations and modifications that many students with disabilities need to access instruction as well as assessments. I would further ask that approved accommodations do not invalidate a test and be permitted for use in both local and state assessments.

Expand your research in technology, which has flexibility and capacity to support authentic assessments, and can become integrated into a state's assessment system. Universally designed assessment systems may incorporate a variety of accessibility tools and features, which are customizable for each student, and would permit students to use those tools and/or features needed while taking a test.

Expand your research and determine the best approaches for educators to use performance-based assessments as a way to gain 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. This is sound instruction; it is what teachers do everyday. It is what I do everyday. Blending such sound instructional methods with existing technology such as a computer-based assessment is more likely to increase access and participation of students with disabilities.

Consider expanding the use of assessments embedded in curriculum, which allows 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. Again, this is what teachers do and validates current practice.

All of this will require appropriate, not *adequate,* but appropriate levels of funding across multiple levels – district, state, and national. Educators will need training at both the pre-service level with ongoing training at the in-service level to match changes in technology as they occur. Technology adaptations require infrastructures to be in place, hardware and software updates, maintenance, and personnel to monitor data and outcomes in a timely manner.

I believe if we work together on this, we can increase access and the rates of participation of students with disabilities in assessment systems. I know the AFT is committed 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. I thank you and appreciate this opportunity to speak on behalf of the United Federation of Federation of Teachers and our national organization, the American Federation of Teachers. If you need additional information, I may be reached at jdownes@ufttc.org or Dr. Lisa Thomas at lthomas@aft.org.

Respectfully,

Jeannette Downes
PS 138 UFT Teacher Center
Manhattan, New York

Race to the Top Fund Meeting
Assessment of Students with Disabilities
Registered Speaker: Stephen N. Elliott, PhD
Atlanta, GA / November 18, 2009

Specific Topics & Questions Addressed

As written in the Federal Register (Vol. 74, No. 204) on Friday October 23, 2009 under Notices, I am addressing two specific topics with my oral comments.

- Design of Assessment Systems-Required Characteristics part (2) *Be accessible to the broadest possible range of students, with appropriate accommodations for students with disabilities and English language learners;*
- 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?

If time allows, I would also like to respond to the multi-part Question 3 regarding Project Management. It reads as follows:

- 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?

Focus of Comments

My comments virtually all concern **access for students with disabilities;** access to an enacted curriculum that is well aligned with the intended grade level curriculum (i.e., content standards) and access to the assessments aligned with both these curricula. I wish to assert that **optimal accessibility** is rarely, if ever achieved, on our current large-scale assessments. This is the result of at least four factors: (1) teachers' limited training and use of materials that facilitate access to the intended curriculum, (2) students having insufficient opportunities to learn the intended curriculum, (3) testing accommodations being implemented inconsistently or with poor integrity, and (4) tests that are comprised of many items with extraneous information which results in construct irrelevant variance. *Figure 1* depicts how these four factors create barriers along the pathway to optimal access.

Over the past decade, I (in collaboration with colleagues) have conducted research on two of these factors (1) students' opportunity to learn the breadth and depth of the intended grade-level curriculum and (2) tests characterized by a preponderance of items with substantial extraneous information that creates unnecessary cognitive load which interferes with the valid measurement of the grade-level construct. I have also published eight experimental studies of testing accommodations, but these studies did not manipulate implementation integrity. *Figure 2* illustrates actions or strategies based on research that can be used to overcome each of the four barriers to access.

Figure 1

Barriers to the General Curriculum that Limit Access for Students with Disabilities

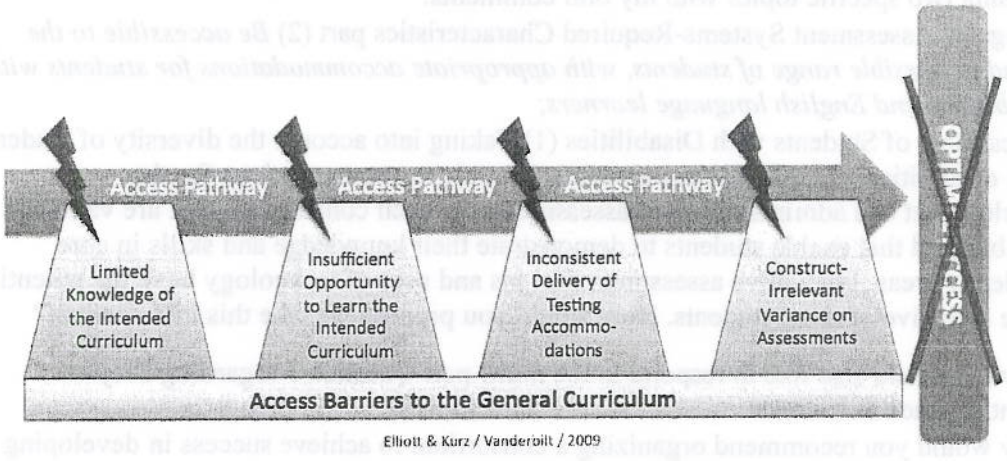
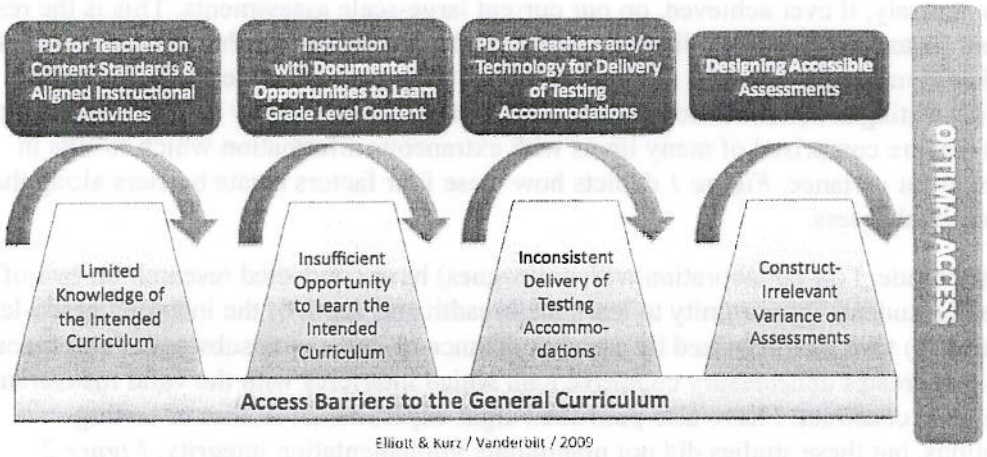


Figure 2

Access Actions To Overcome Barriers to the General Curriculum for Students with Disabilities



Recommendations for Improving Access & Measurement of Students' Academic Achievements

I recommend that consortia involved in developing and validating large-scale assessments with Race to the Top funding should be required to demonstrate (a) that their assessments are based on content that students have had opportunities to learn, (b) the enacted curricula for students with and without disabilities are comparable, and (c) the items on statewide achievement tests of reading and mathematics are highly accessible.

To achieve these quality assessment goals, I recommend:

1. The Required Characteristics for the Design of Assessment Systems listed for Race to the Top Assessments add the following characteristic: Contain only content which has been empirically documented that students with and without disabilities have had a reasonable opportunity to learn.

2. The second Required Characteristics for the Design of Assessment Systems listed for Race to the Top Assessments be revised to separate the concepts of item accessibility and testing accommodations, thus facilitating clarity and documentation efforts. The revised characteristics should read: (2a) Contain only test items which have been empirically determined to meet high standards of accessibility for students with disabilities and English language learners. (2b) Be compatible with accommodations proven to facilitate access and improve the validity of the intended test score inference.

3. To advance the development of assessments that are sensitive to the diversity of students with disabilities that all items undergo an Accessibility Review by an expert panel using a structured process coupled with item indices (i.e., item difficulty, item discrimination, readability estimates, depth of knowledge index, and distractor response frequencies for high and low performing groups of students). Such an Accessibility Review is intended to have similar weight to well-established Fairness Review requirements.

4. To advance the administration of assessments that are sensitive to the diversity of students with disabilities that individuals providing testing accommodations to students document the actual accommodations used during the testing event and note (a) which ones were essential for use in future assessments and (b) which ones were difficult to implement with integrity.

The theoretical foundation and the majority of the research that supports these recommendations have been directly influenced by universal design (UD) principles and testing accommodation research. Quality research on testing accommodations has been available for several years (e.g., Lang, Elliott, Bolt, & Kratochwill, 2008; Sireci, Scarpati, Li, 2005) and will not be reviewed in this document; however, much more research is needed to advance UD theory and practices. I believe the work of research teams like mine at Vanderbilt brings the federally endorsed UD principles to live in very practical and measureable ways. An examination of key concepts and research is summarized later in this paper and web links are provided to assessment and validation research resources resulting from three US Department of Education funded projects.

Background for Recommendations Concerning Curricular Access and Test Accessibility*

Key terms and concepts in this presentation are: *access*, *opportunity to learn, accommodations*, *modifications,* and *test accessibility*. These terms have all have been used for more than a decade when discussing educational testing and the validity of resulting test scores. They represent key concepts in the world of testing and federal assessment and accountability policies.

Current federal legislation requires the application of universal design principles to the development of all state and district-wide achievement tests. *Universal design*, as defined in the Assistive Technology Act (P.L. 105-394, 1998), is "a concept or philosophy for designing and delivering products and services that are usable by people with the widest possible range of functional capabilities, which include products and services that are directly usable (without requiring assistive technologies) and products and services that are made usable with assistive technologies" (§3(17)). While the term *accessibility* is not used in this definition, universal design principles as applied to assessment technology clearly are intended to address issues of access while responding to the concern raised in the *Testing Standards* that the use of individualized accommodations may increase measurement error. The legislation provides the rationale for the use of universal design principles as follows:

> The use of universal design principles reduces the need for many specific kinds of assistive technology devices and assistive technology services *by building in accommodations* for individuals with disabilities before rather than after production. The use of universal design principles also increases the likelihood that products (including services) will be compatible with existing assistive technologies. These principles are increasingly important *to enhance access* to information technology, telecommunications, transportation, physical structures, and consumer products ((PL105-394(§3(10)); emphasis added).

In the next two sections of this document, I provide overviews of *opportunity to learn* and *test accessibility* concepts and research issues.

Opportunity to Learn and Access to the General Curriculum

Comparable opportunity to learn (OTL) the intended curriculum (i.e., grade-level content standards) represents a fundamental assumption in the NCLB accountability system. Approximately 85% of students with disabilities are included in the same large-scale assessments as their general education peers and their resulting test scores are used to draw inferences about students' academic achievement and the extent to which teachers and schools can be held accountable for the observed achievement. Although federal legislation has underscored the rights of students with disabilities to have access to the general curriculum and instructional content aligned with standards (Individuals with Disabilities Act [IDEA], 1997; No Child Left Behind [NCLB], 2001), little is known about the extent to which standards-based instruction at grade level is delivered to student with disabilities. Not knowing whether students with or without disabilities are afforded the opportunity to learn the instructional content for which they are held accountable leaves fundamental questions regarding educational equity and the validity of test score interpretations unanswered. Evidence of only the alignment between content standards and test content does not address the OTL question. Alignment, as commonly measured, is a limited proxy for students' opportunity to learn. <u>In short, the validity of test score inferences for the present accountability system is error prone without further documentation of all students' opportunity to learn (OTL) the intended curriculum.</u>

Over the last four decades, researchers interested in school and classroom process indicators of OTL have focused repeatedly on three specific instructional variables: (a) *time on instruction* (e.g., Fisher & Berliner, 1985; O'Sullivan, Ysseldyke, Christenson, & Thurlow, 1990; Vannest & Hagan-Burke, 2009); (b) *content of instruction* (e.g., Husén, 1967; Porter, 2002; Rowan & Correnti, 2009); and (c) *quality of instruction* (e.g., Brophy & Good, 1986; Pianta & Hamre, 2009; Wang, Haertel, & Walberg, 1993). In addition, researchers have provided empirical support for the relation between each of these teacher variables and student achievement (e.g., Gamoran, Porter, Smithson, & White, 1997; Rowan, 1998; Thurlow, Ysseldyke, & Graden, 1984; Wang, 1998).

Research data on the relation between OTL and student achievement for student with disabilities is presently very limited. Kurz , Elliott, Wehby, and Smithson (2009) examined students' opportunity to learn the intended curriculum via the SEC's alignment methodology. The relation between alignment and student achievement averages was examined separately for general and special education teachers. The results indicated that the content of instruction delivered by general and special education teachers did not differ significantly between the two groups. The correlation between OTL and (class averages of) student achievement was .64 ($p < .05$). When general and special education teachers were examined separately, the correlation between alignment and achievement remained significant only for the special education group with .77($p < .05$). Elliott, Roach, Kettler, & Beddow (2008) and colleagues in a cognitive lab study used to validate items for a new modified achievement test in Arizona documented students with disabilities receiving instruction on grade-level mathematics in the general education classroom were provided significantly fewer sessions on state performance objectives. This data was collected using teachers' self-reports about the amount of instruction provided for individual students with and without disabilities. In general, students with disabilities received a third fewer instructional sessions on performance objectives listed in state grade-level content standards and subsequently tested.

The paucity of research notwithstanding, documentation of OTL is particularly important in the context of special education. The inclusion of students with disabilities in standards-based reform has established legal mandates that guarantee them access to general curriculum and content aligned with standards (IDEA, 1997; NCLB, 2001). Special education teachers (whose students are participating in the same assessments as their general education peers) are thus required to provide their students with the opportunity to learn the assessed and intended curriculum. By virtue of being administered the *same* assessments, students with disabilities have to be afforded at least the *same* opportunity to learn the content they are expected to know on these tests, regardless of where or by whom they receive their instructional content. Failure to do so undermines educational equity and the validity of test score interpretations. The latter, of course, is critical to the assessment and accountability provisions of standards-based reform: test scores are presumed to provide accurate information on the extent to which students have attained the standards and about how well teachers and schools have promoted student achievement to that end.

The need to document OTL in the context of special education is further strengthened by a lack of systematic evidence about the effects of inclusion in large-scale assessments on the opportunities afforded to students with disabilities (Koretz & Barton, 2003). Although current data on the assessment of student achievement against standards for purposes of NCLB permit

only limited conclusions, a common theme across states is discernible: students with disabilities are falling well below the performance objectives for AYP in reading and mathematics (Malmgren, McLaughlin, & Nolet, 2005). According to the National Center on Educational Outcomes, approximately 70% of students with IEPs who took the regular or alternate state assessment in reading and mathematics failed to score proficient according to state criteria (Thurlow et al., 2006). Additional issues regarding low expectations for students with disabilities (e.g., Wehmeyer, Sands, Knowlton, & Kozleski, 2002), highly variable quality of instruction (e.g., Wehby, 1997; Wehby, Symons, Canale, & Go, 1998; Wehmeyer et al., 2002), chronic underachievement (National Council on Disability, 2004), students' instructional under-preparedness for high-stakes graduation assessments (Thompson & Thurlow, 2003), and low teacher time on instruction (Vannest & Hagan Bruke, 2009) further underscore the need to document OTL for students with disabilities.

Test Accessibility and the Validity of Achievement Score Inferences

For educational testing, *access* is the opportunity for test-takers to demonstrate proficiency on the target construct of a test (e.g., language arts, mathematics, science) or a test item (e.g., synonyms, homonyms, and homographs). Complete access is manifest when a test-taker is able show the degree to which he or she knows the tested content. Access, therefore, must be understood as an interaction between individual test-taker characteristics and features of the test.

The purpose of both *testing accommodations* and *modifications* is to increase individuals' access to tests. The definitions of these two access-enabling strategies, however, have been the subject of debate, in part because of their inconsistent use in the *Standards for Educational and Psychological Testing* (*Standards*; AERA, APA, & NCME, 1999) and some states' testing guidelines. Recently, the inclusion of the term *modified* in the AA-MAS has added to the debate about these terms. An examination of the *Standards for Testing* finds the term *modification* is cited nearly a dozen times in the index. A number of the standards refer to a modification as an accommodation. For example, in the *Testing Standards'* section entitled "Testing Individuals with Disabilities," it is stated:

> The terms *accommodation* and *modification* have varying connotations in different subfields. Here accommodation is used as the general term for any action taken in response to a determination that an individual's disability requires a departure from established testing protocol. Depending on circumstances, such accommodation may include modification of test administration processes or modification of test content. No connotation that modification implies a change in construct(s) being measured is intended. (AERA et al., 1999, p. 101)

The implementation of testing accommodations for students with disabilities is a universally endorsed policy in all states. The modification of test content, however, is inconsistent with the definition of a testing accommodation in the majority of state testing accommodation guidelines (Lazarus, Thurlow, Lail, Eisenbraun, & Kato, 2006). Accommodations are widely recognized in state testing guidelines as changes to the setting, scheduling, presentation format, or response format of an assessment. Accommodations are made to increase the validity of inferences that can be made from a student's scores, so that those scores can be meaningfully compared to scores of students for whom testing accommodations are not needed.

The federal policy on AA-MAS policy extends the notion of access and the spirit of individualized accommodations to changes made to item content. Such changes are defined as modifications by most test developers and administrators. When item and test alterations are made (e.g., by changing the layout, reducing the length of the reading passage, adding graphic support), it is not always clear without test results whether the changes affect only access to the test or, in fact, also affect the construct being measured and the subsequent inferences that are drawn from scores. If the content of an item or test has been changed and the scores are no longer comparable to the original construct to be measured, it has been customary to consider the alteration a *modification* (Phillips & Camara, 2006; Koretz & Hamilton, 2006). To ensure the modifications are acceptable under AA-MAS policy, research is needed to confirm the modified test measures the same construct(s) as the original test. Indeed, the policy assumes AA-MAS for eligible students measure the same grade-level content standards and performance objectives (constructs) as general assessments for students without disabilities. It should be noted that some modifications may result in an AA-MAS that yields scores that are not comparable to scores obtained from the original test. These modifications could still be permissible for an AA-MAS, assuming the content is commensurate with the intended grade-level of the test (Kettler, Russell, et al., 2009).

In summary, the application of the term *modification* to the context of recent legislative regulation is not acceptable to all parties and admittedly the use of these terms is subject some controversy. There does not, however, exist a consensus alternative term for alterations made to item or test content that has been shown to result in tests of unchanged constructs that yield scores from which comparable inferences can be made. Given the approach we recommend for developing an AA-MAS, we have used the term *modification* to refer to a process by which the test developer starts with a pool of existing test items with known psychometric properties, and makes changes to the items, creating a new test with enhanced accessibility for the target population. When analyses indicate inferences made from the resulting test scores are valid indicators of grade-level achievement, the modifications are appropriate. Conversely, if analytic evidence suggests the inferences made from resulting scores are invalid indicators of grade-level achievement, the modifications are inappropriate. Thus, like individualized testing accommodations, modifications must be studied to determine their appropriateness. Unlike accommodations, modifications as conceptualized within AA-MAS policy are intended to afford an entire group of students access and better measurement of their achieved knowledge.

Test accessibility is the extent to which a test and its constituent item set permits the entirety of the test taking population to demonstrate proficiency on the target construct(s) of the test. To the extent a test demands physical, material, or cognitive resources in excess of those required to measure the intended construct, inferences made from the scores on the test likely reflect in some part the test taker's degree of access to the test (Beddow, Kettler, & Elliott, 2009). While accessibility is a characteristic of the test itself and not of the test-taker or test user, it is dependent both on individual differences among test-takers and on factors controlled by test users. The use of non-standard changes in test conditions (i.e., accommodations) to moderate the negative effects of individual test-taker needs on test results can reduce the validity of inferences. According to the *Standards for Educational and Psychological Testing*: "Each step toward greater flexibility almost inevitably enlarges the scope and magnitude of measurement error. However, it is possible that some of the resultant sacrifices in reliability may reduce construct irrelevance or construct underrepresentation in an assessment program" (AERA, APA, &

NCME, 1999, p. 26). Thus, an ideal assessment is one that can be administered under the same conditions across the population, the results of which yield inferences that are equally valid across the range of that population.

In anticipation of aforementioned regulations under NCLB, permitting states to develop AA-MAS, the CAAVES project team conducted a series of studies to examine the differential effects of item modifications on the performance of students across disability and eligibility groups. To assist in the item modification process with the goal of developing more accessible tests, Beddow, Kettler, and Elliott (2008) developed the Test Accessibility and Modification Inventory (TAMI). The TAMI was designed to assist with the process by which general assessment items are modified in response to the demand for universally-designed assessments to be used for the AA-MAS. Thus, the instrument has two purposes: the first is to meaningfully quantify the construct of accessibility with regard to tests and items; the second is for use as a training tool to guide item writers in the development and modification of tests and items that are highly accessible. To wit, the development of the TAMI was influenced by (a) universal design principles for learning and assessment, (b) cognitive load theory, (c) research on test and item development, and (d) guidance on web and computer accessibility.

Universal Design for Learning and Assessment. The Center for Universal Design (CUD; 1997) lists seven primary aspects of universal design: (a) equitable use, (b) flexibility in use, (c) simple and intuitive use, (d) perceptible information, (e) tolerance for error, (f) low physical effort, and (g) size and space for approach and use. The last decade has seen a shift in focus toward applying universal design principles across the educational arena, called universal design for learning (UDL; e.g., see Rose & Meyer, 2006) and universal design for assessment (UDA). Several documents from the National Center on Educational Outcomes informed the development of the TAMI, specifically with respect to UDA principles (e.g., Thompson, Johnstone, & Thurlow, 2002; Johnstone, Thurlow, Moore, & Altman, 2006).

Cognitive Load Theory. As conceptualized by Sweller and his colleagues, cognitive load theory (CLT; Chandler & Sweller, 1991) is a logical extension of Miller's (1956) work on processing limitations and working memory, exemplified in Miller's famous "Magical Number Seven, Plus or Minus Two" article. Based on an information-processing framework, CLT has heretofore been conceived as a model for understanding the demands of learning tasks and is grounded in three primary assumptions about how the mind works: (a) the mind processes information through two separate channels: the auditory/verbal channel and the visual channel; (b) each of these channels has a limited capacity for processing information; (c) both channels are required for most learning tasks. Proponents of CLT posit that to properly gain knowledge from instruction, students must: (a) attend to the presented material, (b) mentally organize the material into a coherent structure, and (c) integrate the material with existing knowledge. Thus, the efficiency of instructional tasks depends on the extent to which element interactivity and the requisite cognitive resources needed to process essential information are minimized. Accordingly, CLT disaggregates the cognitive demands of learning tasks into three load types: *intrinsic load*, *germane load*, and *extraneous load*. Intrinsic load refers to the amount of mental processing that is requisite for completing a task. Germane load refers to cognitive demands that are not necessary for gaining essential knowledge, but enhance learning by facilitating generalization or automation (e.g., lessons that require learners to extend learned concepts to arenas outside the classroom or apply them to novel situations). Extraneous load refers to the demand for cognitive

resources to attend to and integrate nonessential elements that are preliminary to actual learning, but are nonetheless required for a learning task. Proponents of CLT argue that learning tasks should be designed with the goal of minimizing the demand for cognitive resources that are extrinsic to the goals of instruction. The triune model of cognitive load was encapsulated by Paas, Renkl, and Sweller (2003): "Intrinsic, extraneous, and germane cognitive loads are additive in that, together, the total load cannot exceed the working memory resources available if learning is to occur"(p.2).

Notwithstanding the broad overlap between instruction and testing, CLT heretofore has had little research application to school-age students with or without special needs or to the assessment of student learning. Considering the numerous similarities between instructional tasks and the variety of tasks required in many forms of tests, the development of the Test Accessibility and Modification Inventory or TAMI (Beddow et al., 2008) focused explicitly on the degree to which cognitive load demands may impact a test-taker's ability to demonstrate performance on assessments. Particular attention was paid to how CLT has been used to understand the cognitive demands of multimedia learning. To the extent the cognitive demands of an assessment are intrinsic to the target constructs of the assessment, inferences made from test results are likely to represent the person's actual competence on the constructs. Extraneous load demands by an assessment item interfere with the test-taker's capacity to demonstrate performance on the target construct, and should be eliminated from the assessment process.

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" (Elliott, Rodriguez, Roach, & Kettler, 2009). These item reviews were part of two USDE funded projects (CMAADI and OAASIS) and involved multiple independent ratings with the TAMI ARM of multiple-choice reading, math, and science items representative of 4th, 8th, and high school operational tests. The trained raters frequently reported that items passages or stimuli contained numerous extraneous words, had visuals that were poorly integrated with the other item elements, and offered excessive numbers of implausible answer choices. Related research by Beddow (2009), a Vanderbilt researcher, also indicated that overall accessibility ratings correlated .70 with the difficulty indices for operational reading items and .35 with the difficulty indices of operational math items. These findings, if replicated across other state tests, provide further evidence that reading test scores, in particular, may be influenced by substantial construct irrelevant variance.

Efforts to improve the accessibility of test items have been recognized by a number of other educators and psychometricians. For example, a report summarizing work by the National Accessible Reading Assessment Projects (Thurlow, Laitusis, Dillion, Cook, Moen, Abedi, and O'Brien, 2009) documenting the need for more accessible tests and provided five principles to guide the development of such tests. These principles are:

- Reading assessments are accessible to all students in the testing program, including students with disabilities.
- Reading assessments are grounded in a definition of reading that is composed of clearly specified constructs, informed by scholarship, supported by empirical evidence, and attuned to accessibility concerns.
- Reading assessments are developed with accessibility as a goal throughout rigorous and well-documented test design, development, and implementation procedures.

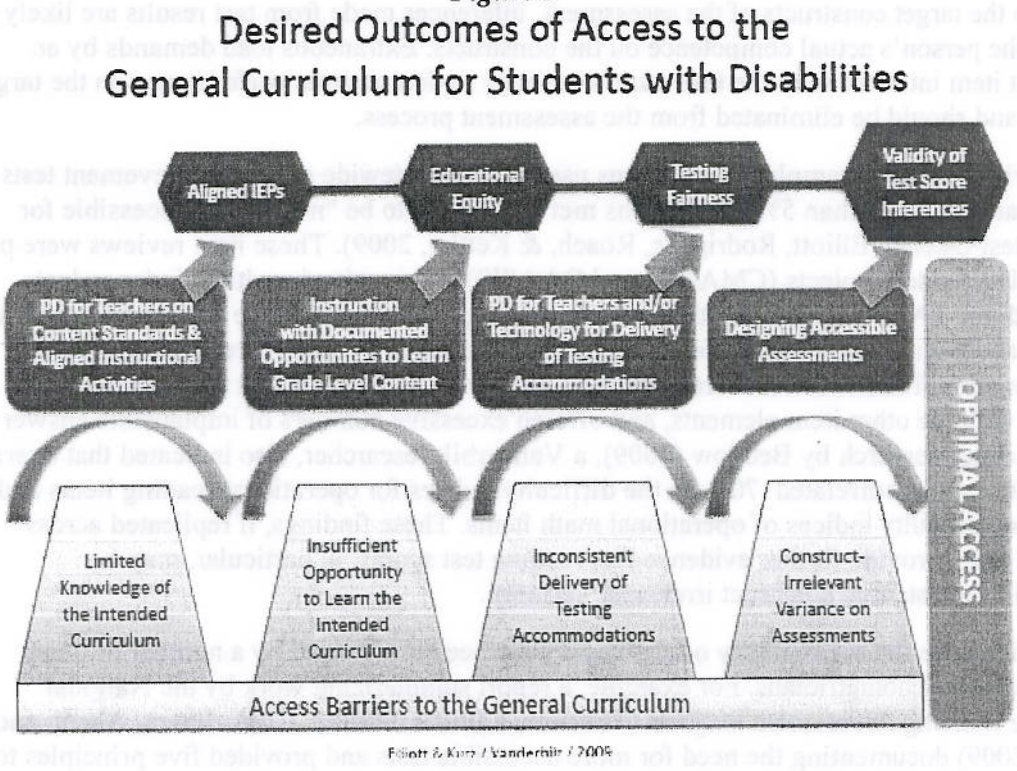
- Reading assessments [*that are highly accessible*] reduce the need for accommodations, yet are amenable to accommodations that are needed to make valid inferences about a student's proficiencies.
- Reporting of reading assessments results is designed to be transparent to relevant audiences and to encourage valid interpretation and use of these results.

(Thurlow, et al., 2009, page 4)

These accessibility principles, along with the use of practical tools like the TAMI ARM, should guide the design and evaluation of test items to ensure they meet high standards for accessibility. Without such a commitment to accessibility, it is doubtful new assessments will advance the inclusive practices demanded by federal legislation or overcome barriers to valid assessment results for students with disabilities.

Figure 3 summarizes my vision of the work ahead and the desired outcomes for all involved in the next generation of large-scale assessments.

Figure 3



Published Research and Technical Reports to Support Access Recommendations

Support for my claims comes from a number of research studies conducted by Vanderbilt and University of Minnesota researchers in collaboration with state department of education leaders in states like Arizona, Hawaii, Idaho, Indiana, Mississippi, South Carolina, South Dakota, Wisconsin, and Wyoming. Key conceptual papers and published research studies used for this presentation are listed below. Three articles from which a majority of my comments about access are based are indicated with an *.

Center for Universal Design. (1997). *The principles of universal design*. Retrieved August 4, 2008 from http://www.design.ncsu.edu/cud

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*, 293-332.

Clark, R., Nguyen, F., & Sweller, J. (2006). *Efficiency in learning: Evidence-based guidelines to manage cognitive load.* San Francisco, CA: Pfeiffer.

*Beddow, P.A., Kettler, R.J., & Elliott, S.N. (2009). *TAMI: Test Accessibility and Modification Inventory.* Nashville, TN: Vanderbilt Univerity.

*Elliott, S.N., Kettler, R.J., Beddow, P.A., Kurz, A., Compton, E., McGrath, D., Bruen, C., Hinton, K., Palmer, P., Rodriguez, M., Bolt, D., & Roach, A.T. (in press). Effects of using modified items to test students with persistent academic difficulties. *Exceptional Children.*

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15(3)*, 309-344.

*Kettler, R.J., Elliott, S.N., & Beddow, P.A. (2009). Modifying achievement test items: A theory-guide and data-based approach for better measurement of what students with disabilities know. *Peabody Journal of Education.*

Kettler, R.J., Elliott, S.N., Beddow, P.A., Compton, E., McGrath, D., Kaase, K., Bruen, C., Ford, L., & Hinton, K. (in press). What do alternate assessments of alternate achievement standards measure? A multitrait-multimethod analysis across six states. *Exceptional Children.*

Kettler, R.J., Russell, M., Camacho, C., Thurlow, M., Ketterlin Geller, L., Godin, K., et al. (2009). Improving reading measurement for alternate assessment: Suggestions for designing research on item and test alterations. Manuscript in revision for publication.

*Kurz, A., Elliott, S.N., Wehby, J., & Smithson, J. (in press). Alignment of the intended, planned, and enacted curriculum in general and special education and its relation to student achievement. *Journal of Special Education.*

Lang, S.C., Elliott, S.N., Bolt, D.M., & Kratochwill, T.R. (2008). The effects of testing accommodations on students' performances and reactions to testing. School Psychology Quarterly, 14 (3), 107-124.

Roach, A.T., Beddow, P., Kurz, A., Kettler, R.J., & Elliott, S.N. (in press). Using student responses and perceptions to inform item modification for an alternate assessment based on modified achievement standards. *Exceptional Children.*

Rodriguez, M.C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*, 3-13.

Rodriguez, M.C. (2009, April). The role of item response attractors in modified items for students with disabilities. In S.E. Elliott (Chair), *Designing and evaluating modified items for students with disabilities: Research results.* Symposium conducted at the annual meeting f the National Council on Measurement in Education, San Diego, CA.

Sireci, S.G., Scarpati, S.E., & Li, S. (2005). Testing accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, Winter 2005, Vol. 75 (4),* 457-490.

Thurlow, M.L., Laitusis, C.C., Dillion, D.R., Cook, L.L., Moen, R.E., Abedi, J., & O'Brien, D.G. (September 2009). *Accessibility principles for reading assessments*. Minneapolis, MN: National Accessible Reading Assessment Projects. (available: www.narap.info)

About the Speaker

Stephen N. Elliott, PhD, is a Professor of Special Education and the Dunn Family Chair of Educational and Psychological Assessment in Peabody College at Vanderbilt University. He is the Director of the Learning Sciences Institute, a trans-institutional center for externally funded research and also directs Peabody College's Interdisciplinary Program in Educational Psychology. Steve teaches courses on measurement and assessment and his research focuses on (a) test development and validation and (b) the use of testing accommodations and alternate assessment methods for evaluating the academic performance of students with disabilities for purposes of educational accountability. He currently co-directs three USDE research grants concerning test accessibility and the validity of item modifications on the performance of students with disabilities. Another IES funded research project he directs concerns the validation of scores for the Vanderbilt Assessment of Leadership in Education, a measure of principal's learning focused leadership behaviors. Steve has worked with 9 states on the design and validation of alternate assessments (AA-AAS and AA-MAS) for students with disabilities. He was recently honored by the American Psychological Association as a Senior Scientist for this assessment research on students with disabilities. He can be contacted at steve.elliott@vanderbilt.edu.

Acknowledgements

I wish to acknowledge my Vanderbilt research team and colleagues' contributions to this presentation. In particularly, **Alexander Kurz** contributed to the insights and evolving research on the assessment of students' opportunity to learn and **Peter Beddow** and **Ryan Kettler** have offered a number of important insights and helped create a method for measuring test item accessibility. **Andrew Roach**, now at Georgia State University, also has consistently helped me advance my understanding of teachers' assessment lives and the measurement of access to the general curriculum. Our collaborative work is guided by what we have come to call *Accessibility Theory*; a powerful organizing framework for examining the interactions among instructional and test access barriers and students' striving to learn and demonstrate what they know and can do. Collectively, these colleagues and I have benefited from the ongoing support of the USDE via Enhanced Assessment Grants and General Supervision Enhancement Grants. The research from these various projects can all be accessed by visiting the Learning Sciences Institute's website at Vanderbilt University.

Relevant Website Links

http://www.vanderbilt.edu/lsi

http://peabody.vanderbilt.edu/tami.xml

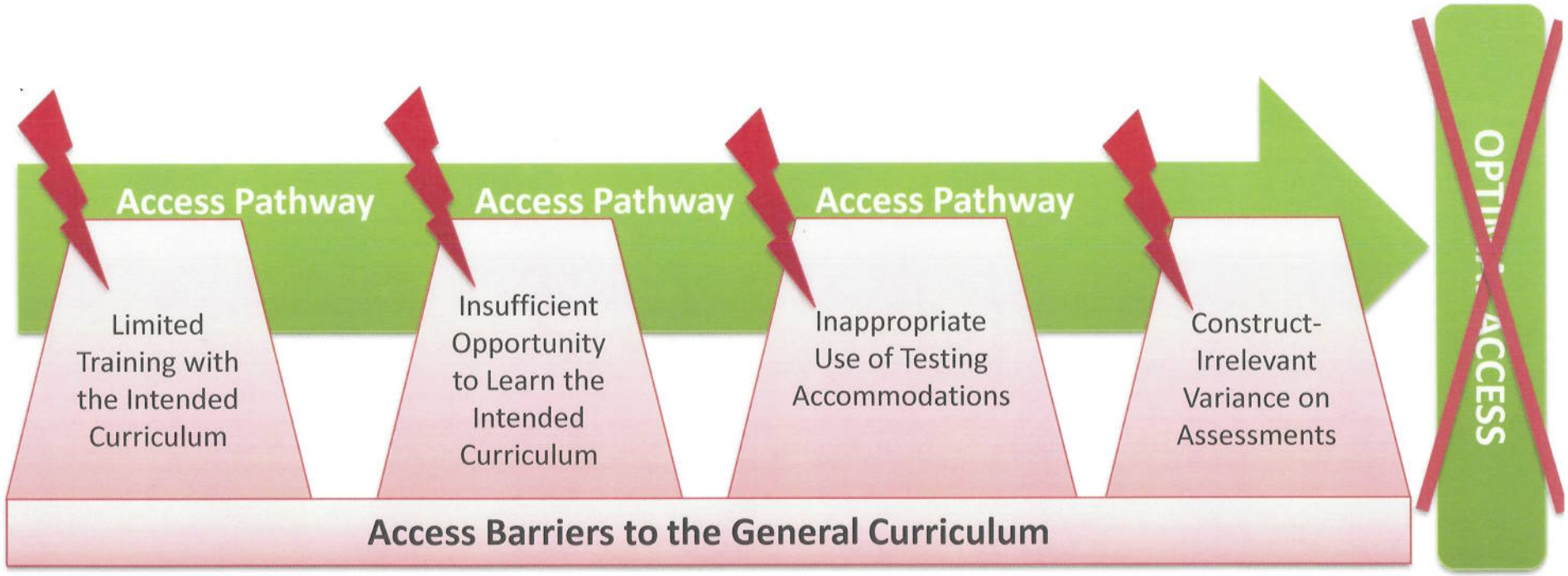
http://peabody.vanderbilt.edu/LSI_Projects/C-MAADI_Project_Home.xml

http://peabody.vanderbilt.edu/LSI_Projects/CAAVES_Project_Home.xml

http://peabody.vanderbilt.edu/LSI_Projects/MAAPS_Consortium_Project_Home.xml

VANDERBILT UNIVERSITY Peabody College

Figure 1.
Barriers to the General Curriculum that Limit Access for Students with Disabilities

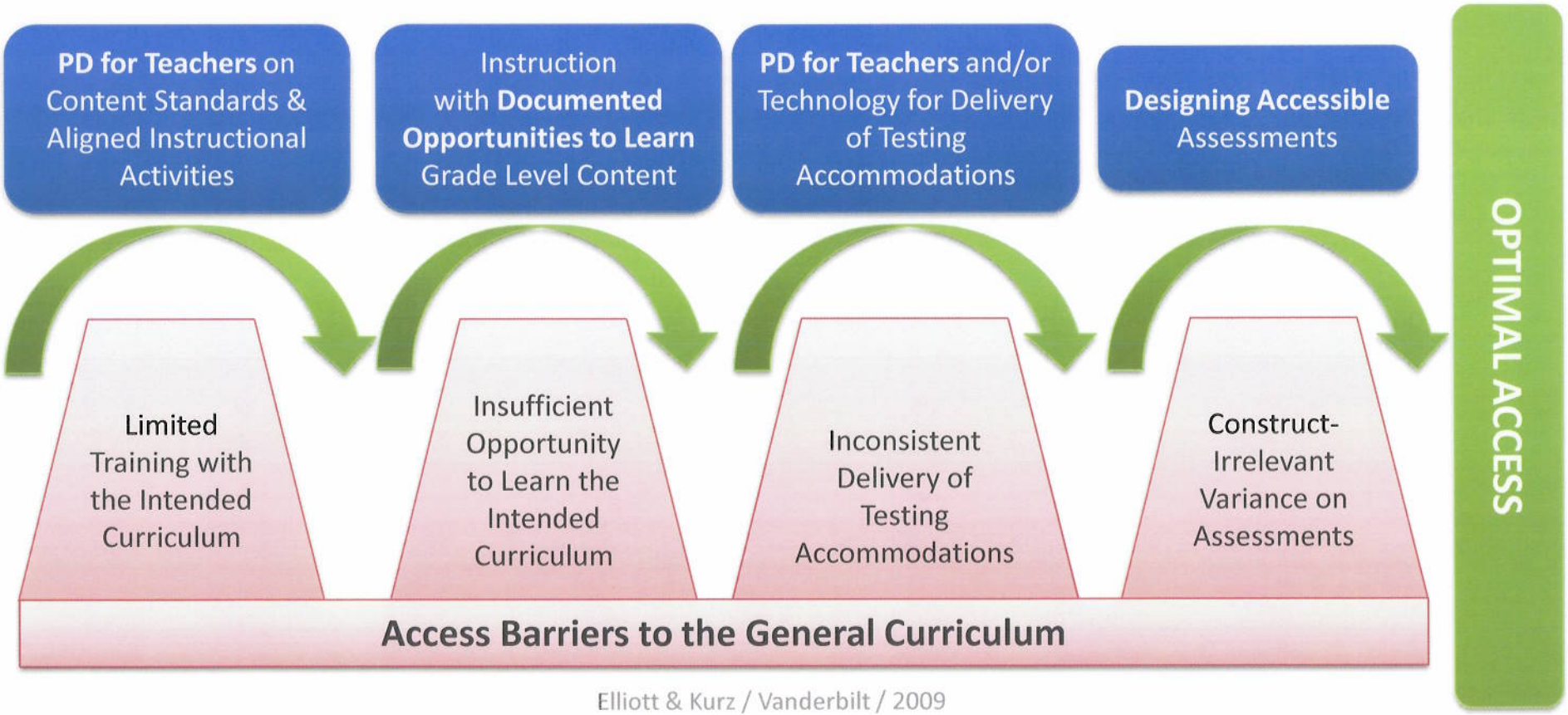


Elliott & Kurz / Vanderbilt / 2009

VANDERBILT UNIVERSITY Peabody College

Figure 2.

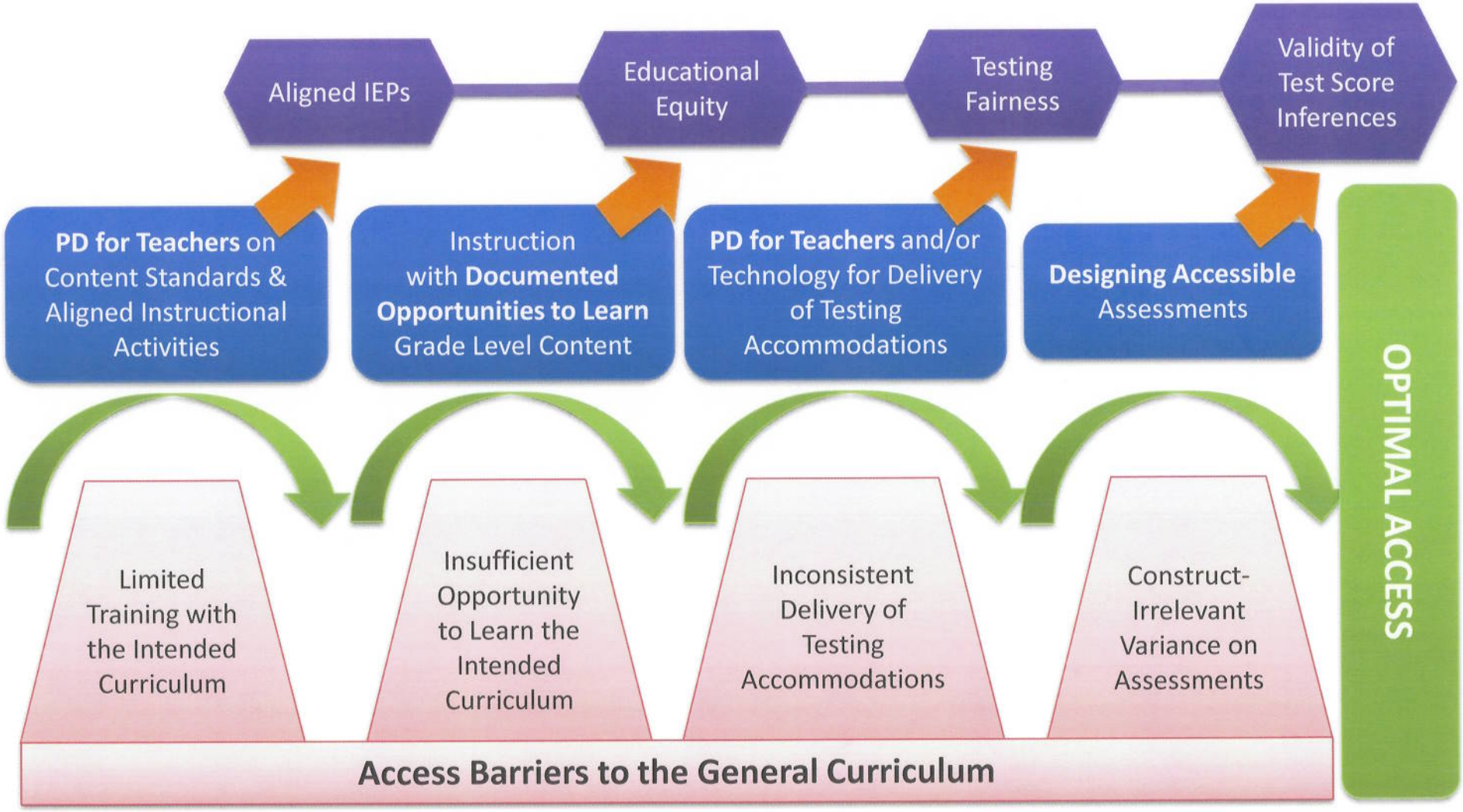
Access Actions to Overcome Barriers to the General Curriculum for Students with Disabilities



Elliott & Kurz / Vanderbilt / 2009

VANDERBILT UNIVERSITY Peabody College

Figure 3. Desired Outcomes of Access to the General Curriculum for Students with Disabilities



Elliott & Kurz / Vanderbilt / 2009

Test Accessibility and Modification Inventory

TAMI™

Accessibility Rating Matrix

Technical Supplement

Peter A. Beddow, Stephen N. Elliott, & Ryan J. Kettler

VANDERBILT UNIVERSITY

November 2009

Purpose and Rationale

The Test Accessibility and Modification Inventory (TAMI) is a multi-purpose training and evaluation tool with two components. First, the inventory component of the TAMI provides practical guidance for item writers and test developers to ensure tests and test items are optimally accessible. Second, the Accessibility Rating Matrix (ARM) is a measurement tool for evaluating the accessibility of tests and test items accessibility and suggesting modifications that may improve accessibility for more test-takers. The authors of the TAMI argue that the by applying the principles of accessibility theory and systematically using the TAMI and TAMI ARM during the test development process will result in more accessible tests and improved measurement test-takers with a broad range of abilities and needs. Students identified with disabilities comprise approximately 10-12% of the student populations of most states. Common characteristics of these students include reading problems, attention difficulties, and feelings of anxiety and frustration with schoolwork and testing. As a result, many of these students experience persistent academic difficulties and learn at a rates slower than their peers. Additionally, there are an estimated 5-10% of students across states who have not been identified with disabilities but for whom the demands of school and tests are similarly challenging.

Recent federal legislation permits states to report proficiency for calculating AYP for up to 2% of students using results from an alternate assessment based on modified academic achievement standards (AA-MAS). To be eligible for participation in an AA-MAS, a student must be identified with a disability and must have scored below-proficient on the statewide assessment in prior years with the data-based prediction that he or she is unlikely to achieve proficiency on the general state assessment during the current year. When an IEP team agrees to permit the student to participate in the AA-MAS, the implication is that because of this student's special needs, the general state assessment is unlikely to yield scores from which valid inferences can be made about the student's knowledge of the academic content standards.

The Test Accessibility and Modification Inventory (TAMI; Beddow, Kettler, & Elliott, 2008) and the TAMI Accessibility Rating Matrix (Beddow, Elliott, & Kettler, 2009) represent the combined effort of a team of educators, assessment leaders, researchers and test developers across several states to evaluate the accessibility of test items with the goal of improving testing for all students.

Accessibility Theory

Accessibility theory provides a framework for improving tests for all individuals by offering a perspective on the measurement of this target construct in terms of three sets of variables: the test-taker, the test, and the test event. The theory is based on three primary areas of research and theory: universal design (UD) principles, cognitive load theory (CLT), and research on test and item design.

Current federal legislation requires the application of universal design principles to the development of all state and district-wide achievement tests. Universal design, as defined in the Assistive Technology Act (P.L. 105-394, 1998), is "a concept or philosophy for designing and delivering products and services that are usable by people with the widest possible range of functional capabilities, which include products and services that are directly usable (without requiring assistive technologies) and products and services that are made usable with assistive technologies" (§3(17)). While the term *accessibility* is not used in this definition, universal

© 2009 Vanderbilt University. All rights reserved.

design principles as applied to assessment technology clearly are intended to address issues of access while responding to the concern raised in the Testing Standards that the use of individualized accommodations may increase measurement error. Guidance from researchers at the National Center on Educational Outcomes (NCEO) provides recommendations for applying UD to testing (Thompson, Johnstone, & Thurlow, 2002; Johnstone, Thurlow, Moore, & Altman, 2006).

Originally conceived and applied as a model for understanding the demands of instructional activities, cognitive load theory has clear implications for the development of accessible tests. The model disaggregates the cognitive demands of learning tasks into three load types: intrinsic load, germane load, and extraneous load. The triune model of cognitive load was encapsulated by Paas, Renkl, and Sweller (2003) as follows: "Intrinsic, extraneous, and germane cognitive loads are additive in that, together, the total load cannot exceed the working memory resources available if learning is to occur"(p.2). With regard to assessment, application to CLT focuses on certain aspects of test items that may require the test-taker to use cognitive resources in excess of those needed to demonstrate knowledge of the tested content. Excessive cognitive load may be reduced by eliminating extraneous material from items, reducing reading load, and using visuals efficiently.

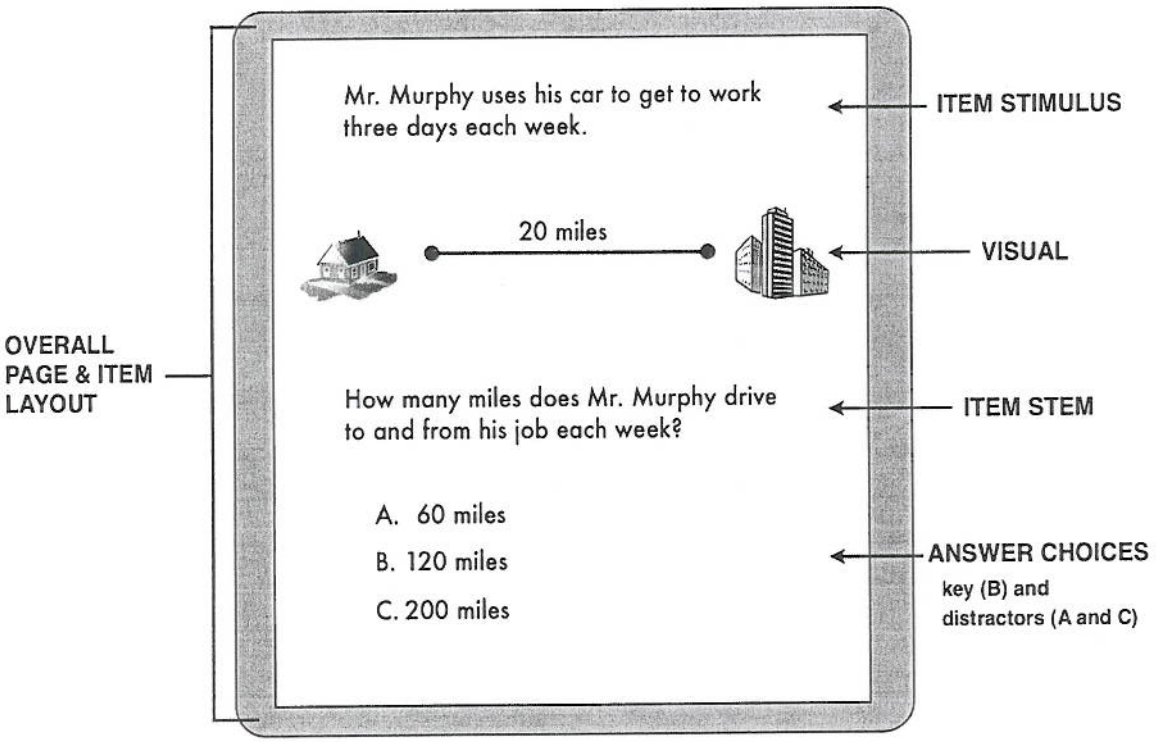
Further, evidence based on the collective expertise of test and item development scholars has informed the development of the TAMI and has advanced our understanding of test and test item accessibility. Among these is Haladyna, Downing, and Rodriguez's (2002) research-based taxonomy consisting of a number of recommendations for writing multiple-choice items. Another is Rodriguez's (2005) meta-analysis of over 80 years of research on item development which led to his conclusion that three answer choices are optimal for multiple-choice items. Rodriguez' results indicate that reducing items from 4 or 5 answer choices to 3 tends to result in small, nonsignificant changes in item difficulty, discrimination, and reliability, and such modifications are sensible in that the reduced reading load results in a test that is more manageable for a group of students who typically take much longer to finish. Likewise, the decreased reading load may permit test developers to add more items and increase the reliability of the test. Accordingly, the authors of the TAMI support the practice of reducing the number of response options from 4 to 3, when doing so eliminates an implausible distractor and results in a set of answer choices that are balanced and free of cueing.

Description of the Instruments

The TAMI and TAMI ARM are user-friendly tools that have been used by over 100 educators and assessment experts to improve the accessibility of test items. These tools provide item developers and evaluators a systematic structure for creating new items or enhancing existing items. The two parts of this pragmatic approach to improving tests for all students are described next.

© 2009 Vanderbilt University. All rights reserved.

Figure 1. ***Anatomy of a multiple-choice item.***



Test Accessibility and Modification Inventory. The inventory component of the TAMI consists of two sections: Item Analysis and Computer-Based Test Analysis. The TAMI Item Analysis section contains 46 descriptors organized into five primary categories based on the key elements of most test items: the item passage and/or stimulus, item stem, visuals, answer choices, and layout (see Figure 1.) An additional Fairness category consists of a distillation of the Fairness Review Guidelines (Educational Testing Service, 2009). The Computer-Based Test Analysis section follows the same format as the Item Analysis section and contains 35 descriptors across four categories: Test Delivery System, Test Layout, Training, and Audio. Guidance on web accessibility standards and current research and practice regarding computer-delivered tests (e.g., NimbleTools; see http://www.nimbletools.org) were primary influences for this section of the TAMI.

TAMI Accessibility Rating Matrix. The TAMI ARM consists of two scoring rubrics: the Item Analysis and the Overall Analysis. After writing the item number on the ARM Record Form, the rater begins by using the Item Analysis rubric to evaluate the accessibility of the item according to the key elements of a multiple-choice test item. While individual test items may or may not include each of these elements, the ARM was designed with the assumption that this anatomic structure can be used to evaluate most current assessment item formats.

Using the Item Analysis rubric, the rater determines the accessibility level of each item element on a 4-point scale (see Figure 2). After assigning a rating to each item element, the rater uses the Modification Guide to select modifications that are likely to improve the accessibility of

© 2009 Vanderbilt University. All rights reserved.

the item. Additionally, there is a space at the bottom of the ARM Record Form for raters to code suggested modifications that are not listed on the Modification Guide. After rating the individual item elements, the rater reviews the item element ratings and selects an Overall Accessibility rating using the Overall Analysis rubric.

Figure 2. ***TAMI accessibility levels.***

Level	Description	Heuristic
4	Maximally Accessible for Nearly All Test-Takers	Optimal accessibility for between 95-99% of the population
3	Maximally Accessible for Most Test-Takers	Optimal accessibility for between 90-95% of the population
2	Maximally Accessible for Some Test-Takers	Optimal accessibility for between 85-90% of the population
1	Inaccessible for Many Test-Takers	Optimal accessibility for fewer than 85% of the population

Evidence for the Reliability of Accessibility Ratings

Accessibility involves an interaction between item features and test-taker characteristics; raters, therefore, should have both expertise in the test content and an intimate understanding of the range of abilities and needs across the intended test-taker population. Levels of accessibility on the ARM are based on the extent to which the item is *maximally accessible* for a given portion of the intended test-taker population. A maximally accessible item is an item that contains no barriers that would limit or hinder the test-taker from demonstrating his or her knowledge of the target construct. An item that is maximally accessible for nearly all test-takers, therefore, requires few, if any, cognitive resources in excess of those needed to show what the test-taker knows. If the extraneous cognitive load demand of an item differentially impacts performance on the item across the test-taker population, the item's accessibility is less than optimal.

The ARM rubrics are intended to guide professional judgments about test accessibility rather by providing a flexible algorithm for assigning accessibility ratings to test items. It is essential, therefore, that raters are trained in evaluating accessibility prior to using the TAMI to ensure sufficient reliability of scores and validity of subsequent diagnostic inferences and recommendations. Based on the current data, the reliability of accessibility ratings by evaluators with extensive training in accessibility principles is high. Specifically, across three item reviews totalling over 350 test items, perfect or adjacent agreement for ratings by an evaluation team of assessment researchers and professors trained to use the ARM was over 94%.

Evidence for the Validity of Accessibility Ratings

To date, several studies have yielded empirical data supporting the use of TAMI to facilitate the development of accessible tests and contributing to an increased understanding of test accessibility. These studies have emerged from three federal grant projects aimed at supporting states' efforts to develop an AA-MAS for students with persistent academic

© 2009 Vanderbilt University. All rights reserved.

difficulties: (a) the Consortium for Alternate Assessment Validity and Experimental Studies (CAAVES; Elliott & Compton, 2006-2008), (b) the Consortium for Modified Alternate Assessment Development and Implementation (CMAADI; Elliott, Roach, & Rodriguez, 2008-2010) and (c) Operationalizing Alternate Assessment of Science Inquiry Skills (OAASIS, 2009-2011). Results have indicated students perform comparatively better on items modified with a focus on accessibility than on unmodified sibling items (Elliott, Kettler, Beddow, Kurz, Compton, et al., In press). Evidence also indicated the magnitudes of differences in item difficulty across modified items for students identified with disabilities who would be eligible for an AA-MAS are greater than those observed for their general education peers who would not be eligible for a modified assessment. Further, students who completed surveys and cognitive interviews following testing reacted positively to the item changes (Roach, Beddow, Kurz, Kettler, & Elliott, In press).

Additionally, results of two field tests of the TAMI indicate accessibility ratings are sensitive to items across a range of accessibility levels. In both studies, raters were randomly assigned a set of items to evaluate and results were analyzed to determine whether ratings were higher for items that had undergone modification to improve their accessibility. In the second of these studies, members of a team of raters who had recently received comprehensive training in the use of the ARM were randomly assigned to rate items from two item pools: (a) publicly-released items from the Texas Assessment of Knowledge and Skills (TAKS) and (b) corresponding items in modified form for the modified version of the TAKS (called the TAKS-M). Results indicated accessibility ratings of original (unmodified) versions of the items had significantly lower accessibility ratings than those on modified items. Further, ratings by the evaluation and training team corresponded highly with ratings by the novice raters, which is another indicator that when used by trained raters, the TAMI yields reliable ratings of accessibility.

Accessibility Reviews

Most states developing an AA-MAS have made modifications to the pool of items used for the general assessment to enhance their accessibility for the new test. The procedures used for modifying items vary across states and until recently, no state has attempted to quantify the accessibility of their assessment items. Without a thorough examination of the accessibility across a pool of test items, predictions about how the items may perform when used for the group of test-takers for whom the items are intended likely will be unreliable or invalid.

To date, the TAMI evaluation team has conducted several reviews of large item pools across a number of content domains and grade levels for several state education departments. For each review, the team examined each item in a representative item pool based on the five elements that are common to the majority of large-scale test items, including the item passage or stimulus, item stem, visuals, answer choices, and layout (Figure 1), noting patterns across the reviewed items and suggesting modifications that may improve the accessibility of the items for more test-takers. The team provided a report following each review to summarize features of the item pool that may enhance or reduce the accessibility of the items for portions of the test-taker population and to provide recommendations for improvement. Evidence suggests if test developers follow the team's suggested modifications in revising the items, the changes should result in optimally accessible items for nearly all test takers.

© 2009 Vanderbilt University. All rights reserved.

Contact Information

For copies of the TAMI or the TAMI Accessibility Rating Matrix, please visit the TAMI webpage at http://peabody.vanderbilt.edu/tami.xml. For further information about accessibility theory or the accessibility review process, please contact the senior author of the TAMI by mail, phone, or email:

Peter Beddow
Peabody Box #59
230 Appleton Place
Nashville, TN 37203-5721
(615) 403-8206
peter.beddow@vanderbilt.edu

© 2009 Vanderbilt University. All rights reserved.

References

Beddow, P. A., Elliott, S. N., & Kettler, R. J. (2009). *TAMI Accessibility Rating Matrix*. Nashville, TN: Vanderbilt University.

Beddow, P. A., Kettler, R. J., & Elliott, S. N. (2008). *Test Accessibility and Modification Inventory (TAMI)*. Nashville, TN: Vanderbilt University.

Educational Testing Service. (2009). ETS Guidelines for Fairness Review of Assessments. Retrieved on November 13, 2009 from http://www.ets.org/Media/About_ETS/pdf/overview.pdf

Elliott, S. N., Kettler, R. J., Beddow, P. A., Kurz, A., Compton, E., McGrath, D., Bruen, C., Hinton, K., Palmer, P., Rodriguez, M. C., Roach, A. T., & Bolt, D. (In press). Using modified items to test students with persistent academic difficulties. *Exceptional Children.*

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15(3)*, 309-344.

Johnstone, C., Thurlow, M., Moore, M., & Altman, J. (2006). *Using systematic item selection methods to improve universal design of assessments* (Policy Directions 18). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Kettler, R. J., Rodriguez, M. C., Bolt, D. M., Elliott, S. N., Beddow, P. A., & Kurz, A. Modified multiple-choice items for alternate assessments: Reliability, difficulty, and the interaction paradigm. Manuscript under review by *Applied Measurement in Education* for publication.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent develpments. *Educational Psychologst, 38,* 1-4.

Roach, A. T., Beddow, P. A., Kurz, A., Kettler, R. J., & Elliott, S. N. (In press). Using student responses and perceptions to inform item development for an alternate assessment based on modified achievement standards. *Exceptional Children.*

© 2009 Vanderbilt University. All rights reserved.

Suggested Requirements
for
Race to the Top Assessment Consortia

Expertise/Experience	Key Personnel		
	Expert 1	Expert 2	Expert 3
Designing large scale assessments			
Measuring opportunity to learn core content			
Measuring alignment among intended curriculum, enacted curriculum, and assessed curriculum at each grade level			
Writing highly accessible test items			
Assessment & measurement of children with disabilities			
Documenting accommodations & their implementation integrity			
Working with states to develop assessments			
Development of criterion referenced tests & instruments			
Psychometric evaluation			
Developing performance level descriptors			
Conducting standard settings			
Conducting studies of the technical adequacy of assessment instruments			
Research & publishing in area of assessment & psychometrics			

Validity Evidence Framework

Type of Validity Evidence	Sources of Validity Evidence	Brief Description of Validity Studies
Content	**Evidence Based on Test Content**	
Concurrent; Convergent & Discriminant; Predictive	**Evidence Based on Relations to Other Variables**	
Construct	**Evidence Based on Internal Structure**	
Consequential	**Evidence Based on Consequences of Testing**	
Content & Scoring Consistency	**Evidence Based on Response Process**	
Reliability Estimates		

Speaker #4

Received @ 702 AM 11·18·09

Common State Assessments
Including Students with Disabilities

November 18, 2009

Good morning, my name is Lori Hagen and I am one of 3800 members of Local 1420 , in Albuquerque, New Mexico led by Dr. Ellen Bernstein. As president of the Albuquerque Teacher's Federation, Ellen was instrumental in collaborating to support our school district in creating the Beginning Teacher Mentorship Program to assist first year teachers and the Peer Assistance and Review Program to assist experienced teachers in need of support. Our local collaborated to support the New Mexico State Education Department in creating the current 3 Tiered Licensure system for all teachers.

I am a National Board Certified Teacher as an Exceptional Needs Specialist. I work for the Albuquerque Public Schools district which currently has over 86,000 students with approximately 14% being students with disabilities. This is my 31st year teaching students with disabilities. I have worked in self-contained, inclusion, and departmentalized classrooms from elementary to high school. I have worked with students with learning disabilities, emotional and behavioral concerns, autism, speech articulation needs, communication disorders, dyslexia, twice exceptional, and other health impairments. I've supported my district in serving on the Standards Writing Committee representing students with disabilities and in the Peer Assistance and Review Program going into classrooms and assisting teachers with lesson plans, behavior management, data keeping and formative assessments, as well as, helping first year teachers prepare for conferences, manage paperwork, write IEPs, communicate with parents and provide professional collaboration. I am the local coordinator of the Albuquerque Teachers Federation, National Board Candidate Support Program encouraging teachers to strengthen their profession by looking reflectively at their practice to improve student learning. National Board Certification requires teachers to prepare a portfolio including videos and evidence that students of all levels and abilities are challenged and learning. I have been a national trainer of trainers in the American Federation of Teachers Education, Research and Dissemination Program in Managing Antisocial Behavior for over 14 years. I also serve as a team member of the National IDEA Cadre in supporting all stakeholders working with students with disabilities. Mostly, I have witnessed students whose self esteem has been shattered and hard work ignored through their participation in State Based Assessments which are not at their functioning level, nor does this testing take into consideration their multisensory mode of learning. Currently, assessment testing of students with disabilities does not provide educators, such as myself, worthwhile information to make informed educational decisions based on individual student needs.

Today, I stand before you, on behalf of the 1.4 million members of the American Federation of Teachers to share my thoughts and recommendations on how the US Department of Education and Secretary Duncan, given this unique opportunity to develop high quality assessments, may be aligned to a common set of K12 standards.

This is a daunting task and I am encouraged by the initial steps in making this happen, however, like many past initiatives, I remain concerned that in the absence of systemic change, appropriate funding, ongoing monitoring and appropriate adjustments at various levels, this initiative will fall to the way side as so many in the past have.

I strongly recommend that you take the following recommendations into consideration as you move forward:

- Convene a panel of assessment experts, educators, district and state assessment administrators to review existing research on how to best assess students with disabilities using technology that is flexible enough to accommodate a number of modifications and adaptations for students with disabilities.
- Conduct a review of teacher preparation programs and determine how to best prepare pre-service teachers coming into the field. This generation of teachers will have had much more exposure to recent advances in technology and will need to incorporate this into their preservice training.
- Consider the use of universal design principles – multiple means of presentation, expression, and engagement. This type of assessment design will meet the needs of a wider range of all students, including students with disabilities and English language learners.
- Invest in the development of computer-based assessments and the structures and systems necessary to support the use of such assessments. Educators will need on-going training, buildings will need updated hardware and software, and parents and students will need to be involved by understanding how the assessments will be used and how to prepare for them.

I appreciate this opportunity to speak on behalf of the New Mexico Federation of Teachers and our national organization, the American Federation of Teachers. If you need additional information, I may be reached at Lori1523@aol.com or hagen@aps.edu.

Signed,

Lori Hagen

Lori Hagen, NBCT
Special Educator and PAR Consultant Teacher
Albuquerque Public Schools
Albuquerque Teachers Federation

National Association of State Directors of Special Education, Inc.
1800 Diagonal Road, Suite 320, Alexandria, VA 22314
Tel: 703/519-3800 Fax: 703/519-3808 www.nasdse.org

THE NATIONAL ASSOCIATION OF STATE DIRECTORS OF SPECIAL EDUCATION (NASDSE)

TESTIMONY ON RACE TO THE TOP ASSESSMENT PROGRAM

NOVEMBER 18, 2009
ATLANTA, GEORGIA

I am Marcia Harding, President of the National Association of State Directors of Special Education (NASDSE), the national association that represents the state directors of special education in the states, the federal territories, the Bureau of Indian Education, the Department of Defense Education Agency and the Freely Associated States. I am also the Director of Special Education for the Arkansas Department of Education. Thank you for this opportunity to speak to you today about assessing students with disabilities.

Assessment of Students with Disabilities

As an organization, NASDSE called for the inclusion of students with disabilities in an inclusive accountability system as far back as the mid-1990's – long before it was fashionable or popular to do so. Our members have continued to support the inclusion of students with disabilities in the general education curriculum, recognizing that students with disabilities are general education students FIRST. We believe that all students, including students with disabilities should be held to high standards and that most students with disabilities can take and should be able to pass the regular assessments that are now required by the No Child Left Behind Act (NCLB).

However, we also recognize that for some students with disabilities, the regular assessment may be a mountain that is simply too high for them – even with all of the supports that a school can provide. For some of these students – those now referred to as the most significantly cognitively impaired – they will need to be assessed on an alternate assessment that is based on alternate achievement standards. That system was put into place by regulations after passage of the NCLB and appears to be well understood. Most states and school districts appear comfortable with the 1% limitation on those students who can be counted as proficient on this type of assessment recognizing that some flexibility may be appropriate in some school districts if they can demonstrate a high percentage of these students.

There is considerably more controversy over other students with disabilities. Some feel that <u>all</u> remaining students with disabilities should be able to take the regular assessment based on regular achievement standards – with or without accommodations. Students (and their parents) have expressed frustration that their particular accommodation has not been permitted for the NCLB assessment however, lending an additional concern to the assessment process. And others – including NASDSE – believe that some students with disabilities – even with all of the best instructional supports and services that we can offer – cannot be successful on the regular assessment based on regular achievement standards. NASDSE believes that these students need an alternate assessment based on modified achievement standards in order to demonstrate their capabilities.

It is equally important that assessments for students with disabilities be able to reflect their educational gains and growth. Because not all students with disabilities will be able to make – and demonstrate – a year's worth of progress on a regular assessment, it is critically important that assessments be accessible so that the student can be recognized for whatever growth s/he has made.

We know that we have the technology available that can provide students with access to both curriculum content and the most appropriate assessment. We can accomplish this through the use of materials and assessments that are developed using the constructs of Universal Design for Learning or UDL. I'll talk a little more about this later in my remarks. Technology advances are helping teachers implement the formative assessments needed to implement Response to Intervention in schools as well as summative assessments. We need to take advantage of the tools that are available. NASDSE urges you to consider proposals that take advantage of technology to provide access to both of these types of assessments that will ensure that students with disabilities will be able to demonstrate their knowledge and skills.

Given this brief overview, I would like to make the following points:

1. Special education – including disability advocates AND local educators – must be at the table when critical policy decisions are being made at the federal level regarding assessments.

2. It is important to remember that students with disabilities span a broad range of disabilities, even within a specific category of disability. For example, within the category of specific learning disability, the disability may range from mild to severe. Assumptions about the severity of a student's disability or what kind of assessment or accommodation a student needs cannot be made based on the 'category' of the disability.

3. When designing assessments, the federal government can use its bully pulpit and funding to encourage and support strategies that incorporate the concepts of universal design for learning so that all students can access the assessments. In giving out Race to the Top grants, we encourage you to consider UDL as one of the factors that you look for in considering grant applications.

4. We are concerned that if the Department intends to tie these assessment grants to the common core initiative currently underway that this initiative has, to date, made no accommodation for students with students. NASDSE believes that the common core initiative must accommodate extended standards for students who need them. Likewise, any assessments that are tied to the standards must reflect this need as well.

5. Finally, we note that some states have been working diligently to develop alternate assessments based on modified achievement standards based on the No Child Left Behind Act, yet the Department has found fault with virtually all of them and yet offered little or no guidance as to how to fix them. If the Department intends to continue this Peer Review construct in the reauthorization of the Elementary and Secondary Education Act, then we urge you to provide better guidance to states on what you are looking for in modified assessments, as well as technical assistance and competitive grants that would speed up the process for development of these assessments.

Again, I thank you for the opportunity to speak before you today and I'm happy to answer any questions that you might have.

Testimony Submitted to
U.S. Department of Education
Kim Hymes
Director, Policy & Advocacy
Council for Exceptional Children
November 18, 2009

Good afternoon, I am Kim Hymes here on behalf of the Council for Exceptional Children. This issue is of great importance to CEC's 40,000 members who are special education teachers, early intervention providers, administrators, researchers, and higher education faculty.

As you know, No Child Left Behind Act, has revolutionized how students with disabilities participate in our national accountability system. However, because our accountability system can only be as strong as the assessment on which it is based, CEC has advocated for revamping current assessments.

As the Department considers the development of the Race to the Top Assessment Competition, CEC urges the Department to focus on the following areas:

1. Creating assessments that are accessible to diverse learners;
2. Creating better Alternate Assessments based on Alternate Achievement Standards (AA-AAS) and Alternate Assessments based on Modified Achievement Standards (AA-MAS); and
3. Creating assessments that provide meaningful feedback to educators and families

Creating Assessments that are Accessible to Diverse Learners

CEC urges the Department to fund assessments and assessment systems that consider the needs of diverse learners from the first stages of creation and development. Current assessments were not created to address the diverse learning needs of students and, as a result, attempts have been made through the use of accommodations and other strategies to retrofit current assessments so they are more accessible. 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 a disability – from the beginning.

Specifically, CEC urges the Department to fund grants that create assessments which:

- Use multiple measures that are norm referenced on students with disabilities
- 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
- Include computer adaptive testing

Creating better Alternate Assessments based on Alternate Achievement Standards (AA-AAS) and Alternate Assessments based on Modified Achievement Standards (AA-MAS)

As you know, 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) for certain students with 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[1] within the Institute Of Education Sciences, found 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 determine a student's achievement.

Currently, the Department has only approved only Texas's AA-MAS. States clearly need assistance in this area. Students with disabilities have the right under law to access these exams and have their scores counted. But there are 49 states where this remains a right in name only. The RTTT Assessment Grants are a perfect opportunity for the Department to fund research and creation of assessment systems which will expand the number of students that can access these systems.

Therefore, CEC urges the Department to use this opportunity to fund alternate assessments based on both alternate and modified achievement standards as part of the RTTT Assessment Competition. Also, CEC believes it is critical for grants to include other elements that contribute to effective assessments and administering this type of assessment such as professional development, potential for scaling-up, dissemination practices, and additional research needed.

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

---

[1] 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.

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

In conclusion, CEC appreciates this opportunity to provide feedback as the Department moves forward in funding grants through the RTTT Assessment Competition and we will submit written comments that expand on these issues and discuss students with gifts and talents. 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.

Good morning and thank you for the opportunity to speak with you today. My name is Cara Cahalan Laitusis and I am a senior research scientist at Educational Testing Service in Princeton New Jersey.

Currently I am project director for two research projects funded by the U.S. Department of Education to investigate improving state reading assessments for students with learning disabilities and students who are blind or visually impaired. In addition I manage the research on improving measurement for individuals with disabilities that ETS funds through our Validity Research Initiative.

My colleagues at ETS and I have analyzed the test data for several state assessments and have conducted experimentally designed research studies to examine the impact of testing accommodations. Our 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 the state assessments that I have worked with the proportion of students without disabilities responding at chance level (or below) is less than 3 percent, but this percentage jumps to 10 to 20 percent for students with learning disabilities. This difficulty level is required to assess state standards but results in test scores that are less reliable. In addition, it 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. Currently, several states are now exploring adaptive testing models. 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 of the projects I am working on 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 non-trivial 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 math 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 are 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.

In conclusion I believe that adaptive testing, particularly multistage adaptive testing, holds promise for students with disabilities; however it is not a panacea. In my written comments I have included several recommendations for how the Race to the Top Common Assessment funds could be used to develop the infrastructure for delivering accessible assessments that target test questions to student achievement.

Thank you for your time today.

Recommendations:

- Specify that some portion of the Race to the Top Common Assessment Funds be devoted towards 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 U.S. Department of Education has already invested funds to develop accessible computer based testing platforms (NAEP Writing and NimbleTools) so I would encourage you to consult with your colleagues in NCSER, NCES, and 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 disability subtypes that are likely to exhibit diverse cognitive profiles.

- 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 you consider innovative test items and design features I encourage you to consider carefully 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 which 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 NCSER, OSEP, and NSF; such as the National Accessible Reading Assessment projects. The NARAP Accessibility Principles for Reading Assessments which includes supporting research evidence can be found at http://www.narap.info/publications/reports/NARAPprinciples.pdf

Speaker # 8

11·18·09 received @ 838 AM

Dear Madam and Sirs:

My name is Dr. Ted Mauro. I am the educational director of ED101 Inc of South Carolina. We are an educational consulting and advocacy group for disadvantaged youth focused on the issues of drop-out, at-risk and disability. I do need to state that I appear today on my own accord and not as a representative of ED101 or its clients.

I stand before you with over 15 years experience in the classrooms including in residential treatment, private and incarcerated settings. I was a national assessor of the portfolio system for National Boards and proctored numerous test and assessments in private practice. I have received a Masters in Special Education and a PH.D in educational administration and have tirelessly worked in some of the most challenging educational environments including at OSEP in Washington DC. I am presently on the staff of Drexel University in Philadelphia, PA.

I thank you for this opportunity to address your panel on the incredible challenges of creating a assessments system for Special education students, a diverse and exceptional group, while attempting to maintain the components of a objective, valid and relevant student progress measure. While the use of accommodations for high incident students with disabilities such as LD or physical challenges is well documented - the ultimate challenges of creating a system to evaluate the academic progress of those with low-incident disabilities (particularly those with various forms of Development Disorders that fall on the autism spectrum, intellectual challenges such as profound MR or emotional/behavioral disorders such as depression is challenging since it must attempt to not only engage the student to attempt the measure but then also solicit the desired performance on cue for a task that must be relevant to the needs of the child while taking into account the numerous challenges of being a special educator in today's ever changing school system. This is not to say that a person with Autism cannot take assessments with appropriate accommodations since many have and many will attend Major Colleges and Universities in the future which we all applaud. Flexibility is the most desired trait in today's special educator and one of the few constant in the classroom is that tomorrow will not be like today. The system of assessment must be flexible too.

I personally have administered 5 different Special Education assessments at the High School level and found some much more appropriate than others. Nothing was more frustrating than giving a paper test to a student who couldn't spell their name or a performance based measure on a task that wasn't taught or would be unnecessary for this student. Because of the variety of individual performances and the varied factors in each classroom I found the portfolio based assessment as the best devise not only for myself as the instructor but also as the best understood by parents, administrators, school boards and potential employers. I had learned this when I became an assessor of the National Boards and it opened my eyes to the large variety of outstanding programs across the country. It allowed me to further my expertise as I learned and emulated some of the best teachers and programs in the country. This system is also support by the National association of secondary school principals NASSP

I had used a portfolio-based system in South Carolina and saw some merit to it but received little feedback from it and found it confused in intent. While I would seek to create and maintain electronic portfolios, all portfolio assessments are flexible enough to adapt to different students needs. It also has the added benefit of providing stratified evidence of relevant classroom instruction, assist researchers who have latitudal and longitutal data of various real world measures. Portfolios' also adopts particularly well to mainstream and co-teaching environments, produces cultural evidential artifacts of not only a student but a program and can document Part C to B transitions or other agency or community involvement. Like the personnel file in every Fortune 500 company office a collection of data, evaluation reports, IEP, statistical data, teacher qualifications, restraint and seclusion reports, vocational instruction, parental involvement and other assessments and supporting documentations would be part of the portfolio. I see these portfolios be an integral part of merit based pay for Special Educators and along with the IEP itself and as an extension of the State Accountability System for SPP Indicators.

Components of these portfolios can be dictated by the age and classification of the student. As the needs of students change with maturity the basic academic subjects can be transitioned Indicator 13 & 14 as well as independent living skills. I would include a bar-coded tracking system that will link each portfolio so components can follow educational transitions, migrant students as well as assist post secondary outcomes. What could be more valuable to me as a

teacher from another district or working in Vocational Rehabilitation than to receive a collection of student produced work to go with a IEP.

This is not to say that portfolios have their challenges. They can be very difficult to store, hence out desire to create an electronic version. With today's IPhones video, electronic IEPs and massive download speeds it the ability to maintain these portfolios with direct student produced work is better than ever. The electronic components will also allow the redacting of personal information and encryption of data.

Portfolio based assessments are labor intensive – unlike a paper test it is a daily consideration however like the NCATE or National Boards system (both portfolio systems) it also causes one to question how ones instruction matters to this student and what type of impact I make to a student's life beyond school. An effective special educator makes assessment an integral seamless part of instruction. The ultimate challenge is for the assessor but, as mentioned, these issues have been addressed by other organizations and are well researched. The use of these portfolios to be training tools for teachers is a very important component and one that should not escape SEA's or teacher training programs. Assessments are only one part of effort to address shortfall and need to tied to differenced instruction or personnel retraining. The special education field is always changing and has made great strides however the professional development of some education communities has been slow and ineffective; the special education assessment component of the Race to the Top program must be a catalyst to reflective and innovative instructional programs.

The key to the overall effectiveness of this program is going to come down the effectiveness of the portfolio design, an appropriate rubric, the training of educator/administrator/assessor and the guidelines that are shared with the SEA and LEA. I would also stress that Federal Guidelines need to be strict, that a national portfolio model and a universal national test system is created so there is consistency across the country, that LEA numbers match the numbers submitted to SEA and SPP are the same numbers of alternative portfolio assessments allowed. Similar to other educational records alternative assessments need to be available for parental inspection.

By creating a National Alternative Assessment that is portfolio based you allow you to adapt a system that will allow the flexibility you desire for the wide scope of programs and settings that exists while maintaining the accountability standards of assessments for every students, educator and SEA across the country. The student performance itself is important however, not only is individualized performances measures by this method but the performance of school programming in addressing behavior, parental involvement, disproportionality, transition program information is important for everyone.

I thank you for your time today and I invite any questions that the panel may have for me at this time or in the future.

Ted David Mauro

Theodore Darid Mauro PhD
Educational and Training Director
www.ED101.org
ted@ed101.org
tdm53@drexel.edu
864.982.2381

11.18.09 Speaker #5 Received @ 8:05 AM

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

Paula E. Woodard
National Education Association
NEA IDEA Resource Cadre

Assessments and Students with Disabilities

My name is Paula Woodard, and I am a teacher with the Marietta City Schools System. I am also a member of the National Education Association (NEA) and a member of the NEA IDEA Resource Cadre. I am here representing NEA, which is the nation's largest professional employee organization. NEA's 3.2 million members work at every level of education, from pre-school to K-12 public schools to university graduate programs, which includes general and special education teachers, administrators, college educators, specialized instructional personnel and education support professionals in every state, as well as in more than 14,000 communities across the United States.

NEA strongly supports the inclusion of students with disabilities in state and local assessments. However, many students with disabilities are struggling to pass their states' high-stakes assessments. If the students don't pass the assessments, they may not receive a high school diploma. For example, just recently, I saw a former student at one of the local malls. He graduated from high school two years ago with the correct number of credits and a special education diploma; however, he is still retaking one section of the Georgia High School Graduation Test (GHSGT) because he didn't make the minimum score on all sections. In fact, in a few days, he was getting ready to retake the math section once again because he is unable to obtain the minimum score. At this point, this student has probably taken that section at least 10

times. Obviously, if there had been an alternative valid assessment, his situation might be drastically different. Therefore, it's critical that assessments for students with disabilities be valid, reliable, and meaningful.

I attended this student's exit IEP meeting, and he stated that he wanted to attend a two-year technical college. Since he doesn't have a high school diploma, he cannot attend a two-year technical school because a special education diploma won't allow him entry into any post-secondary institution. He's currently in a holding pattern until he passes the last section of the GHSGT.

Unfortunately, this student's story is not unusual. Due in large part to the fact that accessible, meaningful assessments are not readily available to all students with disabilities, these students are currently subjected to high-stakes assessments that limit their future options leaving them frustrated and demoralized.

Obviously, this situation must change. To address these concerns, NEA recommends the following:

1) Assessments for students with disabilities should focus on evaluating their growth through student progress monitoring and other student growth measures.
2) Accessible assessments should be developed for all students using the principles of universal design for learning (UDL). [For information on UDL, see the UDL guidelines published by the National Center on UDL at http://www.udlcenter.org/aboutudl/udlguidelines.]
3) Students should be provided with the widest range of assessment accommodations feasible. To do this, test designers must run validation studies using a broad range of

accommodations, including the read aloud accommodation and Braille. [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 and are available at http://www.narap.info/publications/reports/NARAPprinciples.pdf.]

4) Educators need to be provided with professional development on the use of accommodations in instruction and assessments. 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.
5) 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 need.
6) Computer-adaptive assessments can provide in-depth measurement of knowledge and skills of students who have difficulty showing their abilities on typical paper-pencil tests. We encourage the U.S. Department of Education to fund research on creating valid and reliable computer-adaptive assessments.

Thank you for allowing me this opportunity to share with you today. As an NEA member and educator, I want ALL students to have the opportunity to participate in assessments, which are not only valid, but reliable and meaningful as well.

Comp. by: JNagalakshmi Date:5/3/07 Time:09:45:54 Stage:First Proof
File Path://spiina1001z/womat/production/PRODENV/0000000005/0000001817/
0000000016/0000556949.3D Proof by: QC by:

JAMAL ABEDI

UTILIZING ACCOMMODATIONS IN ASSESSMENT

Au1

An equitable assessment and accountability system requires that all students be included in large-scale national and local assessments. However, there is a substantial performance gap between those for whom the assessment language is a second language and those students who are native speakers of the assessment language, particularly on academic subjects that are high in language demand (Abedi, 2006a). The literature suggests that this performance gap is explained by many different factors including parent education level and support, SES, the challenge of second language acquisition (Hakuta, Butler, and Witt, 2000; Moore and Redd, 2002) and a host of inequitable schooling conditions (Gándara, Rumberger, Maxwell-Jolly, and Callahan, 2003). Yet, it is also often the case that the measurement tools are ill-equipped to assess the skills and abilities of second language learners. To offset these challenges, nonnative speakers of the assessment language are provided with "*test accommodations.*"

Test accommodations refer to changes in the test process, in the test itself, or in the test response format. The goal of accommodations is to provide a fair opportunity for nonnative speakers of the assessment language and students with disabilities to demonstrate what they know and can do, to level the playing field, so to speak, without giving them an advantage over students who do not receive the accommodation.

The issues concerning accommodations are important in all countries where there are students who do not have high proficiency in the language of instruction and assessment in schools; usually these are immigrants and indigenous groups. Since the USA has conducted more research on accommodations than many other countries, in this chapter we present an overview of major research findings that are reported in the American research journals for English language learners.

EARLY DEVELOPMENTS

Historically, the concept of *accommodations* was first introduced in the field of special education. Many students with disabilities need specific forms of assistance in the classroom setting to deal with their disabilities, i.e., to level the playing field. For example, deaf and hard

E. Shohamy and N. H. Hornberger (eds), Encyclopedia of Language and Education, 2nd Edition, Volume 7: Language Testing and Assessment, 1–17.
©2008 Springer Science+Business Media LLC.

Comp. by: JNagalakshmi Date:5/3/07 Time:09:45:54 Stage:First Proof
File Path://spiina1001z/womat/production/PRODENV/0000000005/0000001817/
0000000016/0000556949.3D Proof by: QC by:

of hearing students need hearing aids to offset the effect of their inability to hear at the same level as regular students. Similarly, blind or visually impaired students need to use the brail version of a test to be able to read the test items. These accommodations are used to increase equity in the classroom as well as during assessment conditions. The concept of accommodations was then extended to English language learners (ELL). Unfortunately, however, not only the concept but the actual accommodation strategies that were created and used for students with disabilities were used for ELL students, many of which may not be relevant for these students.

By definition, accommodations are used for students with disabilities (SD) to assist them with their disabilities. For ELL and nonnative speakers of the assessment language the goal of accommodations is to help with second language needs. Another goal is to reduce the performance gap between SD/ELL and non-SD/non-ELL students, without jeopardizing the validity of assessments. In the USA, there are many forms of accommodations which are used for both ELL students and students with disabilities in different states (Abedi, Kim-Boscardin, and Larson, 2000; Rivera, Stansfield, Scialdone, and Sharkey, 2000; Thurlow and Bolt, 2001). Yet, as will be shown below, there is little evidence to support the effectiveness and validity of assessments using these accommodations. Au2

There are many policy-related issues concerning accommodations that need to be carefully reviewed. For example, one must examine the validity of criteria used for decisions on the provision of accommodations and the appropriateness of the type of accommodations used for particular groups of students. Different criteria are used for the selection of accommodations and the lack of a national or local master plan for accommodations creates controversies over the use of accommodations.

Most accommodations that are currently offered to ELL students and to a lesser extent to SDs, have a limited empirical research base and states in the USA often use a "common sense" approach in choosing accommodations. With regard to the validity of accommodations, research results are available only for a few forms of accommodations, while the validity for others has yet to be examined. For some accommodations there may not be enough research to judge the effectiveness and validity of the accommodations for a particular group of students.

MAJOR CONTRIBUTIONS

As noted above, the main focus of this chapter is on accommodations for ELL students in the USA. However, a short discussion on accommodations for students with disabilities is included as well, due to some

Comp. by: JNagalakshmi Date:5/3/07 Time:09:45:54 Stage:First Proof
File Path://spiina1001z/womat/production/PRODENV/0000000005/0000001817/
0000000016/0000556949.3D Proof by: QC by:

historical connections between the accommodation policies and practices for these two subgroups of students. In fact, some accommodations that are currently used for ELL students were initially proposed and used for students with disabilities (see, for example, Rivera, Stansfield, Scialdone, and Sharkey, 2000).

Review of literature on accommodations suggests that: (1) existing research on some forms of accommodations is not conclusive, and (2) for many forms of accommodations used by different states there is very limited empirical data to support their validity. It should be noted that the term "validity of accommodations" is used here within the general framework of assessment; therefore, validity of accommodations refers to the "validity of accommodated assessments." In other words, an accommodation strategy may not be valid or invalid unless it is considered within the assessment framework. In presenting the research summary it will be shown that: (1) some accommodations that are used for ELL students are designed for students with disabilities and are not relevant to ELL students, and (2) in some cases, findings from different studies about accommodations are not consistent. Below is a list along with a summary of research for some commonly used accommodations.

Braille is used for students with blindness or significant visual impairments. Braille versions of a test may be more difficult for some items than other items such as items with diagrams and/or special symbols (Bennett, Rock, and Kaplan, 1987; Bennett, Rock, and Novatkoski, 1989; Coleman, 1990). This is clearly an accommodation for SD (blind) students only.

Computerized Assessment is used for students with physical impairments that have difficulty in responding to items in a paper-and-pencil format. Some studies suggest that this accommodation increases the performance of students (Russell, 1999; Russell and Haney, 1997; Russell and Plati, 2001). Other studies have not found computerized assessments to be effective (MacArthur and Graham, 1987), or not as effective as traditional assessments (Hollenbeck, Tindal, Stieber, and Harniss, 1999; Varnhagan and Gerber, 1984; Watkins and Kush, 1988). In a study with grade 4 and 8 students in mathematics, Abedi, Courtney, and Leon (see, Abedi, Hofstetter and Lord, 2004) found that computerized assessments can be highly effective in making tests more accessible to ELL students. The study did not find any validity issues with the computerized assessment suggesting that the computerized assessment was not different than the traditional assessment for non-ELL students.

Dictate Response to a Scribe (someone writes down what a student dictates with an assistive communication device). This accommodation has been shown to have an impact on the performance of students with learning disabilities (Fuchs, Fuchs, Eaton, Hamlett, and Kams, 2000; MacArthur and Graham, 1987). Tippets and Michaels (1997) found

Comp. by: JNagalakshmi Date:5/3/07 Time:09:45:54 Stage:First Proof
File Path://spiina1001z/womat/production/PRODENV/0000000005/0000001817/
0000000016/0000556949.3D Proof by: QC by:

this accommodation, in combination with other accommodations, such as read aloud and extended time helps students with disabilities. However, there are concerns over the validity of this accommodation. Koretz (1997) found this accommodation helped students with learning disabilities; however, Thurlow and Bolt (2001) recommended that if students are unable to handwrite but can efficiently use a computer, the use of a computer should be considered.

Extended Time. This is one of the most commonly used accommodations. Under this accommodation, students receive extra time (usually 50% more time) to respond to the test items. It is used for both English language learners and students with different types of disabilities. Thurlow, House, Boys, Scott, and Ysseldyke (2000) suggested that disagreement between states may be a concern regarding the validity of extended time accommodation. Chiu and Pearson (1999) found extended time to be an effective accommodation for students with disabilities, particularly for learning disabilities. Some studies found extended time to help students with disabilities in Mathematics (Chiu and Pearson, 1999; Gallina, 1989). However, other studies did not show an effect of extended time on students with disabilities (Fuchs, Fuchs, Eaton, Hamlett, and Karns, 2000; Marquart, 2000; Munger and Loyd, 1991). Studies on the effect of extended time in language arts did not find this accommodation to be effective (Fuchs, Fuchs, Eaton, Hamlett, and Karns, 2000; Munger and Loyd, 1991). Some research studies showed that extended time affects the performance of both SD and non-SD students, and therefore makes the validity of this accommodation suspect. For ELL students, research on extended time has produced mixed results. Abedi, Hofstetter, and Lord (2004) found no effect of extended time for ELL students. On the other hand, Hafner (2000) found extended time to be an effective accommodation for ELL students.

It must be noted at this point that many school districts in the USA allow unlimited time in taking both Title I and Title III (Rivera and Collum, 2006) assessments under the No Child Left Behind (NCLB, 2001) accountability requirements. That is, the state tests are often considered as *power* tests and not as *speed* tests. Therefore, extended time is not viewed as an accommodation and consequently there is no concern over the validity of assessments using extended time since everyone receives extra time in testing.

Interpreter for Instructions. In this accommodation an interpreter translates test instructions in sign language. This accommodation is recommended for students with hearing impairments. Adaptations in the directions may help deaf children score the same as other students (Sullivan, 1982).

Large Print is used for students with visual impairments. Research has indicated that this accommodation has helped reduce the performance gap between students with visual impairments and students without

Comp. by: JNagalakshmi Date:5/3/07 Time:09:45:54 Stage:First Proof
File Path://spiina1001z/womat/production/PRODENV/0000000005/0000001817/
0000000016/0000556949.3D Proof by: QC by:

disabilities (see for example, Bennett, Rock, and Jirele, 1987). The results of a study by Bennett, Rock, and Kaplan (1987) revealed that using this accommodation for visually impaired students does not affect the construct under measurement. Other studies suggest that extra time may be needed with this accommodation (Wright and Wendler, 1994). Large print has also been used for students with learning disabilities. Several studies have shown no impact of this accommodation for students with learning disabilities. One study, however, showed that large print helps students with learning disabilities (Perez, 1980). This accommodation has also been used for ELL students (Rivera, 2003; Sireci, Li, and Scarpati, 2003) although it is not clear how relevant this accommodation is to ELL students.

Mark Answer in Test Booklet in which students write their answers directly on the test booklet rather than on an answer sheet. This accommodation can be used for students who have a mobility coordination problem. Some studies on the effectiveness of this accommodation did not find significant difference between those students tested under this accommodation and those using separate answer sheets (Rogers, 1983; Tindal, Heath, Hollenbeck, Almond, and Harniss, 1998). However, other studies found lower performance for students using this accommodation (Mick, 1989). In fact, many school districts in the USA have used this accommodation for ELL students (Rivera, 2003), yet there is no evidence on the relevance or effectiveness of this accommodation for ELL students.

Read Aloud Test Items are used by students with learning disabilities and students with physical or visual impairments. While some studies found this accommodation to be valid in mathematics assessments (Tindal, Heath, Hollenbeck, Almond, and Harniss, 1998), others have concerns over the use of this accommodation on reading and listening comprehension tests (see for example, Burns, 1998; Phillips, 1994) since this accommodation may impact the validity of assessment by altering the construct (see also Bielinski, Thurlow, Ysseldyke, Freidebach, and Freidebach, 2001; Meloy, Deville, and Frisbie, 2000). Read aloud as an accommodation has also been used for ELL students in the USA (Rivera, 2003), again, without any indication of the relevance or effectiveness of this accommodation for this group of students.

Read or Simplify Test Directions is appropriate for students with reading/learning disabilities. A study by Elliot, Kratochwill, and McKevitt (2001) suggested that this accommodation affects performance of both students with disabilities and students without disabilities. There are therefore concerns over the validity of this accommodation especially since it has also been used frequently for ELL students; the use of this accommodation is of particular concern in reading assessment.

Test Breaks where students receive multiple breaks during the testing session can help students with different forms of disabilities. A study by DiCerbo, Stanley, Roberts, and Blanchard (2001) found

Comp. by: JNagalakshmi Date:5/3/07 Time:09:45:55 Stage:First Proof
File Path://spiina1001z/womat/production/PRODENV/0000000005/0000001817/
0000000016/0000556949.3D Proof by: QC by:

that students tested under the *multiple-breaks* administrations obtained significantly higher scores than those tested under standard testing conditions with no additional breaks. The study also showed that middle and low-ability readers benefited more from this accommodation than high-ability readers. However, another study (Walz, Albus, Thompson, and Thurlow, 2000) found that students with disabilities did not benefit from a multiple-breaks test administration while students without disabilities did. These results show quite the opposite of what is expected of valid accommodations. Sometimes *test breaks* as a form of accommodation has been recommended for ELL students (Rivera, 2003) as it may help some ELL students but may not be relevant for other ELLs since it does not address their English language needs.

Providing an English Dictionary and extra time (Abedi, Hofstetter, and Lord, 2004; Hafner, 2000; Thurlow, 2001) was found to affect performance of all students (see also, Maihoff, 2002; Thurlow and Liu, 2001). This suggests that the results of accommodated and nonaccommodated assessment may not be aggregated.

Translation of Assessment Tools into Students' Native Language may not produce desirable results and may even provide invalid assessment results if the language of instruction and assessment is not aligned (Abedi, Hofstetter, and Lord, 2004.)

As noted earlier, in spite of the concerns expressed by researchers over the validity, effectiveness, and feasibility of some forms of accommodations, these accommodations are used frequently by states and districts across the USA. That is, decisions on the type of accommodations for English language learners and students with disabilities do not seem to have been influenced much by the research findings.

Accommodation Issues for English Language Learners: Accommodations are meant to "level the playing field" for ELL students by accommodating their potential language limitations in an assessment. Unfortunately, there are major equity issues with many of the accommodations used for ELL students. The practice of using accommodations for ELL students that are initially developed for students with disabilities (Rivera, Stansfield, Scialdone, and Sharkey, 2000) is extremely problematic as some accommodations that are used for students with disabilities are not relevant for ELL students. For example, using large print may be an effective accommodation for some students with visual impairments while ELL students need specific accommodations to address their linguistic needs. As discussed above, there are major issues concerning accommodations for both ELLs and students with disabilities. While these issues deserve equal attention for both SD and ELL students, the focus in the next section will be on accommodation issues for ELL students.

Comp. by: JNagalakshmi Date:5/3/07 Time:09:45:55 Stage:First Proof
File Path://spiina1001z/womat/production/PRODENV/0000000005/0000001817/
0000000016/0000556949.3D Proof by: QC by:

WORK IN PROGRESS

As the number and percentage of English language learners increase in the USA, assessment equity and validity are becoming priorities for educational policymakers. Between 1990 and 1997, the number of US residents born outside the country increased by 30%, from 19.8 million to 25.8 million (Hakuta and Beatty, 2000). According to the National Clearinghouse for English Language Acquisition, over 4.5 million Limited English Proficient (LEP) students were enrolled in US public schools in 2000–2001, representing nearly 10% of the nation's total public school enrollment for prekindergarten through Grade 12 (Kindler, 2002).

To reduce the impact of language factors on the assessment outcome of ELL students, assessment in students' native language has been proposed as an accommodation. While this seems to be an attractive idea and many districts and states in the USA use this approach, research results do not support its fairness (Abedi, Hofstetter, and Lord, 2004). One major issue here is the possibility of lack of alignment between the language of instruction and language of assessment. If the language of assessment is not the same as the language of instruction, then the assessment outcome may be even less valid, again raising fairness as a serious issue. For example, when a native Spanish speaker learns content-area terminology in English, but is tested in Spanish, the outcome of the assessment may not be valid due to the student's lack of content terminology knowledge in Spanish. A student may be a fluent speaker of a language but not necessarily proficient in the academic language of his or her native language.

Some educational researchers and policy makers suggest that rather than testing students in their native languages (L1), they should be assessed by providing them with language accommodations such as a customized dictionary or a linguistically modified version of the test to help them with their English language needs. This seems to be a reasonable approach if the focus is on learning English as quickly as possible. However, others argue that students' knowledge of their first language could benefit their academic progress and testing them in English may not properly utilize their knowledge of L1.

PROBLEMS AND DIFFICULTIES

The purpose of testing accommodations is to assist students with certain limitations that they might have and provide them with a fair assessment. It is therefore important to examine the appropriateness, effectiveness, validity, and feasibility of accommodations for the targeted student populations.

Comp. by: JNagalakshmi Date:5/3/07 Time:09:45:55 Stage:First Proof
File Path://spiina1001z/womat/production/PRODENV/0000000005/0000001817/
0000000016/0000556949.3D Proof by: QC by:

Appropriateness. How appropriate are accommodations that are provided for ELL students? Since the common characteristic that distinguishes ELLs from non-ELL students is their possible limitation in English proficiency, it is reasonable to expect that accommodations that help ELL students with their language barrier would be the most relevant. However, in many places, the current practice of accommodations for ELL students is to simply use accommodations that are easily available or those that decision makers find relevant. These accommodations may not always be appropriate for these students. For example, Rivera (2003) presented a list of 73 accommodations that are used nationwide for ELL students. Our analyses of these accommodations (Abedi, 2006b) revealed that of these 73 accommodations, only 11 (15%) of them were highly relevant for ELL students in providing assistance with students' language needs. The list included accommodations such as:

- Subtests flexibly scheduled
- Tests administered at a time of day most beneficial to test-taker
- Tests administered in small groups
- Tests administered in a familiar room
- Colored stickers or highlighters for visual cues provided
- Copying assistance provided between drafts
- Test-taker types or uses a machine to respond (e.g., typewriter/word processor/computer)
- Test-taker indicates answers by pointing or other method
- Test-taker verifies understanding of directions

Since none of these accommodations address ELL students' language needs, they may not be adequate or appropriate for these students.

The National Assessment of Educational Progress (NAEP) also uses some accommodations that, at face value, are not very relevant to ELL students' language needs. For example, among the accommodations NAEP used for ELL students in the 1998 civics assessment were large print, extended time, reading questions aloud, small group testing, one-on-one testing, and scribe or computer testing (see Abedi and Hejri, 2004). While some of these accommodations may be helpful for students with disabilities, they may not be effective for ELL students.

Studies have found that the provision of accommodations in NAEP increased the *inclusion* rate of these students (Mazzeo, Carlsom, Voelkl, and Lutkus, 2000). However, research has shown that accommodations did *not* increase ELL student scores on the NAEP; that is, providing accommodations did not reduce the performance gap between ELL and non-ELL students. For example, no statistically significant differences were found between the performance of accommodated and nonaccommodated ELL students in the 1998 NAEP main assessments in reading, writing, and civics for students in fourth and eighth grades (Abedi and Hejri, 2004). Among the most likely

Comp. by: JNagalakshmi Date:5/3/07 Time:09:45:55 Stage:First Proof
File Path://spiina1001z/womat/production/PRODENV/0000000005/0000001817/
0000000016/0000556949.3D Proof by: QC by:

explanations for this is the lack of *relevant* accommodations. As indicated earlier, if the accommodations provided to ELL students have no relevance to their needs (mainly English language proficiency), then one would not expect any positive impact of accommodations on the outcome of assessments. Examples of relevant accommodations for ELLs and nonnative speakers of the assessment language include providing a glossary of noncontent terminology or modifying complex linguistic features as these accommodations directly address ELL students' language needs.

Another major issue in the provision of accommodations in NAEP was the very small number of ELL students who were accommodated. In the main NAEP assessments, the number of ELL students who were included in the study comprised between 7 and 8% of the sampled students, but only a fraction of these students, who had been accommodated by their schools in earlier assessments, received NAEP accommodations. For example, in the main assessment of the 1998 Grade 4 reading test, 934 ELL students were included, but only 41 (4%) of the included ELL students were provided with accommodations. In the Grade 8 sample, 896 ELL students were included, but only 31 (3.5%) were accommodated. Similarly, in the 1998 main assessment in civics, 332 ELL students in Grade 4 were included and only 24 (7%) were accommodated. In the same assessment, 493 ELL students were included in Grade 8 but only 31 (6%) were accommodated (Abedi and Hejri, 2004).

Validity. Invalid accommodations affect the outcome of assessments for individual students as well as for the group in which students belong. If accommodations affect the construct, then the accommodated and nonaccommodated assessments cannot be aggregated. Studies have found that some forms of accommodations may alter the construct being measured (see, for example, Abedi, Hofstetter, and Lord, 2004). For example, providing a published dictionary may affect the measurement of the construct, since it may provide content-related information which students can use to answer the questions. Abedi, Hofstetter, and Lord (2004) found that providing a glossary plus extra time increased performance of non-ELL students for whom the accommodation was not intended, thereby increasing the performance gap between ELL and non-ELL students. Thus, the validity of many commonly used accommodations is questionable. Unfortunately, research on the validity of accommodations is very limited and the validity of only a handful of accommodation strategies used for ELL students have been experimentally examined (Abedi, Hofstetter, and Lord, 2004; Francis, Lesaux, Kieffer, and Rivera, 2006; Sireci, Li, and Scarpati, 2003).

Feasibility. Feasibility of implementation is another issue with accommodation strategies. Accommodation strategies should not be a

Comp. by: JNagalakshmi Date:5/3/07 Time:09:45:55 Stage:First Proof
File Path://spiina1001z/womat/production/PRODENV/0000000005/0000001817/
0000000016/0000556949.3D Proof by: QC by:

logistical or financial burden to implement. For example, computer testing could be a burden if a school lacks funding for adequate computer resources. One-on-one testing may also be logistically challenging in large-scale assessments. Providing a dictionary as a form of accommodation is another example of a feasibility concern. To control for extraneous factors (e.g., differences in the type of dictionary and experience working with the dictionary), it is important that the same dictionary be provided to every student, which obviously creates major logistical issues.

FUTURE DIRECTIONS

Research-supported accommodations. The main goal of an accommodation is to make assessments more accessible across subgroups of students who otherwise could be affected unfairly by many nuisance variables that would make the assessment unfair and invalid. The discussion above casts doubt over the ability of many of the current accommodation practices to reach this important goal. There is no firm evidence to suggest that the accommodations used widely by school districts are effective, feasible, and valid. However, results of recent studies introduce some accommodation strategies for ELL students that, in addition to being valid, are also effective in reducing the performance gap between ELL and non-ELL students in content area assessments.

One major assessment issue is that a student's level of proficiency in the language of assessment may severely impact the validity of the assessment results. Students may have the content knowledge (e.g., in math and science) in their native language but may not be fluent enough in the language of assessment to express their knowledge on a test. To reduce the impact of language factors on the assessment outcomes of students, the linguistic modification of test items has been proposed in the literature (see, for example, Abedi, Lord, and Plummer, 1997). A linguistic-modification approach helps test developers reduce the level of unnecessary linguistic complexity in test items by controlling for sources of linguistic complexity (for a detailed description of linguistic modification approach, see Abedi, 2006a).

Earlier in this chapter research-based evidence about accommodations was presented. This evidence raises concerns about the validity of the accommodations used in schools for ELL students. The main question for the future is whether there are accommodations that would be beneficial to ELL students but do not affect the construct under measurement. Below is a short survey of accommodations that studies have shown to be effective and valid.

Recent studies at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) have examined several

Comp. by: JNagalakshmi Date:5/3/07 Time:09:45:55 Stage:First Proof
File Path://spiina1001z/womat/production/PRODENV/0000000005/0000001817/
0000000016/0000556949.3D Proof by: QC by:

different forms of accommodation. Abedi, Hofstetter, and Lord (2004) and Maihoff (2002) examined the linguistic modification approach and found it to be an effective and valid accommodation in the assessment of ELL students. Rivera and Stansfield (2001) found this accommodation to have no impact on the non-ELL student group suggesting that the accommodation is valid for ELL students. With this approach, simpler versions of items with language that might be difficult for students were drafted; the task remained the same but noncontent vocabulary and unnecessary linguistic complexity were modified (see Abedi, 2006a, for further discussion of the nature of and rationale for the linguistic modifications). These studies compared student scores on NAEP test items with comparable modified items in which the mathematics tasks and mathematics terminology were retained but the language and/or linguistic structures were modified.

Following are a few examples of studies on the effectiveness and validity of the linguistic modification approach as a form of accommodation for ELL students. Abedi and Lord (2001) examined the effects of this accommodation with 1,031 eighth grade students in southern California. Test booklets with either original English versions or modified English versions of the items were randomly assigned to the students. The results showed significant improvements in the scores of students in low- and average-level mathematics classes who received the booklets with linguistic modifications. Among the linguistic features that appeared to contribute to the differences were low-frequency vocabulary and passive voice verb constructions. English language learners and low-performing students benefited the most from the linguistic modification of test items.

In another study, Abedi, Hofstetter, and Lord (2004) examined the impact of linguistic modification on the mathematics performance of English learners and non-English learners. Using items from the 1996 NAEP Grade 8 Bilingual Mathematics booklet, three different test booklets (Original English, Modified English, and Original Spanish) were randomly distributed to a sample of 1,394 eighth grade students in schools with high enrollments of Spanish speakers. Results showed that language modification of items contributed to improved performance on 49% of the items. The students generally scored higher on shorter problem statements.

A third study (Abedi, Hofstetter, and Lord, 2004) examined the impact of four different forms of accommodation on a sample of 946 eighth grade students tested in math. The accommodations were (1) Modified English, (2) Extra Time only, (3) Glossary only, and (4) Extra Time plus Glossary. These four accommodation types, along with a standard test condition, were randomly assigned to the sampled students. Findings suggested that some accommodations increased

Comp. by: JNagalakshmi Date:5/3/07 Time:09:45:55 Stage:First Proof
File Path://spiina1001z/womat/production/PRODENV/0000000005/0000001817/
0000000016/0000556949.3D Proof by: QC by:

performance of both English learners and non-English learners, compromising the validity of the assessment. Among the different options, only the Modified English accommodation narrowed the score gap between English language learners and other students.

Other studies have also employed the language modification approach Kiplinger, Haug, and Abedi (2000) found linguistic modification of math items helpful in improving the math performance of ELL students. Maihoff (2002) found linguistic modification of content-based test items to be a valid and effective accommodation for ELL students in math. Rivera and Stansfield (2001) compared English language learner performance on regular and modified fourth and sixth grade science items. Although the small sample size did not show significant differences in scores, the study demonstrated that linguistic modification did not affect the scores of English-proficient students, indicating that linguistic modification is not a threat to score comparability.

While the current prevalent trends in accommodation practices are not supported by research (Solano-Flores and Trumbull, 2003), there is growing evidence that states are paying more attention to research findings on the effectiveness and validity of accommodations. The increasing use of research-supported accommodations for ELL students (such as linguistic modification of items) is encouraging. This trend may result in fairer assessments for ELL students.

English language learners and students with disabilities are faced with many challenges in their academic career and need special attention. For ELL students, the challenge of learning English and at the same time competing with their native English speaking peers in learning academic concepts in English is enormous. Similarly, for students with disabilities, it is quite challenging to learn at the same rate as their nondisabled peers given their disabilities. Even more serious is the case of ELL students with disabilities. These students are faced with dual challenges—learning a new language and dealing with their disabilities. Such inequity in educational opportunity creates a substantial performance gap between these students and their peers. While accommodations are provided to offset these challenges it has been shown that these accommodations are often not relevant or helpful and have limited supported research. It is especially important that accommodations for ELL students must be language-related in order to be effective in making assessments more accessible for these students.

One of the major issues for the future is the need to expand the research in the area of accommodations as there is not enough research to judge the effectiveness and validity of many of the existing accommodations for both SD and ELL students. For example, score comparability is highly related to the outcome of accommodated assessment. If provision of accommodation alters the construct being measured,

Comp. by: JNagalakshmi Date:5/3/07 Time:09:45:56 Stage:First Proof
File Path://spiina1001z/womat/production/PRODENV/0000000005/0000001817/
0000000016/0000556949.3D Proof by: QC by:

then accommodated assessment outcomes may not be valid and as a result the accommodated and nonaccommodated assessment outcomes cannot be aggregated. Additional research is needed to help schools choose the best accommodations and to ensure that the outcome of accommodated and nonaccommodated assessments can be aggregated. Recent publications reporting results of research on accommodations for ELL and SD students, including the taxonomy of accommodations provided in Rivera and Collum (2006), could help schools make better choices in selecting existing accommodations rather than using a common sense approach in their decisions.

It must be noted at this point, however that some accommodations may have a limited impact on assessment outcomes and may only be considered a quick fix because they may not be able to systematically address the underlying issue of equitably assessing immigrants and ELL students in providing an assessment in the appropriate language. Other accommodations, however, may help make assessments more accessible—and consequently more valid and fair—for immigrants and ELL students. For example, Levin, Shohamy, Spolsky (2003) found that if bilingual students were able to take a test in both of their languages, their performance improves because they construct meaning in two languages rather than one. Findings of a study by Levin and Shohamy (2006) indicated that "immigrants, rather than being deficient [in terms of their language resources] have a clear advantage that should be included in an expanded view of the construct of academic language" (p. 19). Obviously, the native language assessment is effective under the condition that the language of instruction and the language of assessment are aligned.

It is also important to understand how instruction and assessment interact. Students can benefit more when accommodations are provided under both assessment and instruction conditions. This combination provides an opportunity for bilingual and ELL students to become familiar with the accommodations that are used in their assessments.

The concept of academic language is an extremely important consideration when dealing with the assessment of immigrants and English language learners in content-based areas such as math and science. While everyone, particularly immigrants and ELL students, can greatly benefit from assessments with clear language, these students must also be familiar with the language that facilitates content learning, i.e., academic language. For example, as Levin and Shohamy (2006) pointed out, content literacy rather than language per se greatly impact students' performance in content-based areas. For example, Levin and Shohamy indicated that, "not only the vocabulary and symbols but also the norms, values, and conventions that are characteristics of the discipline" (p. 18).

Comp. by: JNagalakshmi Date:5/3/07 Time:09:45:56 Stage:First Proof
File Path://spiina1001z/womat/production/PRODENV/0000000005/0000001817/
0000000016/0000556949.3D Proof by: QC by:

Finally, the differential item functioning (DIF) approach may help identify specific items that discriminate against students who are not proficient in the language of assessment. The effectiveness of accommodations can then be examined on test items that exhibited a high level of DIF (C-DIF) (see, for example, Uiterwijk and Vallen, 2003).

REFERENCES

Abedi, J.: 2006a, 'Language issues in item-development', in S.M. Downing and T.M. Haladyna (eds.), *Handbook of Test Development*, Lawrence Erlbaum Associates, Mahwah, NJ.

Abedi, J.: 2006b, '*Are accommodations used for ELL students valid*?', paper presented at the 2006 Annual Meeting of the American Educational Research Association in San Francisco.

Abedi, J. and Hejri, F.: 2004, 'Accommodations in the national assessment of educational progress for students with limited English proficiency', *Applied Measurement in Education* 17(4), 371–392.

Abedi, J., Hofstetter, C., and 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., Kim-Boscardin, C., and Larson, H.: 2000, *Summaries of Research on the Inclusion of Students with Disabilities and Limited English Proficient Students in Large-Scale Assessment*, University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.

Abedi, J. and Lord, C.: 2001, 'The language factor in mathematics tests', *Applied Measurement in Education* 14(3), 219–234.

Bennet, R.E., Rock, D.A., and Jirele, T.: 1987, 'GRE score level, test completion, and reliability for visually impaired, physically handicapped, and nonhandicapped groups', *Journal of Special Education* 21, 9–21.

Bennet, R.E., Rock, D.A., and Kaplan, B.A.: 1987, 'Level, reliability, and speededness of SAT scores for nine handicapped groups', *Special Services in Schools* 3/4, 37–54.

Bennett, R.E., Rock, D.A., and Novatkoski, I.: 1989, 'Differential item functioning on the SAT-M Braille Edition', *Journal of Educational Measurement* 26(1), 67–79.

Bielinski, J., Thurlow, M., Ysseldyke, J., Freidebach, J., and Freidebach, M.: 2001, *Read-aloud accommodation: Effects on multiple-choice reading and math items* (Technical Report 31). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Burns, E.: 1998, *Test Accommodations for Students with Disabilities*, Charles C. Thomas, Publisher, Ltd., Springfield.

Chiu, C.W.T. and Pearson, D.: 1999, June, 'Synthesizing the effects of test accommodations for special education and limited English proficient students', paper presented at the National Conference on Large Scale Assessment, Snowbird, UT.

Coleman, P.J.: 1990, Exploring visually handicapped children's understanding of length (math concepts). (Doctoral dissertation, The Florida State University, 1990). *Dissertation Abstracts International*, 51(0071).

DiCerbo, K., Stanley, E., Roberts, M., and Blanchard, J.: April, 2001, 'Attention and standardized reading test performance: Implications for accommodation', paper presented at the annual meeting of the National Association of School Psychologists, Washington, DC.

Elliott, S., Kratochwill, T., and McKevitt, B.: 2001, 'Experimental analysis of the effects of testing accommodations on the scores of students with and without disabilities', *Journal of School Psychology* 31(1), 3–24.

Comp. by: JNagalakshmi Date:5/3/07 Time:09:45:56 Stage:First Proof
File Path://spiina1001z/womat/production/PRODENV/0000000005/0000001817/
0000000016/0000556949.3D Proof by: QC by:

Francis, D.J., Lesaux, N., Kieffer, M., Rivera, H.: 2006, *Practical guidelines for the education of English language learners. Research–based recommendations for the use of accommodations in large-scale assessments*, Texas Institute for Measurement, Evaluation, and Statistics at the University of Houston for the Center on Instruction, Houston.

Fuchs, L.S., Fuchs, D., Eaton, S.B., Hamlett, C., and Karns, K.: 2000, 'Supplementing teacher judgments about test accommodations with objective data sources', *School Psychology Review* 29(1), 65–85.

Gallina, N.B.: 1989, Tourette's syndrome children: Significant achievement and social behavior variables (Tourette's syndrome, attention deficit hyperactivity disorder) (Doctoral dissertation, City University of New York, 1989). *Dissertation Abstracts International, 50*(0046).

Gándara, P., Rumberger, R., Maxwell-Jolly, J., and Callahan, R.: 2003, 'English learners in California schools: Unequal resources, unequal outcomes', *Education Policy Analysis Archives* 11(36), Retrieved June 14, 2006 from http://epaa.asu.edu/epaa/v11n36/

Hafner, A.L.: 2000, April, 'Evaluating the impact of test accommodations on test scores of LEP students & non-LEP students', paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Hakuta, K. and Beatty, A. (eds.): 2000, *Testing English-language Learners in U.S. Schools*, National Academy Press, Washington, DC.

Hollenbeck, K., Tindal, G., Stieber, S., and Harniss, M.: 1999, *Handwritten vs. Word Processed Statewide Compositions: Do Judges Rate them Differently?*, University of Oregon, BRT, Eugene, OR.

Levin, T. and Shohamy, E.: 2006, 'The role of academic language in understanding the mathematics achievements of immigrant students', in C.S. Sunal and K. Mutua (eds.), *The Enterprise of Education: Research on Education in Africa, the Caribbean, and the Middle East*, Info Age Publishing, The University of Alabama, Tuscaloosa, AL.

Kindler, A.L.: 2002, *Survey of the states' limited English proficient students and available educational programs and services, 2000–2001 Summary Report.* Washington, DC: National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs.

Kiplinger, V.L., Haug, C.A., and Abedi, J.: 2000, April, 'Measuring math – not reading – on a math assessment: A language accommodations study of English language learners and other special populations', presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Koretz, D.: 1997, The assessment of students with disabilities in Kentucky (Tech. Rep. No. 431), University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles.

Levin, Shohamy, Spolsky (2003) Au3

Mac Arthur, C.A. and Graham, S.: 1987, 'Learning disabled students' composing under three methods of text production: Handwriting, word processing, and dictation', *The Journal of Special Education* 21(3), 22–42.

Maihoff, N.A.: 2002, June, 'Using Delaware data in making decisions regarding the education of LEP students', paper presented at the Council of Chief State School Officers 32nd Annual National Conference on Large-Scale Assessment, Palm Desert, CA.

Marquart, A.: 2000, 'The use of extended time as an accommodation on a standardized mathematics test: An investigation of effects on scores and perceived consequences for students of various skill levels', paper presented at the annual meeting of the Council of Chief State School Officers, Snowbird, UT.

Mazzeo, J., Carlson, J.E., Voelkl, K.E., and Lutkus, A.D.: 2000, *Increasing the participation of special needs students in NAEP: A report on 1996 NAEP research*

Comp. by: JNagalakshmi Date:5/3/07 Time:09:45:56 Stage:First Proof
File Path://spiina1001z/womat/production/PRODENV/0000000005/0000001817/
0000000016/0000556949.3D Proof by: QC by:

activities (NCES Publication No. 2000–473), National Center for Education Statistics, Washington, DC.

Meloy, L.L., Deville, C., and Frisbie, C.: 2000, 'The effect of a reading accommodation on standardized test scores of learning disabled and non-learning disabled students', paper presented at the annual meeting of the National Council on Measurement in Education (New Orleans, LA).

Mick, L.B.: 1989, 'Measurement effects of modifications in minimum competency test formats for exceptional students', *Measurement and Evaluation in Counseling and Development* 22, 31–36.

Munger, G.F. and Loyd, B.H.: 1991, 'Effect of speededness on test performance of handicapped and non-handicapped examinees', *Journal of Educational Research* 85(1), 53–57.

No Child Left Behind Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425 (2002).

Perez, J.V.: 1980, Procedural adaptations and format modifications in minimum competency testing of learning disabled students: A clinical investigation (Doctoral dissertation, University of South Florida, 1980). *Dissertation Abstracts International, 41*(0206).

Phillips, S.E.: 1994, 'High stakes testing accommodations: Validity vs. disabled rights', *Applied Measurement in Education* 7(2), 93–120.

Rivera, C.: 2003, 'State assessment policies for English language learners', presented at the 2003 Large-Scale Assessment Conference.

Rivera, C. and Collum, E.: 2006, 'Including and accounting for English language learners in state assessment systems', in C. Rivera and E. Collum (eds.), *State Assessment Policy and Practice for English Language Learners: A National Perspective*, Lawrence Erlbaum Associates, Mahwah, NJ.

Rivera, C. and Stansfield, C.W.: 2001, April, 'The effects of linguistic simplification of science test items on performance of limited English proficient and monolingual English-speaking students', paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Rivera, C., Stansfield, C.W., Scialdone, L., and Sharkey, M.: 2000, *An Analysis of State Policies for the Inclusion and Accommodation of English Language Learners in State Assessment Programs During 1998–1999*, The George Washington University, Center for Equity and Excellence in Education, Arlington, VA.

Rogers, W.T.: 1983, 'Use of separate answer sheets with hearing impaired and deaf school age students', *B.C. Journal of Special Education* 7(1), 63–72.

Russell, M.: 1999, 'Testing writing on computers: A follow-up study comparing performance on computer and on paper', *Educational Policy Analysis Archives* 7.

Russell, M. and Haney, W.: 1997, 'Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil', *Educational Policy Analysis Archives* 5(3).

Russell, M. and Plati, T.: 2001, Effects of computer versus paper administration of a state-mandated writing assessment. *TCRecord.org.* Retrieved January 23, 2001, from the World Wide Web: http://www.tcrecord.org/PrintContent.asp?ContentID=10709.

Sireci, S.G., Li, S., and Scarpati, S.: 2003, *The Effects of Test Accommodations on Test Performance: A Review of the Literature*, Center for Educational Assessment (Research Report No. 485), School of Education, University of Massachusetts Amherst, Amherst, MA.

Solano-Flores, G. and 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.

Sullivan, P.M.: 1982, 'Administration modifications on the WISC-R performance scale with different categories of deaf children', *American Annals of the Deaf* 127(6), 780–788.

Thurlow, M.L.: 2001, April, 'The effects of a simplified-English dictionary accommodation for LEP students who are not literate in their first language',

Comp. by: JNagalakshmi Date:5/3/07 Time:09:45:56 Stage:First Proof
File Path://spiina1001z/womat/production/PRODENV/0000000005/0000001817/
0000000016/0000556949.3D Proof by: QC by:

paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Thurlow, M. and Bolt, S.: 2001, *Empirical Support for Accommodations Most Often Allowed in State Policy* (Synthesis Report 41), University of Minnesota, National Center on Educational Outcomes, Minneapolis, Retrieved on July 13, 2002, from the Website: http://education.umn.edu/NCEO/OnlinePubs/Synthesis41.html

Thurlow, M. and Liu, K.: 2001, *State and district assessments as an avenue to equity and excellence for English language learners with disabilities* (LEP Projects Report 2), University of Minnesota, National Center on Educational Outcomes, Minneapolis, MN.

Thurlow, M., House, A., Boys, C., Scott, D., and Ysseldyke, J.: 2000, *State participation and accommodation policies for students with disabilities: 1999 Update* (Synthesis Report 33), University of Minnesota, National Center on Educational Outcomes, Minneapolis, MN, Retrieved on October 9, 2006 from the Website: http://education.umn.edu/NCEO/OnlinePubs/Synthesis33.html

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., and Harniss, M.: 1998, 'Accommodating students with disabilities on large-scale tests: An empirical study of student response and test administration demands', *Exceptional Children* 64(4), 439–450.

Tippets, E. and Michaels, H.: 1997, 'Factor structure invariance of accommodated and non-accommodated performance assessments', paper presented at the National Council on Measurement in Education annual meeting, Chicago.

Uiterwijk and Vallen: 2003, 'Linguistics sources of item bias for second generation immigrants in Dutch tests', *Language Testing* (2), 211–234.

Varnhagen, S. and Gerber, M.M.: 1984, 'Use of microcomputers for spelling assessment: Reasons to be cautious', *Learning Disability Quarterly* 7, 266–270.

Walz, L., Albus, D., Thompson, S., and Thurlow, M.: 2000, *Effect of a multiple day test accommodation on the performance of special education students* (Minnesota Report 34), University of Minnesota, National Center on Educational Outcomes, Minneapolis, MN.

Watkins, M.W. and Kush, J.C.: 1988, 'Assessment of academic skills of learning disabled students with classroom microcomputers', *School Psychology Review* 17(1), 81–88.

Wright, N. and Wendler, C.: 1994, 'Establishing timing limits for the new SAT for students with disabilities', paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 4–8, 1994). ERIC ID# ED375543.

Comp. by: JNagalakshmi Date:5/3/07 Time:09:45:56 Stage:First Proof
File Path://spiina1001z/womat/production/PRODENV/0000000005/0000001817/
0000000016/0000556949.3D Proof by: QC by:

Author Query Form

Encyclopedia of Language and Education
Chapter - 11650

Query Refs.	Details Required	Author's response
AU1	Kindly provide the Subject Index for this chapter.	
AU2	'Lardon' in "Abedi, Kim-Boscardin, and Larson, 2000" has been changed to 'Larson' to make it consistent with the reference list.	
AU3	Kindly provide the complete reference for "Levin, Shohamy, Spolsky (2003)."	
AU4	Please check and confirm the shortened recto running head "Utilizing Accommodations" is fine.	

Copyright © 2009
Routledge
Taylor & Francis Group
NEW YORK AND LONDON

11

RESEARCH AND RECOMMENDATIONS FOR FORMATIVE ASSESSMENT WITH ENGLISH LANGUAGE LEARNERS

JAMAL ABEDI

Research on the instruction and assessment of English Language Learners (ELL) suggests that these students are faced with dual challenges: learning a new language and learning academic content in a language that they are trying to learn. For ELL students, the outcomes of formative assessment could contribute greatly to their academic success both in learning a new language and mastering academic content knowledge. For example, research has clearly demonstrated that ELL students are in a difficult position when they are presented with instructional and assessment materials that are linguistically complex and culturally biased. The outcomes of formative assessments could help explain how these factors affect their learning and how such sources of bias can be identified and controlled.

The purpose of this chapter is twofold. First, to provide a summary of research on the assessment of ELL students focusing on the factors that interfere with their understanding of instructional and assessment materials, and second, to present a discussion of how formative assessments can be used to improve the quality of education for ELL students. Recommendations for developing more reliable and valid formative assessments for ELLs are offered at the end of the chapter.

SUMMATIVE VERSUS FORMATIVE ASSESSMENTS FOR ENGLISH LANGUAGE LEARNERS

The main purpose of summative assessment is to provide information on what students have learned in a given period within a content area. Since these assessments are typically conducted at the end of formal classroom instruction, the results may not provide constructive feedback to teachers for improving the instruction for those students. As indicated by Herman and Baker (2005), educators "wisely recognize that information from annual state tests is often too little, too late" (p. 1). On the other hand, formative assessments are typically referred to as classroom assessments, or assessment *for* learning

PROOF

Andrade_C011.indd 181 6/16/2009 1:20:36 PM

Copyright © 2009
Routledge
Taylor & Francis Group
NEW YORK AND LONDON

(Stiggins & Chappius, 2006). Formative assessments are ongoing and enable teachers to monitor student progress in order to improve instruction before it is too late for that instruction to have an impact on overall learning. Formative assessments can provide valuable information to help teachers understand students' instructional needs as they are conducted during the instructional term (Heritage, Kim & Vendlinski, 2008; Herman, Osmundson, Ayala, Schneider, & Timms, 2006).

Therefore, the two assessments have different goals and objectives. Shepard (2000) argues that formative assessment should be used to improve learning, and calls for a change in culture in order for this to effectively happen. Shepard also indicates that the social meaning of evaluation should be revised to allow for more interaction between instruction and assessment, considering that the current perception that a single, annual, summative test can adequately identify unique student needs.

In this chapter we will elaborate on how knowledge about the content, technical aspects (reliability, validity, item characteristics) and linguistic structure of summative assessment could help in the development of formative assessments that may be useful tools in informing curriculum and instruction for ELL students. Information from formative assessment is critical for teachers of English language learners in developing an effective instructional program from which these students can benefit.

BENEFITS OF FORMATIVE ASSESSMENT

In a comprehensive review of formative assessment literature spanning kindergarten to college and across all content areas in education, Black and Wiliam (1998a) presented substantial evidence that formative assessment can increase learning outcomes and achievement. They defined formative assessment as "encompassing all those activities undertaken by teachers, and/or by their students, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged" (p. 2). In their follow-up paper, Black and Wiliam (1998b) presented strong evidence that improving formative assessment practices raises standards. Notably, many of the studies indicated that improving formative assessment practices particularly helped low achievers. Similarly, Stiggins (2002) argues that improving classroom assessment is pivotal to advancing student achievement.

Feedback from teachers to students is an important component of formative assessment. The literature calls for specific, descriptive feedback that allows students to learn from their work in order to advance further (Black & Wiliam, 1998b; Sadler, 1989; Stiggins, 2002). Herman and Choi (2008) examined the relationship of students' learning to teachers' judgment of students' understanding. The outcome of this study showed a consistent positive relationship between teachers' judgment and student learning. The study stressed the power of assessment in improving student learning using information for formative assessment.

Formative assessment can also increase students' motivation and self-esteem. In an exploratory study of 370 students, Miller and Lavin (2007) studied whether or not formative assessment helped increase the self-esteem of students aged 10 to 12. The data suggested that students' overall self-perception improved, and that these benefits increased over time. When exposed to formative assessment practices, students dem-

PROOF

Copyright © 2009
Routledge
Taylor & Francis Group
NEW YORK AND LONDON

onstrated a greater degree of self-competence and became more aware of the learning process and how to achieve their academic goals.

The literature summarized above suggests that formative assessment provides useful information for teachers in improving the quality of instruction for all students. The underlying question in this chapter is whether or not the outcome of formative assessment can improve the academic performance of ELL students given the technical issues in the assessment of these students, including the impact of language factors on the validity of their assessments.

ASSESSMENT CONSIDERATIONS FOR ENGLISH LANGUAGE LEARNERS

Many different linguistic features have been identified that negatively affect ELL students' performance. These features include those that slow down the reader, make misinterpretation more likely, and add to the reader's cognitive load, thus interfering with concurrent tasks. Examples of these features include unfamiliar vocabulary, complicated grammatical structures, and styles of discourse that include extra material, abstractions, and passive voice (Abedi, 2006b; Abedi, Lord, & Plummer, 1997). While these linguistic features may impact the academic performance of all students, including native speakers of English, the level of impact on ELL student performance is more profound.

Research findings also show that ELL students have less opportunity to learn when they are compared with their non-ELL peers (Herman & Abedi, 2004). The reduced opportunity to learn for ELL students has been linked with the linguistic complexity of instructional materials (Abedi & Herman, in press). That is, the higher the level of linguistic complexity of instructional materials, the lower the level of opportunity to learn as reported by ELL students.

The substantial performance gap between ELL and non-ELL students in many content areas may be due to factors such as the complex linguistic structure of the assessments (Maihoff, 2002; Solano-Flores & Trumbull, 2003), and the impact of cultural factors (Gándara & Rumberger, in press). Malmberg and Sumra (2001) indicated the sociocultural factors such as socioeconomic status (SES) of the family, parent education, and school-level SES/performance may impact student performance. Spinelli (2008) suggests that cultural and linguistic factors may impact the classification of students and may cause misclassification of ELL students in a learning disability category. She states that assessment measures (such as formative assessments) that fairly and accurately identify individual strengths and areas of need may result in a more accurate classification, and reduce the likelihood of misclassification of ELL students as learning disabled.

Formative assessment can be an effective and efficient tool in the hands of teachers, school officials, and curriculum planners to help identify the most influential factors contributing to their ELL students' learning and inform the design of a curriculum that addresses many of these factors. To provide useful information, formative assessment should be free of linguistic and cultural biases. Therefore, it is imperative to understand the role of language and cultural factors in assessment in general and in formative assessment in particular.

Assessments with complex linguistic structure may provide results that are unreliable,

PROOF

Copyright © 2009
Routledge
Taylor & Francis Group
NEW YORK AND LONDON

invalid, and therefore, misleading. If formative assessments are to be useful in assisting teachers and others involved in the academic career of ELL students, then these assessments should also be free of any linguistic and cultural biases. "Cultural bias" refers to any cultural factors that may differentially impact the performance of ELL students. Solano-Flores and Nelson-Barber (2001) introduced the concept of "cultural validity" as a form of test validity that links many linguistic and cultural factors to ELL students' assessment outcomes.

The Need for Formative Assessments of English Language Learners

Compared to their native, English-speaking peers, ELL students are faced with a more challenging academic career. They have to learn a new language and they have to learn content knowledge in an unfamiliar language. As elaborated by Meskill (this volume), "…there are considerable differences between learning a language in formal settings for limited use outside of the target culture ('foreign language learning') and mastering the language of the culture in which one lives and studies" (p. #). Because of the greater challenges faced by ELL students, proper assessment and appropriate instruction are critical to ensure they are not to be left behind.

The key element in providing appropriate instruction for these students is to understand their academic needs. There are many questions that can be posed with respect to ELL students' academic careers. For example, what are their linguistic needs? What is their current level of proficiency in English? In what content areas do they have difficulty? How do the linguistic factors interact with their understanding of content materials? Are they at the level of proficiency in English where they can meaningfully participate in the statewide summative assessment and benefit from instruction in English? A well-designed battery of formative assessments can address these issues and can provide valuable information for teachers and curriculum designers for these students.

Formative Assessments Inform Instruction for ELL Students

English language learner students are assessed in two different, yet related, areas: (1) their level of English language proficiency (ELP) and (2) their level of content knowledge. Thus, formative assessment covers two major areas for ELL students: assessing their level of ELP and assessing their knowledge in content areas such as mathematics, science, language arts, and social sciences. In both of these areas, the results of formative assessment can be of great value to teachers and curriculum planners in designing instructional materials that best fit their needs.

Formative Assessment of English Proficiency Students' level of English proficiency plays a very important role in their understanding of instruction in content areas (Abedi, 2007). Students who are not at the level of English proficiency necessary to understand academic instruction in English may become frustrated and may lose interest and attention. They may also be misclassified as students with learning disabilities (Abedi, 2006a; Artiles, Rueda, Salazar, & Higareda, 2005). Therefore, assessment of students' ELP level is the most important first step in providing effective and appropriate instruc-

PROOF

Copyright © 2009
Routledge
Taylor & Francis Group
NEW YORK AND LONDON

tion to these students. They must have sufficient proficiency in academic English in all four domains (reading, writing, speaking, and listening) to be able to make academic progress. Results of ELP tests are often not available to teachers at the time when they may be utilized to guide instructional planning. Information on ELL students' level of ELP is needed before instruction begins and during the instruction so that teachers can design and adjust instruction based on the students' linguistic needs.

A comprehensive ELP formative assessment can help teachers of ELL students to determine if they are proficient enough in academic English to benefit from instruction in English. English language proficiency assessment information that is obtained through formative assessment would be the most useful as it provides a timely update of students' levels of proficiency in areas that are essential in understanding instruction. For example, a low level proficiency in reading may result in low test score in math for ELL students not necessarily due to a lack of student knowledge in math, but because of students' poor English reading skills. The outcome of formative assessment in reading will help teachers understand the situation and remedy the problem.

Formative Assessment of Content Knowledge Results of studies of the opportunity to learn have shown that lack of proficiency in English may be a major obstacle in ELL students learning content knowledge (Abedi & Herman, in press; Herman & Abedi, 2004). Formative assessment can identify areas where students have difficulty in instructional and assessment materials. Teachers can then carefully examine the linguistic structure of those sections of the materials that ELL students performed poorly. If complex linguistic structures are identified, then reducing those complexities can help improve student learning.

ELL students may have the content knowledge but may be unable to express and demonstrate such knowledge due to linguistic barriers. It might be necessary to provide formative assessments that demonstrate other ways of knowing such as drawing a diagram of a life cycle to demonstrate scientific knowledge or providing hands-on-performance materials to give them a better opportunity to express their knowledge (Shavelson, Baxter, & Pine, 1991; Solano-Flores & Shavelson, 1997).

CHARACTERISTICS OF HIGH QUALITY FORMATIVE ASSESSMENTS

Traditionally, formative assessment is conducted at the classroom level where there is a direct impact on instruction. It may also be conducted at the district, state, or even national level. There are similarities and differences between formative assessments at each level. At the classroom level, formative assessments can be more focused on the content of materials taught by the teacher, and can take many different forms, such as paper and pencil format, observation of student performance, judging students' performance based on information that teachers collect during the instructional period, and in a computer format. State and national formative assessments, on the other hand, may focus on the overall state and national content standards.

At whatever level formative assessments are conducted and used, there are some basic requirements that they should meet in order to provide accurate information. Herman

PROOF

Copyright © 2009
Routledge
Taylor & Francis Group
NEW YORK AND LONDON

and Baker (2005) discussed six criteria that determine the validity and efficiency of formative assessments. These criteria include: (1) alignment, (2) diagnostic value, (3) fairness, (4) technical quality, (5) utility, and (6) feasibility. We will present our discussion of the characteristics of high quality formative assessment within this framework suggested by Herman and Baker, as it fits well with the focus on formative assessment for ELL students (see also, Linn, Baker, & Dunbar, 1991).

Alignment to Standards, Provision of Diagnostic Information, and Fairness

The first requirement for obtaining accurate information from formative assessments is that they must be aligned to content standards. State content standards define the knowledge, concepts, and skills that students should learn at each grade level. Both summative and formative assessments should represent state content standards, otherwise "their results tell us little about whether students are making adequate progress toward achieving the standard and performing well on the assessment" (Herman & Baker, 2005, p. 2). For ELL students, formative assessment should focus on two different areas: (1) students' level of English proficiency (ELP), which should be aligned with the state ELP standards, and (2) students' level of proficiency in content areas such as mathematics and science, which should be aligned with the state content standards.

The second requirement for accurate performance assessment outcomes is that the assessment is *capable of providing diagnostic information*. One of the major goals of formative assessment is to provide diagnostic information on students' academic performance. "A test with high diagnostic value will tell us not only whether students are performing well but also why students are performing at certain levels and what to do about it" (Herman & Baker, 2005, p. 5). The diagnostic aspect of performance assessment is of paramount importance for ELL students, since their level of proficiency in English determines their success in content based learning. Such information can help teachers to: (1) facilitate student learning in the English language, and (2) reduce unnecessary linguistic complexity of the instructional materials with which students have difficulty. Similarly teachers can help ELL students in the area of writing, where they may have more difficulty (e.g., extended constructed response items where students have to explain their responses).

The third requirement for accurate performance assessment outcome is *fairness***.** As Herman and Baker (2005) indicated, a fair formative test provides accurate information for all students from different cultural and linguistic backgrounds. However, variables such as unnecessary linguistic complexity and cultural factors may introduce bias into the formative assessment outcomes. Such biases may have a more profound impact on ELL students than any other subgroup of students. To provide a fair assessment for all students, including ELL students, all sources of biases should be identified and controlled.

Technical Quality of Formative Assessments

The fourth requirement for accurate performance assessment outcome is technical quality. Assessments that are reliable and valid provide accurate information about

PROOF

Copyright © 2009
Routledge
Taylor & Francis Group
NEW YORK AND LONDON

what students know and are able to do. There are many different factors that could negatively impact the reliability and validity of assessments for all students (e.g., Allen & Yen, 1979; Thorndike, 2005). Assessments for ELL students may have additional sources of threat to reliability and validity, such as the unnecessary linguistic complexity of the assessment.

Results of studies on the assessment of ELL students have demonstrated that the unnecessary linguistic complexity of content-based assessments is a likely source of measurement error, having more impact on the reliability of assessment for the ELL subgroup. The linguistic complexity of test items as a source of construct-irrelevant variance may also influence the validity of assessment for these students (Abedi, 2006b). Results of analyses of existing data show a substantial gap in reliability (internal consistency) and validity (concurrent validity) between ELL and non-ELL students on test items that are linguistically complex (Abedi, 2006b,). A more detailed discussion of reliability and validity of formative assessments and suggestions on how to improve the technical quality of such assessments for ELL students may be helpful to highlight these important characteristics.

Reliability and Validity Reliability problems arise when responses are not consistent across repeated testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Thorndike, 2005). For example, when students do not understand the language of test items, their responses may vary on different occasions of taking the same test (Abedi, 2006b). Results of analyses of data from multiple locations across the United States indicate a large gap in the reliability coefficients obtained from samples of ELL and non-ELL students. This is mainly due to the multidimensional nature of assessment outcomes for ELL students, as these assessments are influenced by language factors.

The gap in reliability decreases as the level of language demand of the assessment decreases. For example, the reliability coefficients (alpha) for native speakers of English students ranged from .898 for math to .805 for science and social science. For ELL students, however, alpha coefficients differed considerably across the content areas. In math, where language factors might not have as much influence on performance, the alpha coefficient for ELL (.802) was slightly lower than the alpha for English-only students (.898). For English language arts, science, and social science; however, where there is more language involved, the gap of the alpha coefficient between English-only and ELL students was substantially larger. Averaging over English language arts, science, and social science, the alpha coefficient for English-only students was .808 as compared to an average alpha of .603 for ELL students.

To improve the reliability of formative assessments for ELL students, all sources of measurement error, including biases due to linguistic and cultural factors, should be identified and controlled. Abedi (2006b) provides information about how to conduct linguistic modification of test items to reduce the level of unnecessary linguistic complexity of the test and increase the reliability of assessments for these students.

The main validity issue is "whether the test measures what we want to measure, all of what we want to measure, and nothing but what we want to measure" (Thorndike, 2005, p. 145). Therefore, if the test measures anything other than the focal construct(s), then

PROOF

Copyright © 2009
Routledge
Taylor & Francis Group
NEW YORK AND LONDON

the validity of the interpretations of scores on the test is diminished. For example, if test items on a mathematics test have a complex linguistic structure, then the test measures not only the construct relevant to the purpose of the test (mathematics), it also measures a construct that is irrelevant to the purpose of the test (language). Thus, linguistic factors may seriously affect the validity of inferences drawn using this assessment. Results of analyses of existing state and national data show that the higher the level of language demand in the test, the higher the performance gap between ELL and non-ELL students due to the impact of construct-irrelevant factors (Solano-Flores, 2008).

Utility and Feasibility

The fifth criterion for formative assessment is utility. That is, formative assessments should provide useful information for teachers, students, and parents. For ELL students, high utility formative assessments provide diagnostic information on English proficiency as a prelude to learning academic content. High utility formative assessments also provide useful feedback on student learning in content areas. Due to the immediacy of formative assessments, such feedback will help teachers to revise instructional and assessment materials to address ELL students' academic needs.

The sixth criterion for efficient formative assessments is *feasibility*. National, state, district, and classroom assessments take a substantial amount of students' time in schools. Teachers complain that too much testing takes time away from instruction. Therefore, assessments should be worth the extra time and resources that are needed for conducting them. High quality formative assessments inform instruction, offer ongoing feedback to students throughout the learning process, and provide useful information for teachers and curriculum planners necessary to the design of effective instruction. Because of their highly challenging academic careers, this is especially important for ELL students.

RESEARCH ON THE ASSESSMENT OF ELL STUDENTS

Due to the importance of formative assessment in shaping and improving instruction for all students, substantial attention has been paid to research in this area for the general student population. However, research on formative assessments for ELL students is scarce. There has been a great deal of attention given to summative assessments for ELL students due to their importance in state and national assessment and accountability requirements (e.g., Abedi, 2004; NCLB, 2002). The findings from research on summative assessment could, to some degree, be applied to formative assessment for ELL students. Therefore, in the absence of any major systematic effort to address issues concerning formative assessments specifically for ELL students, a summary of some of the studies focused on assessment in general of ELL students is provided in this section. Findings from summative assessments of ELL students can then be applied in developing reliable and valid formative assessments for these students.

Language factors greatly influence assessment outcomes for ELL students. The two case scenarios below illustrate this point. These scenarios are based on findings from research on the assessment of ELL students (see, for example, Abedi, 2002, 2006b).

PROOF

Copyright © 2009
Routledge
Taylor & Francis Group
NEW YORK AND LONDON

Case One. Ricardo is a fourth grade student who recently transferred from a high-ranking public school in Mexico. He passed grade 4 math for the first semester with a high score (at the above-proficient level) in Mexico. By the end of the third week in a U.S. school, he was tested again on his math content knowledge. He got a very low score (below proficient). His U.S. teacher is not sure whether Ricardo's low score is due to lack of math content knowledge or lack of understanding of the math test items.

Case Two. Jose is another fourth grade student who transferred to the same U.S. school that Ricardo attends. His math score at the school from his native country was quite low (below proficient). He obtained a similar score in the U.S. school. The U.S. teacher, who does not know enough about Jose's academic background, has difficulty explaining his performance.

As evident from the two case scenarios presented above, it is extremely difficult to interpret assessment outcomes (whether formative or summative) when they are confounded with linguistic and cultural factors. Language factors affect performance outcomes, especially for English language learners. Students' content knowledge in areas such as mathematics, science, or social studies may not be truly assessed if students cannot understand the language of the test (Abedi, 2006b; Kiplinger, Haug, & Abedi, 2000; Maihoff, 2002; Solano-Flores & Trumbull, 2003). Research shows that even minor changes in the wording of content-related test items can change ELL student performance (Abedi & Lord, 2001; Abedi, Lord, Hofstetter, & Baker, 2000; Abedi, Lord, & Plummer, 1997; Cummins, Kintsch, Reusser, & Weimer, 1988; De Corte, Verschaffel, & DeWin, 1985; Hudson, 1983; Riley, Greeno, & Heller, 1983).

Linguistic Complexity

Some studies have focused on the effects of the linguistic complexity of assessment on the performance of ELL students. For example, results of analyses of the National Assessment of Educational Progress (NAEP) data (Abedi, Lord, & Plummer, 1997) show that ELL students had difficulty with the test items that were longer and were more linguistically complex. The study also found that ELL students exhibited a substantially higher number of omitted/not-reached test items since it took them much longer to read and understand assessment questions.

In analyzing test data from four different U.S. locations, Abedi (2002) compared the performance of ELL and non-ELL students in several different content areas. Among these content areas, reading has the highest level of language demand since language is central to the construct being measured. However, in the science and math tests, understanding of the science and math content—not the language—is the focus of assessment. At one of the data sites, ELL students in grade 10 had a mean reading score of 24.0 ($SD = 16.4$) as compared with a mean reading score of 38.0 ($SD = 16.0$) for non-ELL students, a difference of 14 score points. The difference between ELL and non-ELL mean NCE scores for science was 9.7, substantially less than the 14 score points difference in reading. For math, the difference in average scores between ELL and non-ELL students was 2.8. For 11th grade students, the ELL/non-ELL performance difference was 15.9 for reading, 11.2 for science, and close to 0 for math computation. These results were consistent with the results of analyses from the other data sites in the study.

PROOF

Copyright © 2009
Routledge
Taylor & Francis Group
NEW YORK AND LONDON

To reduce the level of impact of unnecessary linguistic complexity on the assessment of ELL students, a linguistic modification approach to the content-based assessment was proposed (Abedi, Lord, & Plummer, 1997). In this approach, guidelines are provided for how to revise assessment questions to be more accessible in terms of linguistic structure, without affecting the construct being measured. The researchers developed a linguistically modified version of the assessment based on the proposed framework and compared student performance taking this modified assessment with the performance of another group taking the original form of the assessment. The original and modified versions were randomly assigned to students. In general, the results suggested that the linguistic modification approach makes assessments (both summative and formative) more accessible to ELL students.

In one study, the impact of linguistic complexity of assessment was tested on a sample of 1,031 eighth grade students in Southern California (Abedi & Lord, 2001). The math items for eighth grade students were modified to reduce the complexity of sentence structures and to replace potentially unfamiliar vocabulary with more familiar words without changing the content-related terminologies (i.e., mathematical terms were not changed). The results showed significant improvement in the scores of ELL students and also non-ELLs in low and average level mathematics classes, but the changes did not affect the scores of higher performing non-ELL students, since those students understand complex math problems without needing much language context.

The outcome of this study cross-validated another study in which the impact of language factors on the mathematics performance of English learners was examined (Abedi, Lord, Hofstetter, & Baker, 2000). This study was conducted on a sample of 1,394 eighth graders in schools with a high enrollment of Spanish speakers. Results showed that modification of the language contributed to improved performance on 49% of the items; the ELL students generally scored higher on shorter/less linguistically complex problem statements. The results of this study also suggest that lower performing native speakers of English benefited from the linguistic modification of the assessment as well.

Other studies were conducted to obtain cross-validation evidence on the impact of language factors on the assessment of ELL students. These studies examined the impact of language factors by focusing on the effectiveness of the language modification approach in reducing the performance gap between ELL and non-ELL students. In one of these studies, which included 1,594 eighth grade students, test items from the NAEP and the Third International Math and Science Study (TIMSS) were used (Abedi, Courtney, & Leon, 2003). The results indicated that the linguistically modified version of the test improved the ELL students' scores without affecting the non-ELL students' scores. Other studies have had similar results (e.g., Maihoff, 2002), including those that involved students in grades 4 and 8 (Kiplinger, Haug, & Abedi, 2000; Rivera & Stansfield, 2001).

In summary, the research evidence shows that linguistic complexity is a major source of measurement error in assessment results for ELL students. Research findings also suggest that reducing the level of unnecessary linguistic complexity of assessments (linguistic modification) may help improve assessment validity and reliability for these students. Some people argue that reducing the complexity of academic content may

PROOF

Copyright © 2009
Routledge
Taylor & Francis Group
NEW YORK AND LONDON

change the construct being taught and assessed. However, in the language modification approach, the language factors that are related to the content of assessment and instruction are distinguished from the unnecessary linguistic complexity of the text in both assessment and instruction, and modifications are focused only on the language that is unrelated to the content being measured. Decisions about what was language-related and what was language-unrelated was made by a team of content and linguistic experts.

Research findings presented in this section showed substantial performance gaps between English language learners and their native English speaking peers. Because there is no evidence to suggest any difference between ELL and non-ELL students in their ability to learn, these gaps are alarming. The evidence suggests that the lower performance of ELL students is mainly due to the impact of language factors on instruction and assessment. While the research on the impact of language on assessment has been conducted mainly in the area of summative assessments, the findings of these studies can be generalized to formative assessments as well.

GUIDELINES FOR CREATING RELIABLE, VALID, AND EFFECTIVE FORMATIVE ASSESSMENTS FOR ELLS

For formative assessments to provide useful information, they must be reliable, valid, and comprehensive in content. There are many factors that may impact the reliability and validity of formative assessments that should be considered (e.g., Allen & Yen, 1979; Thorndike, 2005). Among the most important factors to consider in the development of formative assessments for ELL students is controlling for sources of construct-irrelevant variance. Linguistic and cultural biases could impact both the reliability and validity of inferences from assessments of ELL students. The outcomes of formative assessments which are highly confounded with such sources of bias may not be useful.

General Recommendations for Creating Formative Assessments for ELLs

Some general recommendations should be noted at the outset. First, formative assessments should be based on state content standards, to the extent possible. Formative assessments based on state content standards provide objective and comprehensive information about student levels of achievement on a broader scale. Second, the technical characteristics of formative assessments should be clearly examined. Among these characteristics, data on validity, reliability, and item bias through differential item functioning analysis (DIF) should be provided (e.g., Abedi, Leon, & Kao, 2008; Martiniello, 2008). Information on the items that function differently across ELL/non-ELL categories could be of great value to teachers because such information could help identify possible sources of bias in instructional materials.

A third general recommendation is that the assessment should follow a standard test administration protocol to the extent possible, so that the outcome of these assessments can be comparable across different conditions. Finally, formative assessments for ELL students should include items that address linguistic issues. An English language proficiency (ELP) assessment should include the four major domains (reading, writing,

PROOF

Copyright © 2009
Routledge
Taylor & Francis Group
NEW YORK AND LONDON

listening, and speaking) and content-based assessments should include questions to address academic language proficiency (Bailey, Butler, & Sato, 2007).

Creating Reliable and Valid Formative Assessments

Although it may be difficult to replicate formal research-based testing methods in classroom settings, the following recommendations are suggested when possible. For more informal and frequent formative assessments, some of these suggestions may be impractical, but for more formal and less frequent grade level, school- or district-wide formative assessments, these steps are strongly recommended.

Use a Reasonable Number of Questions Determine the maximum number of questions that can be included in terms of class time period. Remember that an assessment with too many questions may take too much time out of instruction, and assessments with too few test items may not be reliable or have sufficient content coverage to provide accurate information. As a rule of thumb to create a reliable test, we recommend a minimum of 20 test items.

Determine the Format of the Test The format may not need to follow the state summative assessment. It should have different types of items such as multiple-choice and written response items. The written response (open-ended) items should include short essays (extended constructive response) as these items provide an opportunity for teachers to evaluate students' writing ability on content assessments. The test may also include other types of assessment, such as portfolios and performance assessments.

Establish Content and Construct Validity Ensure that the content of the formative assessment corresponds to the state content standards in the relevant subject areas. Such information is usually available through state department of education websites.

Write All Questions Clearly and Concisely Avoid language that is linguistically complex and culturally biased. Table 11.1 provides descriptions of low, moderate, and high linguistic complexity. There are at least five guidelines that can help in reducing the linguistic complexity of assessments. These include: (1) avoid words that are unfamiliar or rarely used; (2) avoid test items that are very long or have complex sentence construction; (3) use concrete terms, since items that are presented in abstract forms are more difficult for ELL students to understand; (4) do not use difficult subordinate, conditional, or adverbial clauses; and (5) use the active voice and concrete presentations of sentences to the extent possible, since ELL students have difficulty with passive voice and abstract or impersonal presentations.

Get Feedback From Students Whenever possible, discuss the assessment with two or three students in another class studying the same content, at the same grade level, who are not likely to take the test. Ask them individually about areas that they have difficulty understanding, due either to language or cultural issues.

PROOF

Copyright © 2009
Routledge
Taylor & Francis Group
NEW YORK AND LONDON

Table 11.1 Continuum of Linguistic Complexity and Item Characteristics

1	Items with no linguistic complexity: ✓ Familiar or frequently used words; word length generally shorter ✓ Short sentences and limited prepositional phrases ✓ Concrete item(s) and a narrative structure ✓ No complex conditional or adverbial clauses ✓ No passive voice or abstract or impersonal presentations
2	Items with a minimal level of linguistic complexity: ✓ Familiar or frequently used words; short to moderate word length ✓ Moderate sentence length with a few prepositional phrases ✓ Concrete item(s) ✓ No subordinate, conditional, or adverbial clauses ✓ No passive voice or abstract or impersonal presentations
3	Items with a moderate level of linguistic complexity: ✓ Unfamiliar or seldom used words ✓ Long sentence(s) ✓ Abstract concept(s) ✓ Complex sentence/conditional tense/adverbial clause(s) ✓ A few passive voice or abstract or impersonal presentations
4	Items with a high level of linguistic complexity: ✓ Relatively unfamiliar or seldom used words ✓ Long or complex sentence(s) ✓ Abstract item(s) ✓ Difficult subordinate, conditional, or adverbial clause(s) ✓ Passive voice/abstract or impersonal presentations
5	Items with a maximum level of linguistic complexity: ✓ Highly unfamiliar or seldom used words ✓ Very long or complex sentence(s) ✓ Abstract item(s) ✓ Very difficult subordinate, conditional, or adverbial clause(s) ✓ Many passive voice and abstract or impersonal presentations

Get Feedback From Colleagues Whenever possible, ask a colleague with a linguistic background to review a test using the information provided in Table 11.1 to identify items that are linguistically complex. Reduce the level of linguistic complexity of items based on the feedback from the colleague and students in step 5.

Check Reliability Whenever possible, estimate the reliability of newly developed formative assessments by giving the test to students in a class taught by a colleague, and give it again to the same students after a week or so. Compare responses given by the same students on the first and second administration of the test and look for consistency between the two administrations. As a rule of thumb, items with 80% (or higher) agreement can be considered reliable.

Providing Feedback Based on the Outcome of Formative Assessments

When providing feedback to ELL students based on the results of formative assessments, the following four guidelines can ensure the feedback is useful to students: (1) Be as specific as possible and avoid general terms such as *good*, *poor*, or *not adequate*; (2) Think

PROOF

Copyright © 2009
Routledge
Taylor & Francis Group
NEW YORK AND LONDON

of providing feedback as consisting of the same steps taken in teaching a lesson, such as identification of a goal, direct instruction, guided practice, independent practice, and assessment. Provide a clear goal for a task, explain the expectations for the task, provide instruction and guided support for achieving the goal, including interacting with the student to check for understanding and whether or not he or she has the prerequisite skills and strategies, and finally, let the student work independently before reassessing progress toward the goal. (3) Present evidence from formative assessment about how the lack of language proficiency could lead to a lack of understanding of the content. For example, show how unfamiliar vocabulary resulted in a lack of understanding of content. (4) When providing ELP feedback, be sure to include the student's areas of need in each of the four domains—reading, writing, speaking, and listening.

SUMMARY AND SUGGESTIONS

The No Child Left Behind (2002) Act, which is the most recent reauthorization of the Elementary and Secondary Act (ESEA) of 1965, mandates the inclusion of all students in statewide accountability to promote higher achievement for every student, including English language learners. However, there are major issues concerning the instruction and assessment of ELL students. As discussed in this chapter, research on the assessment of ELL students shows a substantial performance gap between ELL and non-ELL students. While the inclusion mandate highlights the need for attention to the academic careers of ELL students, it should be considered only the first step. Including ELL students into mainstream instruction and assessment without proper attention to their academic needs could have grave consequences for their academic futures.

English language learner students constitute a heterogeneous group. They are from different countries with different language backgrounds, and different levels of proficiency in English and their native language. Therefore, the same curriculum may not serve all ELL students. Formative assessments can be the best source of information for teachers and others who are involved in their academic careers in recognizing the individual needs of these students and helping them reach the level of academic achievement that every student deserves.

English language learner students face very challenging academic careers because they must learn a new language and learn new content knowledge in that new language. Summative assessment outcomes may provide useful information on the academic progress of ELL students. However, information about the outcomes of these assessments may be too little too late. These outcomes come to light when instruction has officially ended and teachers may no longer be able to use assessment results in addressing ELL needs through instruction.

In addition, the assessment and accountability system for ELL students is more complex than for many other student subgroups. Because ELL students must first learn the English language in order to be able to learn content knowledge in English, the main issue is whether they have reached the level of proficiency in English that is needed to benefit from instruction and assessment in an English-only environment. Formative assessments can shed light on the issues in two areas. First, formative assessment can assess students' levels of proficiency in different English language proficiency domains

PROOF

Copyright © 2009
Routledge
Taylor & Francis Group
NEW YORK AND LONDON

(reading, writing, speaking, and listening) to determine their readiness to participate in mainstream instruction and assessment. Formative assessment can then examine students' performance in content-based areas to determine if the linguistic complexity of instructional materials (teacher lectures, textbooks, etc.) affects students' understanding of instruction.

Formative assessments are typically constructed and used at the classroom level by teachers. While this provides a more direct representation of what the teacher has taught, it may not have the technical quality that such an important assessment should have. Teachers often may not have the technical background that is needed to develop a sound formative assessment system and may not have resources for pilot and field testing these assessments. Furthermore, the teacher-made formative assessments may not cover state content standards that should guide instruction and assessment for all students. On the other hand, formative assessments developed by test publishers or states may not be at the level of specificity that teachers would desire. It is therefore imperative to pay careful attention to both the content and technical characteristics of formative assessments that are used for students.

Test developers and textbook writers must also be prepared to develop formative assessments for ELL students based on findings from research on assessments for these students. Such research points specifically to the impact of unnecessary linguistic complexity of instruction and assessment. It is important for the developers of formative assessments for ELL students to have a good understanding of the impact of linguistic and cultural factors on their assessments and to incorporate suggestions recommended by research in this area into the assessment. Assessments (whether formative and summative) that are free from unnecessary linguistic complexity will be more accessible to all students.

ACKNOWLEDGMENTS

The author acknowledges the contributions of Rita Pope, who contributed substantially with editorial comments and assisted in structuring and revising the paper, as well as Shannon Cannon, who provided valuable comments and suggestions during the revision process.

REFERENCES

Abedi, J. (2002) Standardized achievement tests and English language learners: Psychometrics issues. *Educational Assessment, 8*(3), 231–257.

Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher, 33*(1), 4–14.

Abedi, J. (2006a). Psychometric issues in the ELL assessment and special education eligibility. *Teacher's College Record*, *108*(11), 2282–2303.

Abedi, J. (2006b). Language issues in item-development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 377–398). Mahwah, NJ: Erlbaum.

Abedi, J. (2007) (Ed.). *English language proficiency assessment in the nation: Current status and future practice.* Davis, University of California.

Abedi, J., Courtney, M., & Leon, S. (2003). *Effectiveness and validity of accommodations for English language learners in large-scale assessments* (CSE Tech. Rep. No. 608). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

PROOF

Copyright © 2009
Routledge
Taylor & Francis Group
NEW YORK AND LONDON

Abedi, J., & Herman, J. (in press). Assessing English language learners' opportunity to learn mathematics: Issues and limitations. *Teachers College Record*.

Abedi, J., Leon, S., & Kao, J. (2008). *Examining differential item functioning in reading assessments for students with disabilities.* Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, *14*(3), 219–234.

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.

Abedi, J., Lord, C., & Plummer, J. (1997). *Language background as a variable in NAEP mathematics performance* (CSE Tech. Rep. No. 429). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks-Cole.

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.

Artiles, A. J., Rueda, R., Salazar, J., & Higareda, I. (2005). Within-group diversity in minority disproportionate representation: English language learners in urban school districts. *Exceptional Children*, *71*, 283–300.

Bailey, A., Butler, F., & Sato, E. (2007). Standards-to-standards linkage under Title III: Exploring common language demands in ELD and science standards. *Applied Measurement in Education, 20*(1), 53–78.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-148.

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology*, *20*, 405–438.

De Corte, E., Verschaffel, L., & DeWin, L. (1985). Influence of rewording verbal problems on children's problem representations and solutions. *Journal of Educational Psychology*, *77*(4), 460–470.

Gándara, P., & Rumberger, R. (in press). Immigration, language, and education: How does language policy structure opportunity? In J. Holdaway & R. Alba (Eds.), *Education of immigrant youth: The role of institutions and agency*. New York: Social Science Research Council.

Heritage, M., Kim, J., & Vendlinski, T. (2008, March). *From evidence to action: A seamless process in formative assessment?* Paper presented at the American Educational Research Association Annual Meeting, New York.

Herman, J. L., & Abedi, J. (2004). *Issues in assessing English language learners' opportunity to learn mathematics* (CSE Tech. Rep. No. 633). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Herman, J. L., & Baker, E. L. (2005). Making Benchmark Testing Work. *Educational Leadership, 63*(3), 48–54.

Herman, J. L., & Choi, K. (2008). *Formative assessment and the improvement of middle school science learning: From assessment to evidence.* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Herman, J. L., Osmundson, E., Ayala, C., Schneider, S., & Timms, M. (2006). *The nature and impact of teachers' formative assessment practices* (CSE Technical Report 703). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Hudson, T. (1983). Correspondences and numerical differences between disjoint sets. *Child Development, 54*, 84–90.

Kiplinger, V. L., Haug, C. A., & Abedi, J. (2000, April). *Measuring math—not reading—on a math assessment: A language accommodations study of English language learners and other special populations*. Presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Linn, R., Baker, E. L., Dunbar, S. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20,* 15–21.

Maihoff, N. A. (2002, June). *Using Delaware data in making decisions regarding the education of LEP students.* Paper presented at the Council of Chief State School Officers 32nd Annual National Conference on Large-Scale Assessment, Palm Desert, CA.

Malmberg, L., E., Sumra, S. (2001) Socio-cultural factors and Tanzanian primary school students' achievement and school experience [Special issue]. *Journal: Utafiti,* n.s.*4*, 207–219.

Meskill (in press). Moment-by-moment formative assessment of second language development: ESOL profes-

PROOF

Copyright © 2009
Routledge
Taylor & Francis Group
NEW YORK AND LONDON

sionals at work. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment.* Albany: State University of New York.

Miller, D., & Lavin, F. (2007). But now I feel I want to give it a try: Formative assessment, self-esteem and a sense of competence. *The Curriculum Journal, 18*(1), 3–25.

Martiniello, M. (2008). *Language and the performance of English-language learners in math word problems. Harvard Educational Review, 78*(2). Retrieved August 26, 2008, from http://www.edreview.org/harvard08/2008/su08/s08marti.htm

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Riley, M. S., Greeno, J. G., & Heller, J. I. (1983). Development of children's problem-solving ability in arithmetic. In H. P. Ginsburg (Ed.) *The development of mathematical thinking*. New York: Academic Press.

Rivera, C., & Standsfield, C. W., (2001, April). *The effects of linguistic simplification of science test items on performance of limited English proficient and monolingual English-speaking students.* Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*, 119–144.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education, 4*(4), 347–362. Retrieved August 26, 2008, from http://www.informaworld.com/10.1207/s15324818ame0404_7

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4–14.

Solano-Flores, G. (2008). 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. *Educational Researcher, 37*(4), 189–199.

Solano-Flores, W., & Nelson-Barber, S. (2000, April). *Cultural validity of assessments and assessment development procedures.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Solano-Flores, G., & Shavelson, R.J. (1997). Development of performance assessments in science: Conceptual, practical, and logistical issues. *Educational Measurement: Issues and Practice, 16*(3), 16–24.

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.

Spinelli, C. G. (2008). Addressing the issue of cultural and linguistic diversity and assessment: Informal evaluation measures for English language learners. *Reading & Writing Quarterly*, *24*(1), 101–118.

Stiggins, R. (2002). Assessment crisis: The absence of assessment *for* learning. *Phi Delta Kappan, 83*(10), 758–765.

Stiggins, R., & Chappius, J. (2006). What a difference a word makes. *Journal of Staff Development 27*(1), 10–14.

Thorndike, R. M. (2005). *Measurement and Evaluation in Psychology and Education*. Upper Saddle River, NJ: Pearson, Merrill.

PROOF

Classification System for English Language Learners: Issues and Recommendations

Jamal Abedi, *University of California, Davis*

High-stakes decisions for the instruction and assessment of English language learner (ELL) students are made based on the premise that ELL classification is a valid dichotomy that distinguishes between those who are proficient in the use of the English language and those who are not. However, recent research findings draw a vague picture of the term "ELL" and call for a more valid classification system for ELL students. Thus, the purpose of this paper is twofold: (1) to reveal issues concerning the validity of the current ELL classification system based on the results of several empirical studies, and (2) to initiate a discussion on ways to improve the validity of the ELL classification system by proposing a system that uses existing multiple criteria in a stepwise manner. While the suggested system has its own limitations and controversies, we hope this discussion stimulates thoughts and brings much needed attention to this very important national issue.

Keywords: English language learner, validity, ELL classification, assessment, No Child Left Behind

Recent federal legislation, such as the Improving America's Schools Act of 1994 and the No Child Left Behind Act (NCLB) of 2001, address the need to advance the quality of teaching and learning for *every* child, including those who are English language learners (ELL).[1] On the other hand, research on fair, valid, and effective assessment has brought into question existing ELL classification policies and practices. Improper classification may render assessment results unfair, invalid, and ineffective, which may lead to inappropriate and inadequate instruction for ELL students. Validity problems in ELL classification and assessment may also affect accountability, such as in reporting Adequate Yearly Progress (NCLB, 2002) for ELL students. Misleading results of inaccurate classification and invalid assessment may lead to disproportionately placing ELL students in special education classrooms where it may negatively affect their academic career and may take them a longer time to graduate (Stefanakis, 1998).

Improved validity and consistency in classification of ELL students is of utmost importance as these students continue to be a fast-growing population. According to a recent report by the U.S. Government Accountability Office, about 5 million ELL students were enrolled in schools, representing approximately 10% of all public school students (GAO, 2006). Between 1990 and 1997, the number of U.S. residents born outside the United States increased by 30%, from 19.8 million to 25.8 million (Hakuta & Beatty, 2000). Approximately 1.6 million in the state of California alone are considered English learners (Gándara, Maxwell-Jolly, & Driscoll, 2005). This rapid growth demands that we consistently and accurately determine which students require English language services (Abedi & Gandara, 2006). To make this determination, we first need to accurately identify ELL students.

A discussion of the ELL classification system must consider the validity of information collection methods and how information can be used more effectively. The purpose of this paper is twofold: first, to bring issues concerning validity of current ELL classification to the attention of assessment experts, researchers, educational practitioners and policymakers; and second, to initiate a discussion on how existing information can effectively be used to improve the validity of the ELL classification system.

Revealing the Issues

For the purposes of this paper, students who are not considered ELL are referred to as "non-ELL." The non-ELL group usually consists of native English speakers, students from non-English-speaking homes who are fluent in English at the time of school entry (initially fluent English proficient, IFEP), and students who progress out of the ELL category (redesignated fluent English proficient, RFEP, National Clearinghouse for English Language Acquisition, 2002).

One would expect a uniform approach in assigning an ELL classification code (1 = ELL, 0 = non-ELL) to students across the nation. If states or at least school districts within a state have adopted a uniform definition, then one would expect a student who is classified as ELL at one school to be similarly classified at another school. However, results from several studies have suggested otherwise (see, for example, Abedi, Lord, Hofstetter, & Baker, 2000). A review of data from 12 schools revealed different systems for determining a student's level of English

Jamal Abedi is a Professor, University of California, Davis, School of Education, One Shields Avenue, Davis, CA 95616; jabedi@ucdavis.edu.

 Copyright © 2008 by the National Council on Measurement in Education

proficiency (Abedi, Lord, & Plummer, 1997). Additionally, Rivera, Stansfield, Scialdone, and Sharkey (2000) found in their review of state policies that the ELL definition provided by about half of the participating states differed widely in content. Linquanti (2001) found that the criteria used for initially classifying language-minority students as ELL in California—largely based on English language proficiency (ELP)—are different from the multiple criteria (linguistic and academic) used to reclassify them as RFEP, and that these latter criteria vary across districts within the same state.

Cisneros and Leone (1995) believed that determining the exact number of ELL students in elementary and secondary schools was not easy, since ELL definitions vary so widely from state to state. They indicated:

> Due to the broad definition of "limited English proficient" in the Bilingual Education Act (BEA) and lack of clearly outlined procedures for identifying ELL students, future reauthorization of federal legislation will need to define such terms and clearly outline procedures for identification of ELL students . . . . (p. 362)

In search of a model for classification, one might ask what the national criteria are for including ELL students in large-scale assessments. The National Assessment of Educational Progress (NAEP) does not directly define "ELL," but includes those ELL students who participate in the regular state assessments (NAEP, 2007). Thus, NAEP's identification of ELL students is based on states' classification policy.

The definition of an ELL [LEP] student, as outlined in the No Child Left Behind Act, Title IX #25 (NCLB, 2002) is: (a) age 3 through 21; (b) enrolled or preparing to enroll in an elementary or secondary school; (c) not born in the United States or whose native language is not English; (d) is a Native American, Alaskan Native, or a native resident of the outlying areas; (e) comes from an environment where a language other than English has had a significant impact on an individual's level of ELP; (f) is migratory and comes from an environment where English is not the dominant language; or (g) has difficulties in speaking, reading, writing, or understanding the English language that may deny the individual the ability to meet the state's proficient level of achievement and the ability to successfully achieve in classrooms where English is the language of instruction, or to participate fully in society (NCLB, 2002, Title IX).

The above definition is primarily based on two sources of information: (1) students' language background information and (2) their level of English proficiency. Information on the language background of students (for example, country of birth, native language, and type and amount of a language other than English spoken at home) comes typically from a parent-completed Home Language Survey (HLS) which is described below. Information on the students' level of English proficiency in speaking, reading, writing, listening, and comprehension comes from existing tests of English proficiency. However, research shows major concerns with the reliability and validity of these sources of information.

Home Language Surveys

The HLS determines which students should undergo English language assessment and possibly receive instruction designed for speakers of other languages. The HLS (which may differ across states in term of format and type of questions) is usually used just to identify linguistic minority (i.e., potential ELL) status. The main purpose of the survey is to identify what languages are spoken at home. Some school districts require that the HLS be administered to families of every entering student and that those results be included in every student's permanent file. According to the *Survey of the States' ELL Students*, over 80% of schools made use of some form of HLS (Kindler, 2002). Unfortunately, the validity of HLS data could become questionable. Parents may give inconsistent information for a variety of reasons, including concerns over equity of opportunity for their children, citizenship issues, and poor comprehension of the survey form or interview (Abedi et al., 1997; Littlejohn, 1998).

Littlejohn (1998) questioned the validity of HLS information. He used as an example a student who has always spoken English but who had a relative in the home for a period of time who spoke Spanish would, under the Office for Civil Rights' (OCR), be classified as "PHLOTE" (primary home language other than English) (p. 8). On the other hand, some parents claimed that English was the language spoken at home because their children practiced English as a second language at home. Littlejohn indicated that this was enough of a problem that the Denver Office of Education added the cautionary note of "Do not list languages learned or used only academically" (p. 10).

As suggested above, in addition to the information from an HLS, student assessment outcomes from both content-based and ELP tests are also used as criteria for ELL initial classification (especially in later grades) and reclassification. Below is a short review of these criteria that are used for multiple purposes including for classification of ELL students.

English Language Proficiency Tests

In addition to the HLS, ELP tests are commonly used for identifying ELL students. The *Survey of the States' LEP Students* revealed that 94% of those surveyed used some type of ELP test for ELL classification and placement (Kindler, 2002). While this more objective criterion is a valuable addition to the ELL classification process, there are several major concerns with the conceptual framework and psychometric characteristics of many ELP tests.

ELP tests can be grouped in two different categories: tests prior to the implementation of NCLB (pre-NCLB) and tests that were newly developed based on the NCLB Title III requirements (post-NCLB). To ensure accurate classification of ELL students, it is necessary to examine the validity of ELP assessment outcomes (for both pre- and post-NCLB tests) as criteria for ELL classification. In doing so, we first need to understand the theories behind English language acquisition and then examine validity issues concerning these assessments. We can then demonstrate how improvements in ELP assessments affect the validity of ELL classification.

Theories of Second Language Acquisition

Understanding the process of language acquisition for ELL students is essential to developing a more valid assessment and classification system for these students. However, different views and theories on second language acquisition complicate the issue (Conteh-Morgan, 2002; Francis & Rivera, 2007; Reutzel & Cooter, 2007). The behavioral theorists believe that language development is influenced by environmental stimuli, such as imitation, rewards,

and practice. On the other hand, the innatist theorists believe that learning is a natural process through a human built-in device for learning language. For example, according to Chomsky (1968) language is modeled by internal factors and then shaped through experience. Similarly, Krashen (1988) suggests that humans are born with the ability to learn language (see also Lightbown & Spada, 2000).

Critics of the innatist theory argue against the claim that internal factors fully explain language acquisition process. They believe that environmental factors such as exposure to rich learning environments and interaction with others influence language acquisition (Reutzel & Cooter, 2007). These cognitive theorists believe that the process of language acquisition may in turn affect cognitive and social skill development (Reutzel & Cooter, 2007). Finally, the social interaction theorists believe that language acquisition is impacted by many different factors including cognitive, linguistic, social, and physical ones (Reutzel & Cooter, 2007).

As reflected in many of these well-known theories, the process of language acquisition for all children, including ELLs, is influenced greatly by environmental factors including opportunities to learn and practice at home, at school, and in society. Research literature clearly links environmental factors, such as these rich learning opportunities, with proficiency in English. For example, the number of years students live in the U.S. interacting with native speakers of English (Hakuta, Butler, & Witt, 2000), the number of English-only classes, and student level of proficiency in native language affects students' proficiency in English, helping them to be reclassified as proficient in English. Such information, if collected properly, could improve the validity of the ELL classification system substantially. More importantly, to improve the validity of the classification system for ELL students, teachers, and school officials, including bilingual and ESL/ELD coordinators, must be familiar with students' language needs and backgrounds in order to use the criteria for the ELL classification system properly. Therefore, incorporating such important information into the ELL classification system will help improve the validity of this system. We now discuss ELP assessments that were developed before and after implementation of NCLB.

Pre-NCLB English Language Proficiency Assessments

The pre-NCLB assessments were developed by different organizations at different times based on different needs and requirements. There are major limitations with many of these assessments. First and foremost are the discrepancies in the theoretical bases of these tests. The tests are based on one or more of at least three different schools of thought: (1) the discrete point approach, (2) the integrative or holistic approach, and (3) the pragmatic language testing approach (Del Vecchio & Guerrero, 1995). Consequently, the tests provide very different outcome measures. For example, Valdes and Figueroa (1994) indicated that:

> As might be expected, instruments developed to assess the language proficiency of "bilingual" students borrowed directly from traditions of second and foreign language testing. Rather than integrative and pragmatic, these language assessments instruments tended to resemble discrete-point, paper-and-pencil tests administered orally. (p. 64)

Second, a distinction exists between basic interpersonal communication skills (BICS) and cognitive academic language proficiency (CALP) (Bailey & Butler, 2003; see also Cummins, 2000). In the context of assessments, language proficiency tests could vary in the extent they gauge CALP. Bailey and Butler (2003) defined academic language as "language that stands in contrast to the everyday informal speech that students use outside the classroom environment" (p. 9). In other words a student could score high in BICS but low in CALP. Therefore, it is necessary to determine that language proficiency tests adequately measure the type of language proficiency needed to be successful in mainstream English classrooms.

Zehler, Hopstock, Fleischman, and Greniuk (1994) compared English proficiency test content and structure (productive skills, receptive skills, and reading skills), the test administration procedures, the theoretical bases of the tests, and issues related to the validity and reliability of the tests. They found major differences in all of the areas in which the tests were compared. They also found that tests differed in their approaches to defining language proficiency, the types of tasks and specific item content, the grade level ranges, and the specific time limits (see also, Rossell, 2000).

Del Vecchio and Guerrero (1995) also presented a comprehensive review of some of the commonly used ELP tests prior to NCLB: (1) Basic Inventory of Natural Language (BINL), (2) Bilingual Syntax Measure (BSM), (3) Idea Proficiency Test (IPT), (4) Language Assessment Scales (LAS), and (5) Woodcock–Munoz Language Survey (WMLS). They found major differences between these tests with respect to their purpose, age and language group, administration, cost, items, scoring, test design, theoretical foundation, reliability, and validity of the tests. Such wide ranging disparities in these ELP assessments are a significant cause for concern with regard to the accuracy and consistency of the measures used to classify ELL students.

Post-NCLB English Language Proficiency (ELP) Assessments

NCLB required schools receiving Title I funding to annually assess ELL students' level of ELP using reliable and valid measures. For example, NCLB requires that ELP assessments include four modalities (reading, writing, speaking, and listening), incorporate the concept of academic language, and align the content of ELP assessments with the states' ELP standards.

Four consortia of states carried out the challenging task of developing post-NCLB assessments based on the NCLB Title III requirements (see Abedi, 2007). Test items were aligned with the states' ELP content standards and standard setting was conducted to set language proficiency levels in several categories typically distinguishing beginning, intermediate, proficient, and advanced in all four modalities. ELP tests were often developed for four or more grade clusters (typically K-2, 3–5, 6–8, and 9–12) and included common sets of items across adjacent grade clusters. The newly developed assessments underwent extensive pilot and field testing on large and representative samples of students. The content and psychometric properties of the individual items as well as the total tests were carefully examined and improvements were made where needed.

While these efforts have helped establish a strong foundation for the newly developed ELP assessments, we believe there are still some unresolved issues concerning such assessments

that may impact classification of ELL students. These issues include:

ELP standards

NCLB requires the newly developed ELP assessments to be aligned with state ELP content standards. This poses several concerns. The term *ELP* has not been clearly defined in the literature. Furthermore, many states did not have a set of defined ELP content standards prior to the implementation of NCLB, and it was technically challenging for the consortia to develop a set of standards that are truly common across the participating states in a given consortium of ELP assessment (see, for example, Fast, Ferrara, & Conrad, 2004).

Standard setting

The newly developed ELP assessment consortia conducted standard setting to create language proficiency levels for ELL classification purposes. In addition to the sources of inconsistencies due to the use of different standard-setting approaches and subjectivity involved in the standard-setting process (Giraud, Impara, & Plake, 2005; Jaeger, 1989), other factors may introduce bias into the standard-setting process. For example, inconsistencies in the proficiency levels at the different modalities may affect decisions on classifications of ELLs (see, for example, Bunch, 2006). To illustrate this point, assuming a student was rated as *proficient* in reading and writing but as *below proficient* in listening and speaking, how will this student be rated on the overall proficiency scale? Or, should the level of proficiency of this student be judged based on the total test of all four modalities? If so, then should there be evidence of unidimensionality of the test?

Dimensionality

In addition to the scores from each of these four modalities/subscales (reading, writing, listening, and speaking), composite scores of all subscales as well as other composites are commonly used by states. If the four modalities are highly correlated and if they measure a single construct, the decision to combine the different subscale scores would be less complicated than when the subscales are not measuring the same construct. Therefore, the issue of dimensionality needs to be addressed prior to deciding whether to use subscale or total scores.

The baseline for the NCLB Title III assessment

Since the newly developed ELP assessments were not available at the start of NCLB implementation, states had no other choice but to use whatever existing ELP assessment they found to be relevant for their state. Now that many states have access to new ELP assessments that meet NCLB requirements, they are faced with the quandary of linking ELP assessment results from "off-the-shelf" tests as the *baseline* with the results from their new ELP assessments. The problem is not limited to the tests having different domains of ELP content. Many of the existing ELP tests at the start of NCLB implementation were based on different theoretical emphases prevalent at the time of test development. They were also not aligned with states' ELP content standards and were not based on the concept of academic language. Therefore, even a high correlation between ELP assessments used as the baseline and the new ELP assessment would not be enough to assume a strong link between the two.

Academic English

Clearly, the focus of NCLB Title III ELP assessment is on "academic English." Therefore, many of the newly developed measures of ELP are based on academic English to facilitate learning content knowledge. However, concerns remain as to whether ELP assessment should be focused on the language of the content areas (such as mathematics and science) or the language that facilitates content learning. Fast et al. (2004) indicate that ELP assessments "are not tests of academic content, in other words, no external or prior content-related knowledge is required to respond to test questions" (p. 2). Given these concerns, test item writers for the ELP assessments are not quite certain what constitutes "*academic English*" and how it should be captured within the ELP assessments. (For a more detailed discussion of the pre- and post-NCLB ELP assessments see Abedi, 2007.)

Standardized Academic Achievement Tests

Results from standardized achievement tests which are required for all students are also used in conjunction with scores from language proficiency tests to identify but mainly to reclassify ELL students. The *Survey of the States' LEP Students* revealed that achievement tests were used by 76% of the states surveyed to help identify or reclassify ELL students (Kindler, 2002). Critics believe these tests are not designed for this purpose; rather, they are designed only to assess monolingual English students' content knowledge (Rossell, 2000).

Stefanakis (1998) indicated that a major concern in the assessment of ELL students is the lack of standardized achievement tests specifically designed to assess the content knowledge of these students. Mahoney and MacSwan (2005) argued that the use of academic tests for identifying and reclassifying ELL students is inappropriate. A review of standardized achievement tests by Zehler et al. (1994) found major differences between these tests across different areas, including their content, format, and psychometric characteristics.

Not only do the differences between tests produce inconsistent classification and assessment results for ELL students, but the unnecessary linguistic complexity of many achievement test items that are developed for native speakers of English casts doubt on the validity and reliability of these assessments when used for ELL students. For example, based on the results of many studies on the assessment of ELL students, Abedi (2006a) indicated that unnecessary linguistic complexity of test items may be an additional source of measurement error in using standardized achievement tests for ELL students (see also Figueroa, 1989, 1990; Valdes & Figueroa, 1994). This may seriously undermine the validity of inferences addressed by the assessment because it is a source of construct-irrelevant variance (see also Haladyna & Downing, 2004; Messick, 1994). Additionally, Solano-Flores and Trumbull (2003) found that language factors interact with test items. That is, items that are linguistically complex contribute largely to the measurement error variance observed for ELL students, leading students to misinterpret and misunderstand test questions.

In addition to the content and psychometric concerns with using standardized academic achievement tests as an index for ELL classification, there is disagreement on the level of student performance below which students are considered as ELL. For example, Grissom (2004) reported a cutoff score on the standardized norm-referenced test (NRT) at the 36th percentile point or

above in order for students in California to be RFEP. Gándara (2000) indicated that, "LEP in California is a child who does not understand sufficient English to pass a test of oral proficiency and *who does not score above the 35th to 40th percentile on an English standardized test*" (p. 3, emphasis added). Linquanti (2001) documented reclassification cutscores used in seven California districts that ranged from the 33rd percentile to the 40th percentile, with some utilizing these for reading, language, and/or math NRT sections. A report by the United States General Accounting Office (2001) observed a disagreement about appropriate standards for measuring English proficiency. The report interpreted the age/grade appropriate level as scoring above the 50th percentile on standardized achievement tests but also acknowledged that "some states consider students English proficient when they score at the 40th percentile or even at the 32nd percentile" (p. 14).

As can be seen from the short summary of research presented above, there is no specific indication of which tests or which cutoff score would indicate an acceptable level of English proficiency. Classifying language proficiency by arbitrarily setting a cutoff point on standardized academic achievement test scores (for example in reading/language arts) is also not a good practice since there are large numbers of native English speakers who score below these cutoff points. Should these students also be considered ELL? If the answer to this question is "Yes," then the concept and operational definition of ELL classification becomes even more controversial. On the other hand, if the answer is "No," then one must ask if low-scoring, native English speakers can truly be considered language proficient, classified as "non-ELL," and be deprived of the additional language skill development they deserve. Further, the language intervention strategies would need to be significantly different for these students than for those whose native language is not English.

Empirical Evidence on the Validity of ELL Classification

To examine the validity of current ELL classification, we present empirical data on some of the most commonly used criteria in the classification/reclassification of ELL students. Students' background characteristics, such as students' socioeconomic status (SES) and ethnicity are not directly used as criteria for ELL classifications, but because they have been found to correlate with students' academic performance, these variables may also need to be acknowledged as predictors of ELL classification outcome. Finally, we present some data to help us provide research-based recommendations on how to improve the validity of ELL classification.

Our presentations are based on data from several randomized field studies conducted at the National Center for Research on Evaluation, Standards, and Students Testing (CRESST, see for example, Abedi, 2006a, 2006b; Abedi et al., 2000; Staley, 2005) and analyses of existing data from seven locations nationwide. Table 1 presents summary information, including testing year, grade levels, student population, and type of test for the sites that provided comprehensive databases for analyses. Because of confidentiality agreements with the data providers, state and test names will not be mentioned in this paper but may be revealed by the written permission of the providers.

The sites were in the United States and varied in location and population. The student background variables included gender, ethnicity, free/reduced-price lunch participation, parent education, student ELL status, and Students with Disabilities (SD) status. Item-level standardized achievement test data were also obtained. However, the sites differed in the standardized tests used, the type of language proficiency index used, and the type of background variables provided. Comparisons of the results across the data sites provided cross-validation information. To obtain information on the consistency of results over time, we also included a few locations with very recent data. Of the seven sites, four provided assessment data from 1997 to 1998 (pre-NCLB) and the three others provided data for 2005–2006 (post-NCLB).

Empirical Data Presentation

Home Language Survey

Earlier in this paper, we cited studies that question the validity of the Home Language Survey, a commonly used

Table 1. Site Summary

Site/Grade	Data Year	Number of Students	Number of ELL Students	Percent of ELL Students	Tests Used
Site 1	1998–1999				NRT
Grade 3		36,065	7,270	20.20%	
Grade 6		28,313	3,341	11.80%	
Grade 8		25,406	2,306	9.00%	
Site 2	1997–1998				NRT
Grade 2		414,169	125,109	30.20%	
Grade 7		349,581	73,993	21.20%	
Grade 9		309,930	57,991	18.70%	
Site 3	1997–1998				NRT
Grade 10		12,919	431	3.30%	
Grade 11		9,803	339	3.50%	
Site 4	1997–1998				NRT
Grade 3		13,810	1,065	7.70%	
Grade 6		12,998	813	6.30%	
Grade 8		12,400	807	6.50%	
Site 5	2005–2006				CRT
Grade 5		33,242	5,008	15.10%	
Grade 8		33,106	3,870	11.70%	
Site 6	2005–2006				CRT
Grade 4		102,574	4,219	4.10%	
Grade 8		107,695	3,456	3.20%	
Site 7	2005–2006				CRT
Grade 4		55,724	7,090	12.70%	
Grade 8		52,900	3,026	5.70%	

Note: Data on ELL students from Site 1 is for students receiving bilingual services.

criterion for ELL classification decisions. The findings of our empirical data share similar concerns. In one study, Abedi et al. (1997) formulated the Language Background Questionnaire (LBQ) based on the HLS concept and administered it to 1,031 eighth-grade students. The LBQ included questions about languages other than English spoken at home, the number of years the student had lived in the United States, and the number of English-only classes taken. Students' responses to the LBQ were compared with school rosters reporting the students' official primary languages as identified by the parents on the district's Home Language Survey and their ELL classification where appropriate. Significant discrepancies were revealed, making the accuracy of this single source of language background data highly questionable. In many schools the record of students speaking a language other than English at home, regardless of ELL classification, was significantly lower than what the students reported in the LBQ (see also Abedi et al., 2000).

Language Proficiency Test Scores in Determining ELL Classification

Since ELL classification should distinguish between students who are proficient in English and those who are not, one would assume a high level of association between these two variables. However, research findings do not support this assumption. The lack of a strong relationship between English proficiency test scores and ELL classification may be partly due to content and psychometric shortcomings of the tests but are mainly due to validity issues in defining the ELL/non-ELL dichotomy.

To illustrate the relationship between ELL classification codes and ELP scores, we selected two data sites: one provided data from ELP tests that were developed prior to NCLB (Site 2) and one provided data from the ELP test developed based on the NCLB Title III requirements (Site 7). A comparison of the relationship indices between these two sites may reveal some information on the possible improvement of ELL classification based on the new ELP assessments.

One of the most commonly used ELP tests developed prior to the implementation of NCLB (Loop, 2002) was used for this purpose. A district within Site 2 provided an excellent opportunity to illustrate the power of this test in defining the ELL/non-ELL dichotomy. The test was administered to both ELL and non-ELL students. Performance differences between ELL and non-ELL students were estimated in terms of effect sizes (Cohen, 1988; Kirk, 1995). Additionally, the percent of variance of ELL classification categories explained by the test was also computed and was labeled as ω^2 (see Kirk, 1995, pp. 177–180). Table 2 presents ω^2 as well as effect sizes when comparing the performance of ELL and non-ELL students in grades two through twelve for a district in Site 2. The proportions of variance of ELL/non-ELL explained by the test scores (ω^2) ranged between .03 (3% of the variance for Grade 12) to .09 (9% of the variance for Grade 10), which were not large enough to suggest a strong association between English proficiency test scores and ELL classification. The effect sizes ranged between .179 (for students in Grade 12) to .319 (for students in Grade 10) with an average effect size of .239. Based on Kirk (1995) this average effect size is considered small. Thus, the results do not support the notion that the ELP test score explains much of the variance of the ELL/non-ELL dichotomy.

To compare the association between the pre- and post-NCLB assessments and the ELL dichotomy, we also used data from one of the more recent test administrations. Site 7 administered one of the post-NCLB English language proficiency assessments to both ELL and non-ELL students. At this site the non-ELL students consisted of ELL students who were reclassified as fluent English proficient (FEP). Table 2 presents the results of analyses for Site 7 as well. Unlike Site 2 for which data were available for Grades 2 through 12, for Site 7 we had access to data only for Grades 4 and 8.

As data in Table 2 show, for Grade 4 the proportions of variance of ELL/non-ELL dichotomy explained by the newly developed ELP scores (ω^2) is .142 with an effect size of .407 ($n = 7{,}957$). Comparing with the similar statistics obtained for Site 2 ($\omega^2 = .035$ with an effect size of .190), the level of the relationship between the new ELP and ELL classification code is much stronger with the new ELP assessments. A similar association was found for students in Grade 8. The proportion of variance in ELL dichotomy explained by the new ELP measure was .104 with an effect size of .341, compared with a ω^2 of .064 and an effect size of .260 for Site 2. The average effect sizes of the association between ELP assessments and ELL dichotomy across the two grades for the new test is .374, which according to Kirk (1995) is medium to high as compared with an average effect size of .239 for pre-NCLB which is considered small.

Standardized Academic Achievement Test Scores in Determining ELL Classification

Academic achievement tests were used by 76% of the states surveyed in defining the ELL/non-ELL dichotomy (Kindler, 2002). As noted above, research has identified major sources of construct-irrelevant variance with these tests when administered to ELL students. To illustrate the power of achievement

Table 2. Omega-Square, Effect Size, and Number of Students for Sites 2 and 7

Site 2				Site 7			
Grade	Omega Square	Effect Size	Number of Students	Grade	Omega Square	Effect Size	Number of Students
2	.050	.229	587				
3	.038	.199	721				
4	.035	.190	621	4	.142	.407	7,957
5	.040	.203	1,002				
6	.050	.230	803				
7	.068	.270	938				
8	.064	.260	796	8	.104	.341	4,364
9	.070	.275	1,102				
10	.092	.319	945				
11	.074	.283	782				
12	.031	.179	836				

Table 3. Site 1 and 6 Omega-Square, Effect Sizes, and Number of Students for NRT and CRT Test Scores and ELL Classification

	Site 1			Site 6	
	Reading	Math		Reading	Math
Grade 3			Grade 4		
Omega	.026	.002	Omega	.017	.034
Effect size	.162	.045	Effect size	.132	.188
Number of students	36,006	35,981	Number of students	100,992	101,652
Grade 6					
Omega	.066	.024			
Effect size	.265	.156			
Number of students	28,272	28,273			
Grade 8			Grade 8		
Omega	.067	.028	Omega	.003	.013
Effect size	.266	.170	Effect size	.058	.115
Number of students	25,362	25,336	Number of students	106,700	107,016

test scores in determining the ELL/non-ELL dichotomy, we compared the performance of ELL and non-ELL students on reading/language arts, science and mathematics subscale scores of achievement tests using data from sites 1, 2, 5, and 6. Norm-referenced tests (NRTs) in reading, science, and mathematics were used in sites 1 and 2, and state constructed criterion-referenced tests (CRTs) in reading and mathematics were used in Site 5 and 6.

Once again, performance differences between ELL and non-ELL students on the scores of achievement tests were estimated in terms of percent of variance explained and effect sizes. Table 3 presents the effect sizes and the proportion of the variance of the ELL classification code explained based on test scores for students in Grades 3, 6, and 8 in Site 1 and students in Grades 4 and 8 in Site 6. The two sites (Site 1 with the pre-NCLB data and Site 6 with the post-NCLB data) will provide the opportunity to make comparisons over time and to determine the possible impact of NCLB on the classification of ELL students.

As data in Table 3 show, ω^2 and effect sizes were very small for both sites, suggesting that there is not a strong association between standardized achievement test scores and ELL/non-ELL dichotomy. The ω^2 for the NRT test by ELL classification ranged between .002 (mathematics) for Grade 3 to .067 (reading for Grade 8), indicating that based on the average of the three grade levels (3, 6, and 8), only about 3.5% of the variance of ELL classification is explained by the NRT scores. Effect sizes for NRT across all grade levels ranged from .045 (mathematics for Grade 3) to .266 (reading for Grade 8). Once again, these effect sizes are small (Kirk, 1995).

Results from Site 6 are consistent with those reported for data in Site 1. The ω^2 values ranged from .003 (Grade 8 reading) to .034 (Grade 4 mathematics), which on the average explain about 1.7% of the common variance between test scores and ELL classification codes. The effect sizes ranged from .058 (Grade 8 reading) to .188 (Grade 4 mathematics) with an average of .123 which is quite small (Kirk, 1995).

Table 4 presents ω^2 and effect sizes for NRT and CRT tests in explaining the ELL/non-ELL dichotomy for students in Grades 3, 7, and 9 in reading, science, and mathematics in Sites 2 and for Grades 5 and 8 in Site 5. Site 2 used the NRT test scores directly as a criterion for ELL classification. Therefore, one would expect a higher level of association between the achievement test scores and ELL classification in this site. The index of strength of association (ω^2) for the NRT test ranged from .051 (Grade 9 mathematics) to .203 (Grade 7 reading) with an average of .115, indicating that NRT test scores explained about 12% of the variance of ELL classification. The effect sizes for the NRT test scores ranged from .231 for mathematics in Grade 9 to .504 in

Table 4. Site 2 and Site 5, Omega-square, Effect Sizes, and Number of Students in Different Subscales and ELL Classification

	Site 2			Site 5		
	Reading	Science	Math	Reading	Science	Math
Grade 3				Grade 5		
Omega-Square	.172	.089	.076	.108	.091	.068
Effect size	.456	.313	.286	.348	.316	.270
ELL	104,333	23,555	109,327	5,008	5,008	5,008
Non-ELL	272,653	54,300	277,042	28,118	28,118	28,118
Total	376,986	77,855	386,369	33,126	33,126	33,126
Grade 7						
Omega-Square	.203	.132	.101			
Effect size	.504	.390	.335			
ELL	69,074	24,761	71,227			
Non-ELL	267,235	77,834	268,867			
Total	336,309	102,595	340,094			
Grade 9				Grade 8		
Omega-Square	.150	.087	.051	.068	.061	.057
Effect size	.420	.309	.231	.270	.255	.246
ELL	32,515	33,032	33,311	3,870	3,870	3,870
Non-ELL	192,598	192,639	193,906	33,106	33,106	33,106
Total	225,113	225,671	227,217	36,976	36,976	36,976

reading for Grade 7. These effect sizes fall within the mid-category based on Kirk (1995).

The strength of association (ω^2) between the CRT test scores and the ELL classification code was also obtained for data from Site 5. The results of analyses summarized in Table 4 show a slightly lower level of association between the CRT test scores and ELL classification code. The index of strength of association (ω^2) ranged between .057 for Grade 8 mathematics to .108 for reading in Grade 5 with an average of .076, indicating that only about 7.6% of the variance in the ELL classification code is explained by the scores of standardized achievement tests. Similarly, the effect sizes ranged between .246 (Grade 8 mathematics) to .348 (Grade 5 reading) with an average of .284 which is considered a medium effect size.

The data presented above clearly suggest that standardized achievement test scores, including reading and language arts subscales, were not strongly associated with the levels of ELL classification. More importantly, the post-NCLB assessments did not show much improvement over the pre-NCLB assessments in providing more valid criteria for ELL classification. These data suggest that standardized achievement tests may not be a valid criterion for assessing ELL students for classification purposes as a single criterion or even when combined with other criteria.

Discrepancies in Patterns of ELL Classification in Different Districts

Validity issues and inconsistencies in ELL classification criteria have resulted in large discrepancies between states and districts in ELL classification/reclassification practices. These discrepancies will continue to persist until serious consideration is given to the validity of ELL classification. As indicated earlier in this report, standardized achievement test scores are often used as a criterion for classification of ELL students and their reclassification as "Redesignated Fluent English Proficient" (RFEP) or a similar code. To be reclassified as RFEP at Site 2 of this study, ELL students had to score above the 36th percentile on the NRT reading comprehension test, with some discretion allowed. To study the implementation of this "36th percentile policy," we compared agreement between current classification and performance on the standardized reading test.

Table 5 provides data on the agreement between the NRT reading levels and ELL classifications of students in Site 2 for some ELL students with valid scores in the reading content area. Percentages in the table represent agreement between the actual classification and the reading percentiles. As Table 5 shows, 74.5% of students scoring below the 36th percentile were designated as ELL. Of the students scoring above the 36th percentile, 73.8% were reclassified RFEP and 26.2% were still classified ELL (contingency coefficient = .390, $p = .000$). We understand that standardized achievement test scores are not often used as a single criterion but these results question the usefulness of these scores even when they are used in conjunction with other criteria.

Figure 1 shows the variation in percentages of ELL students scoring below the 36th percentile in reading who remain classified as ELL in districts with at least 200 third-grade ELL students. Each "line" in the figure represents one district. Low-scoring ELL students in grades three through five scoring below the 36th percentile tended to remain classified as ELL, while low-scoring students in higher grades were more likely to be reclassified as RFEP. As grade level increased, however, the variation in agreement among districts also increased. Parrish, Perez, Merickel, and Linquanti (2006) found that some districts used alternative reclassification criteria that lowered reclassification standards for ELL students at later grades. They cited concerns among administrators that long-term ELLs could face aggregated track placement and reduced access to courses needed for postsecondary education.

To measure the overall agreement between the current classification and the next NRT reading performance we computed *kappa* coefficient which measures exact agreement beyond chance. Figure 2 presents the *kappa* for the Site 2 districts. As Figure 2 shows, there was a wide variation in overall agreement among these districts at Site 2. Two districts were very close to the kappa = 0 line that indicates no agreement beyond chance between ELL classification and NRT scores. In comparison, three districts had kappa > .50 in middle school.

These discrepancies once again point to the validity concerns in ELL classification/reclassification practices. The districts used in the above analyses were all using the same standardized achievement test. If the 36th percentile score is a good indication of students' relative proficiency in reading comprehension (the policy set by the state), then the irony is that over 25% of the students scoring above the 36th percentile are still classified as ELL. Obviously, one can expect an even larger discrepancy across different states using different standardized achievement tests based on different content standards, with different percentile score cutscore.

Variables/Factors Unrelated to Level of English Proficiency That Might Have Long-Term Effects on Reclassification

Results of the research presented above suggest that students' level of English proficiency is not the only determinant of ELL classification. Other factors may also influence decisions on ELL classification. Grissom (2004) and Abedi (2004) both found that variables such as gender, socioeconomic status (SES, measured by free/reduced price lunch), ethnicity, and parent education are powerful predictors of ELL classification/reclassification. For example, Grissom (2004) indicated that in California multiple criteria are used for reclassifying from ELL to RFEP. These criteria according to Grissom include: (1) assessment of ELP, (2) teacher evaluation, (3) parent opinion and consultation,

Table 5. Site 2, Agreement Between NRT Reading Levels and ELL Classification

	Current Classification	
NRT Reading Level	**ELL**	**RFEP**
Below 36th percentile (*n*)	56,095	6,040
Below 36th percentile (Pct)	74.5%	26.2%
Above 36th percentile (*n*)	19,172	16,983
Above 36th percentile (Pct)	25.5%	73.8%

Note: Contingency coefficient = .390.

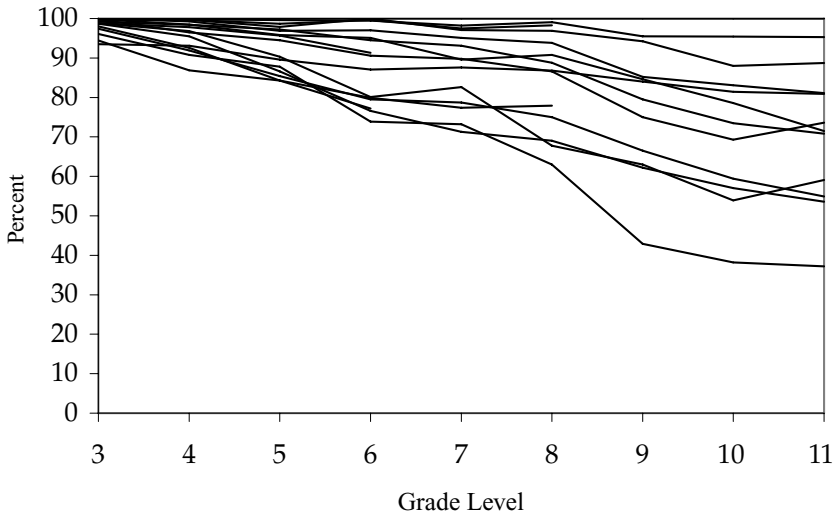


FIGURE 1. Site 2 percent of ELL students scoring below the reading 36th percentile by school district (minimum $N = 200$).

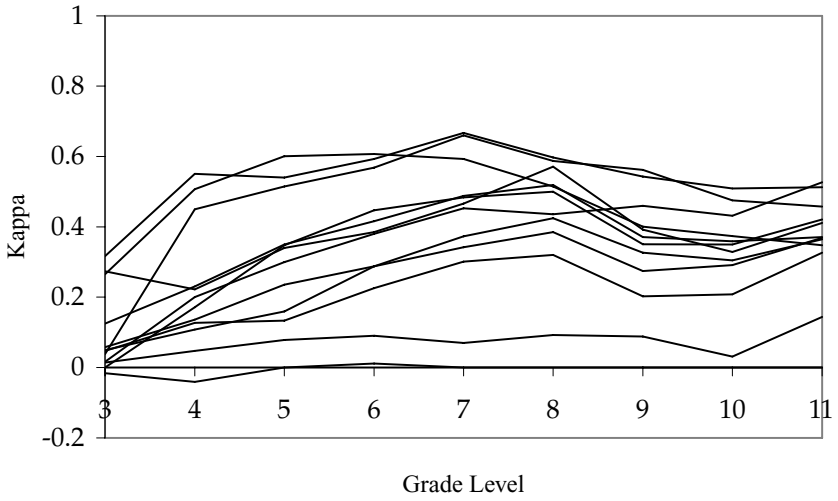


FIGURE 2. Site 2 *kappa* coefficients for agreement between ELL classification and NRT performance (minimum $N = 200$).

and (4) performance in basic skills. Moreover, Parrish et al. (2006), in their examination of nine high- versus low-reclassifying districts in California, noted that variations in the districts' reclassification criteria and cutscore, procedures and systems to monitor students' readiness to reclassify, and the importance of reclassification in local accountability explained much of the observed variation in reclassification rates.

A large district within Site 2 provided a unique opportunity to study the effects of some of these variables on ELL classification/reclassification decisions in a longitudinal setting. We created a cohort of 1993–1994 Grade 7 students (n = 23,856) and followed them for a period of 6 years (12 semesters, fall 1993 to spring 1999), conducting an event history analysis (also referred to as a *survival analysis* approach, Miller & Zhou, 1997).

Results of the event history analysis indicated that, in addition to the students' level of language proficiency, their background variables (such as ethnicity) appear to correlate with the ELL classification. Table 6 summarizes the results of the event history analyses by students' background variables and test scores. Results are reported by gender, ethnicity, free/reduced lunch participation, Title I status, and reading test scores. Table 6 shows the number of ELL students at the start of the cohort and the proportion of redesignated ELL students for the first period (first six semesters) and the second period (second six semesters). The last column of Table 6 shows the median semesters students remained as ELL.

The data show that the percentages of students that were RFEP from the first to the sixth semester vary considerably across categories of some variables. Between males and females, there was not much variation. In the time period between the first and sixth semesters, 27% of female students were redesignated compared to 23% of male students. In the second period, 62% of female students as compared with 53% of male students were redesignated. This difference was not statistically significant. Consequently, the median time in ELL status (number of semesters) was very similar for males and females (9.08 for females and 9.98 for males). Likewise, percentages of RFEP across the free/reduced lunch program were very similar (9.52 for participants versus 9.55 for nonparticipants).

In contrast to gender and SES, there was a much larger variation in percentages and median time spent in the ELL category across racial/ethnic categories. Percentages of RFEP for the first period across ethnic categories ranged between 21% for Hispanics to 55% for Asians and Caucasians. The percentages of RFEP in the second time period were substantially higher for all ethnic categories as expected. However, once again Hispanics had the smallest percent of RFEP (57% Hispanics versus 77% Asians and 68% Caucasians).

It took almost ten semesters for Hispanic students to be reclassified from ELL to RFEP, while it took half as much time for Asian and Caucasian students to be reclassified. Looking at the levels of reading test scores, it took much less time for students with higher reading scores to be reclassified than students with lower reading test scores as one might expect. However, many of these variables may be confounded with other variables. For example, a majority of students who are classified as ELL are Hispanic.

Initiating Changes

The results of analyses summarized above indicate that the current system of ELL classification produces inconsistent outcomes. This may be due to psychometric characteristics of the assessments used for the classification, the resulting accuracy of ELL classification, the weight of other factors influencing

Table 6. Site 2, 1993–1994 Cohort ELL Event History Analysis by Student Background Variables

Subgroup	Number ELL at Start	Percent RFEP Within 0–6 years	Percent RFEP Within 7–12 years	Median Time in ELL Status
Gender				
Female	11,763	27%	62%	9.08
Male	12,093	23%	53%	9.98
Ethnicity				
Asian	1,246	55%	77%	5.45
Hispanic	21,167	21%	57%	9.91
Caucasian	1,019	55%	68%	5.49
Free Lunch				
Free/reduced	19,099	24%	59%	9.52
Nonparticipant	4,757	29%	50%	9.55
Title I				
Title I	14,166	18%	56%	10.13
Non-Title I	9,690	34%	60%	8.45
93–94 NRT Reading Test				
Not tested	10,465	18%	38%	>12.00
Percentile 1–15	3,835	08%	42%	>12.00
Percentile 16–36	5,704	23%	75%	8.83
Percentile 37+	3,852	59%	94%	5.07
96–97 NRT Reading Test				
Not Tested	10,980	17%	38%	>12.00
Percentile 1–15	6,141	15%	52%	10.80
Percentile 16–36	3,747	32%	84%	7.89
Percentile 37+	2,988	58%	94%	5.13

Note: Students with missing data were deleted; therefore, totals are slightly different across different groups.

ELL classification, or to a combination of the above. More importantly, a valid classification system should be based on the theory of language acquisition and should clearly identify the level of academic language proficiency that students should reach in order to be classified as fluent in English and to be able to fully participate in English-only instruction and assessments.

It must be indicated at this point that the term "ELL classification system" that is used frequently in this paper includes initial identification as language minority (via home language survey), initial identification as English learner (typically via ELP assessments) and reclassification to fluent English proficient (typically via both ELP and achievement test scores). While we distinguish between these different purposes, the ELL classification system as discussed in this paper encompasses all these purposes.

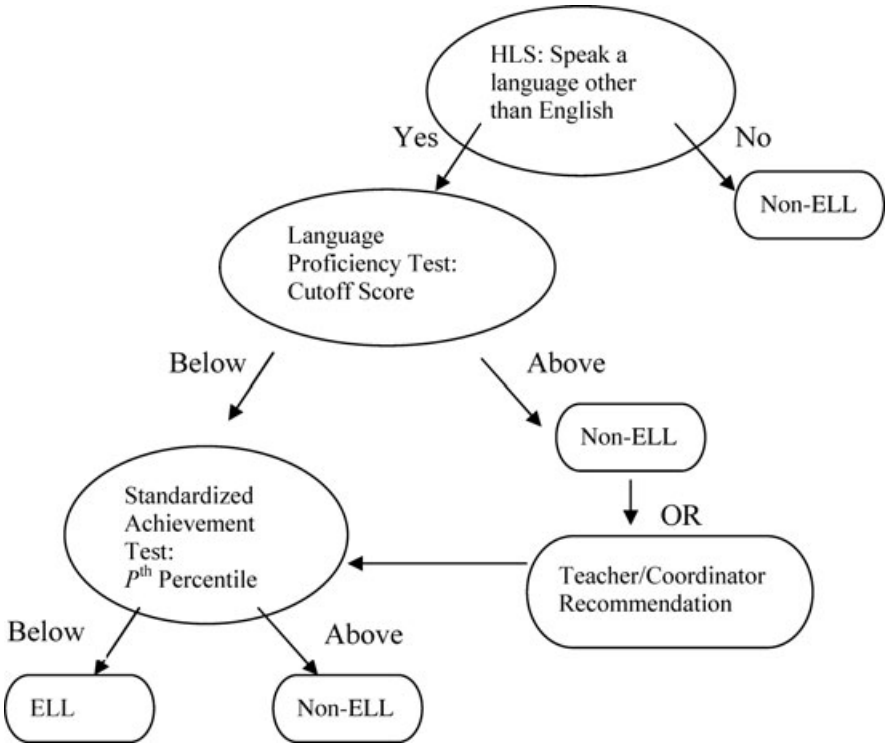
To initiate dialog on creating a model for improving validity of the ELL classification system, we support the use of multiple criteria (Tippeconnic & Faircloth, 2002) with a minimum level of redundancy in a value-added sequence of phases. We build this model on the concept of *academic English* since beyond such a practical concept there is not a commonly acceptable theoretical foundation for ELL classification. We believe the concept of academic English is a sensible base for the model for two reasons: (1) ELL student proficiency in academic English is a prerequisite of their success when both instruction and assessment are offered in English only, and (2) as noted in this paper, measures of students' academic performance are also used as criteria for ELL classification/reclassification.

Figure 3 depicts this stepwise concept. The model uses information currently available and recommends increasing the validity of the classification system by augmenting the knowledge about the student language background with multiple criteria. Obviously, as discussed earlier, the existing data that are used in this model (HLS, ELP, and achievement test scores) may have serious limitations at least in two ways: (1) they do not include the major variables that the literature suggests for increasing the validity of ELL classification (such as number of years in the United States, number of English only classes ELL students take and their proficiency in their native language), and (2) there are serious technical flaws in many of these assessments. However, we want to demonstrate that some improvements to the ELL classification system can be made even under such less than desirable conditions.

The process starts with data from the HLS as one of the most commonly used criteria for identifying linguistic minority status and establishing the need to assess for possible ELL classification. This will establish an initial potential ELL cohort that will then be augmented by additional criteria such as English proficiency and academic achievement test scores. A potential shortcoming at this first level of augmentation, however, is under- or over-identification of potential ELL students based on the HLS data. As an example of over-identification, a native English speaker who does not speak another language may be placed in the potential ELL cohort simply because a family member living with the child sometimes speaks a language other than English. This issue can be avoided by categorizing them as linguistic minority until the ELP assessment identifies them as ELL. On the other hand, under-identification may occur when parents declare the child as English-only because they speak only English at home for practice.

For lack of a better term, we call this process the *Augmented-Classification* (AC) approach. As Figure 3 shows, AC starts with the information from the Home Language Survey. All students who are identified as born outside the United States or who speak a language other than or in addition to English at home will establish the initial potential cohort of language minority students. Information from the next levels of the augmentation (ELP and achievement test scores) will help identify students who are not ELL, despite not being born in the United States and speaking a language other than English at home.

In using test scores (English language proficiency and achievement test scores) as the next levels of augmentation, we need to distinguish



Note: Non-ELL includes FEP and RFEP.

FIGURE 3. Diagram of the Augmented-Classification approach.

between norm-referenced test (NRT) and criterion-referenced test (CRT) data as the two types of data may lead to different classification outcomes. The NRT-based classification system was more common prior to the implementation of NCLB. After NCLB and particularly after development of the new generation of ELP assessments, the CRT-based system was more commonly used. However, since there are still some NRT-based classification practices, we present our proposed AC system based on both alternatives (NRT and CRT).

NRT-Based Augmentation

At this level of augmentation under the NRT-based model, percentile scores are often used. The major flaw with the NRT-based augmentation is the inconsistencies in the cutscore based on which the decision is made to classify/reclassify a student as ELL or non-ELL. Different states adopt different cutscores for identifying their ELL students based on their test scores.

ELP percentile scores are used as the criterion for the second augmentation level. Students who are at or above a given percentile point (to be determined by states based on empirical data) on the ELP tests may exit the ELL cohort. The remaining students in the cohort can then be considered ELL with a higher level of confidence than those identified as ELL based on HLS data alone with teacher's discretion to add an achievement test criterion as another layer of confidence. However, there might be a slight possibility that ELP assessments may have underidentified some students. To control for the possibility of underidentification, achievement test scores can be used as the third level of augmentation. While it may seem unproductive to test ELL students at the lower level of English proficiency in content-based assessment in English, there is evidence that points to such practices. For example, based on the data presented in Table 5, of the total 75,267 ELL students for whom we had score in English language arts, 19,172 or 25.5% of them scored above the 36th percentile on the NRT Reading. Based on these data used as the third level of augmentation, these students can also be reclassified as RFEP.

In the *AC* approach, one can use single or multiple measures at each stage of augmentation. Districts and schools may have multiple measures of English content and multiple measures of English proficiency. To test the improvement level in the validity of ELL classification using multiple criteria at each stage, data from a group of 916 ELL/non-ELL third-grade students were used (Staley, 2005). These students were from a single district within Site 2. For each student the data included four measures of ELP, three standardized English language arts achievement measures, and six standardized mathematics achievement measures. Of the 916 students, 602 had complete data on all the measures used in the analyses.

Composite scores of the four English proficiency scores and three English language scores were created using both latent composite and simple composite approaches. In the latent composite approach, we created a single-factor confirmatory factor model of the individual measures and used the factor score as a latent composite. The simple composite score was computed by converting the measures to standard Normal Curve Equivalent (NCE) scores with a mean of 50 and a standard deviation of 21.06 (Linn & Gronlund, 1995) and then averaging over all the individual measures.

We applied the proposed *AC* approach to the data discussed above by starting with the ELL cohort identified by the district according to their classification policy. Table 7 summarizes the results of these analyses. As the data in Table 7 show, of the 602 students with complete data, 309 or 51.3% were classified as ELL and 293 or 48.7% as non-ELL by their schools.

In the second phase of augmentation, we used the latent composite of ELP

Table 7. Added Classification Power at Different Augmentation Phases

ELL Status Defined	No. of ELL	No. of Non-ELL	% Moved from ELL to Non-ELL	ω^2 English Measures (Effect Size)
By school	309	293	0	.248 (.57)
Based on English proficiency	182	420	41.1	.325 (.70)
Based on English measures	117	485	35.7	.411 (.84)

measures with a cutscore at the median of the English proficiency score distribution. As the data in Table 7 show, of the 309 students who were identified as ELL based on their ELP scores, 127 or 41.1% had ELP test scores above the median of the distribution and thus were reclassified as non-ELL. In this phase, the strength of association (ω^2) was increased from .248 to .325 (an increase of about 8% on the prediction power).

In the next phase of augmentation, we applied the 36th percentile policy on the composite score of NRT English subscales (comprehension, vocabulary, and language). Of the 182 students who were classified as ELL by their school, 65 or 35.7% had scores above the 36th percentile. We reclassified these students as non-ELL. Thus, the AC model improved the strength of association between the ELL classification code and the criteria used for such classification from .248 (24.8% of the variance of ELL classification explained) to .411 (41.1% of the variance) using existing data. We believe more substantial improvements can be observed if the validity of the current criteria for ELL classification is increased. The results of our analyses based on the post-NCLB English language proficiency tests were consistent with the pre-NCLB analyses showing an even higher trend of improvements in the validity of the ELL classification system due to better ELP assessment quality. With the post-NCLB data, we could improve the strength of association between the ELL classification code and the criteria used for classification from 26.4% to 49.6%.

CRT-Based Augmentation

The proposed CRT-Based augmentation model is very similar to the NRT-based model except that in the CRT-based model achievement levels rather than percentile scores are used. Similar to the NRT-based model, this model uses three levels of augmentation. The first level uses data from HLS, the second level is based on ELP test scores (language proficiency levels) and the third level is based on academic achievement test scores (achievement levels). Since the CRT-based classification system has been implemented only recently, there is not enough data to test its effectiveness over the traditional classification system. However, as discussed earlier, the newly developed ELP assessments show higher power in discriminating ELL students at different levels of proficiency in English; therefore, we expect more improvements in the ELL classification system using the new ELP assessments.

There are several potential risks to the validity of the CRT-based classification system. For example, while the collaboration between states in the form of consortium reduced the variation in the ELP assessment outcomes, different states still continue using different ELP assessments. Additionally, differences in the standard setting procedures by states may also create discrepancies in ELL classification system across the nation.

We propose the AC model for two reasons: first, to convey the fact that there are several major concerns with the current ELL classification system that necessitate an urgent remedy and, second, to demonstrate that even the existing data with the limitations and validity issues can be used to provide a more reasonable means of ELL classification. We acknowledge the limitations of the AC model, but we hope that by demonstrating how small steps such as creating a model based on multiple criteria could potentially *improve* the ELL classification system, we can send a message that a more valid ELL classification system is not as far out of reach as many believe. We understand that states may vary in their application of criteria for making decisions about exiting students from ELL status. The AC model could actually help with making such decisions more consistent across and within states.

Note, however, that even the best designed ELL classification system with the most valid criteria may not produce valid outcomes if teachers and school officials (including the bilingual coordinators) are not knowledgeable about assessment and classification systems for these students. To be successful in this area, teachers must know about ELL students, their background characteristics and how their educational needs might be different from the native speakers of English.

Discussion

Because of inherent background differences between ELL and non-ELL students in terms of ELP, the universal use of curricula and assessments designed for non-ELL students may lead to inappropriate instruction and create invalid inferences about ELL academic achievement. The most important prerequisite to providing appropriate instruction and fair and valid assessment for ELL students is to correctly identify them. Inappropriate classification decisions may place students who are at a higher level of English proficiency into remedial or special education programs and may deprive less-proficient students of appropriate curriculum and assessment. Poor placement decisions may affect promotion and graduation, which consequently affects students' academic progress and self-esteem. Misclassification of ELLs may also impact school, district and state accountability systems resulting in negative repercussions. Delay in the reclassification of students who have reached English proficiency may deny them the opportunity to achieve and may reduce access to courses needed for post-secondary education, while premature reclassification may cause ELL students to lose needed specialized academic language instructional services and be placed at greater risk for educational failure (see Linquanti, 2001; Parrish et al., 2006).

Education and assessment communities have raised concerns over the validity of the current ELL classification system. Lack of a strong theoretical foundation and issues regarding the quality of criteria used for such classification are among these concerns. This paper presented empirical evidence substantiating such concerns and initiated discussion on improving the validity of the classification system. Research findings presented in this paper point to the fact that a remedy for this complex problem is urgently needed if ELL students are truly not to be left behind.

As indicated above, the post-NCLB English language proficiency tests improved the validity of the ELL classification system but this trend was not observed in the post-NCLB academic achievement tests. There are at least two possible explanations for this finding. It might be that the performance-gap between ELL and non-ELL students was reduced due to the positive impact of NCLB; therefore, the achievement tests did not show much power in differentiating the two groups in terms of their content-based performance. More likely, this lack of association might be due to ELL classification issues, i.e., large heterogeneity in the ELL population which suppresses the size of performance difference between ELL and

non-ELL students. This finding suggests that states are still facing major challenges in providing more accessible assessments for ELL students.

This paper also cited other contributing factors to inconsistent classification, some of which are unrelated to a student's level of English proficiency, such as ethnicity and schools' Title I status. The impact of such variables on ELL classification decisions may explain large discrepancies in ELL classification within and between states across the nation. While some of these powerful predictors such as ethnicity are not modifiable, knowledge of their impact can inform decisions on a student's ELL status.

The results of studies presented in this paper also raised concerns over the reclassification trends and policies. For example, results indicate that low-scoring students in lower grades (e.g., Grades 3 through 5) tended to remain classified as ELL, while low-scoring students in higher grades were more likely to be reclassified as proficient. Parrish et al. (2006) found that some districts are using alternative, lower reclassification criteria for ELL students in higher grades. If other independent studies confirm this trend, then investigating the potential causes of this trend is important to provide further insight into ELL classification and reclassification practices.

This is a complex situation and a simple solution may not work. Adding more tests to the states already burdened by testing requirements may not be realistic. On the other hand, using existing data as they are used currently for ELL classification may not produce valid and reliable classification outcomes. Thus, the main question is whether the validity of the ELL classification system can be improved using current information with a reasonable level of effort in enhancing the quality of such information.

To initiate a dialog among researchers, educational policymakers, and practitioners, we proposed a model that utilizes the assessment data available from different sources in the state assessment system. The idea is to augment our knowledge of students' English proficiency levels using information from different sources with a minimal level of redundancy.

The model proposed in this paper is tentative and conditional upon improved quality of ELP and standardized achievement tests for ELL students and including other relevant variables such as student's proficiency in L1 and number of years in the United States. For example, academic achievement test scores would help to improve the quality of the ELL classification system when they are more accessible for ELL students linguistically and culturally. Furthermore, the effectiveness of the proposed model is determined by rigorous validation studies that would need to be conducted nationwide. For example the percentile cutscores are set by many states arbitrarily which may not be based on much empirical evidence. Furthermore, some native English speakers score below these cutscores. How should they be treated? More research may be needed to establish the validity of such cutscores nationally.

More importantly, a valid classification system should be based on the theory of second language acquisition and should clearly identify the level of academic language proficiency that is needed for ELL students to function in academic environments where both instruction and assessment are offered only in English. Therefore, we built this model on the concept of *academic English* since beyond such a practical concept there is not a commonly acceptable theoretical foundation for ELL classification.

Improving the validity of the classification system requires both valid criteria and people who are knowledgeable about assessment and classification systems for ELL students to implement the system. The best and the most comprehensive system of ELL classification may not produce desirable outcomes if the implementation phase is not done properly. Therefore, it is imperative for those who are involved in the classification of ELL students to receive proper training and education about these students.

Unfortunately, ELLs are more likely to be taught by teachers without much knowledge on issues concerning classification and assessment of these students and with less classroom experience than teachers of other students (Rumberger & Gándara, 2004; Gándara et al., 2005). Thus, it is clear that not only issues concerning validity of the criteria used in the ELL classification system could create questionable outcomes; problems in the implementation of the classification system also contribute greatly to inconsistencies and problems in the ELL classification system.

We hope this discussion will initiate a national effort in establishing a valid and reliable ELL classification system. Though much legislation mandates equal educational opportunities for every child—including ELL students—very little can be done to improve the academic life of ELL students unless they are validly identified.

Acknowledgments

The work reported here was supported in part under a grant from the U.S. Department of Education, Institute for Education Sciences. The findings and opinions expressed here do not necessarily reflect the position or policies of the Institute for Education Sciences, or the U.S. Department of Education. The author acknowledges the contribution of several people. Robert Linquanti contributed to the quality of this paper by providing excellent comments and suggestions. Rita Pope and Cassandra Hawley assisted in structuring and revising the paper. Seth Leon provided valuable assistance with the data analyses.

Note

[1]The author acknowledges the terms "English language learner (ELL)" or "English learner" (EL) as alternatives to "limited English proficient (LEP)." All refer to students who may be in need of English language instruction, which encompasses a wide range of learners, including students whose first language is not English, students who are just beginning to learn English, and students who are proficient in English but may need additional assistance in social or academic situations (LaCelle-Peterson & Rivera, 1994). "English language learner" has been used as a more positive alternative to the term "LEP," which some regard has having a negative connotation (August & Hakuta, 1998). In this report, we use the term ELL more often since it is more commonly used in research and practice.

References

Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher*, *33*, 4–14.

Abedi, J. (2006a). Language issues in item-development. In S. M. Downing & T. M.

Haladyna (Eds.), *Handbook of test development*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Abedi, J. (2006b). Psychometric issues in the ELL assessment and special education eligibility. *Teacher's College Record*, *108*(11), 2282–2303.

Abedi, J. (Ed.) (2007). *English language proficiency assessment in the nation: Current status and future practice.* Davis: University of California.

Abedi, J., & Gandara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practice*, *26*(5), 36–46.

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.

Abedi, J., Lord, C., & Plummer, J. (1997). *Language background as a variable in NAEP mathematics performance* (CSE Tech. Rep. No. 429). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

August, D., & Hakuta, K. (Eds.) (1998). *Educating language-minority children*. Washington, DC: National Academy Press.

Bailey, A. L., & Butler, F. A. (2003). *An evidentiary framework for operationalizing academic language for broad application to K-12 Education: A design document* (CSE Tech. Rep. No. 611). Los Angeles: University of California, CRESST.

Bunch, M. B. (2006). *Final Report on ELDA Standard Setting*. Durham: Measurement Incorporated.

Chomsky, N. (1968). Language and the mind. *Psychology Today*, *1*(9), 48–68.

Cisneros, R., & Leone, B. (Eds.) (1995). The ESL component of bilingual education in practice: Critical descriptions of bilingual classrooms and programs. *Bilingual Research Journal*, *19*(3&4), 353–367.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Conteh-Morgan (2002). Connecting the dots: Limited English proficiency, second language learning theories, and information literacy instruction. *The Journal of Academic Librarianship*, *28*(4), 191–196.

Cummins, J. (2000). *Language, power and pedagogy: Bilingual children in the crossfire*. Clevedon, U.K.: Multilingual Matters, Ltd.

Del Vecchio, A., & Guerrero, M. (1995). *Handbook of English language proficiency tests*. Albuquerque, NM: New Mexico Highlands University, Evaluation Assistance Center—Western Region.

Fast, M., Ferrara, S., & Conrad, D. (2004). *Current efforts in developing English language proficiency measures as required by NCLB: Description of an 18-state collaboration*. Washington, DC: American Institutes for Research.

Figueroa, R. A. (1989). Psychological testing of linguistic-minority students: Knowledge gaps and regulations. *Exceptional Children*, *56*(2), 145–153.

Figueroa, R. A. (1990). Best practices in the assessment of bilingual children. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (pp. 93–106). Washington, DC: National Association of School Psychologists.

Francis, D., & Rivera, M. (2007). Principles underlying English language proficiency tests and academic accountability for ELLs. In J. Abedi (Ed.), *English language proficiency assessment in the nation: Current status and future practice* (pp. 13–32). Davis: University of California.

Gándara, P. (2000). In the aftermath of the storm: English learners in the post-227 era. *Bilingual Research Journal*, *24*(1&2), 1–13.

Gándara, P., Maxwell-Jolly, J., & Driscoll, A. (2005). *Listening to teachers of English language learners: A survey of California teachers' challenges, experiences, and professional development needs*. Santa Cruz, CA: The Center for the Future of Teaching and Learning (http://www.cftl.org/documents/2005/listeningforweb.pdf) (accessed May 19, 2005).

GAO (2006). *No Child Left Behind Act: Assistance from education could help states better measure progress of students with limited English proficiency*. Washington, DC: United States Government Accountability Office.

Giraud, G., Impara, J., & Plake, B. (2005). Teachers' conceptions of the target examinee in Angoff standard setting. *Applied Measurement in Education*, *18*(3), 223–232.

Grissom, J. B. (2004). Reclassification of English language learners. *Education Policy Analysis Archives*, *12*(36) (http://epaa.asu.edu/epaa/v12n36/) (accessed September 3, 2004).

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

Hakuta, K., Butler, Y. G., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* Santa Barbara: University of California, Linguistic Minority Research Institute.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17–27.

Improving America's Schools Act of 1994, Pub. L. No. 103-382, 108 Stat. 3518 (1994).

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 485–511). Washington, DC: American Council on Education.

Kindler, A. L. (2002). *Survey of the states' limited English proficient students and available educational programs and services, 2000–2001 Summary Report*. Washington, DC: National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Pacific Grove, CA: Brooks/Cole.

Krashen, S. D. (1988). *Second language acquisition and second language learning*. New York: Prentice-Hall International.

LaCelle-Peterson, M. W., & Rivera, C. (1994). Is it real for all kids? A framework for equitable assessment policies for English language learners. *Harvard Educational Review*, *64*(1), 55–75.

Lightbown, P. M., & Spada, N. (2000). *How languages are learned* (revised ed., pp. 38–40). Oxford, U.K.: Oxford University Press.

Linn, R. L., & Gronlund, N. E. (1995). *Measurement and assessment in teaching* (7th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Linquanti, R. (2001). *The redesignation dilemma: Challenges and choices in fostering meaningful accountability for English learners*. Policy Report 2001-1. Santa Barbara: University of California, Linguistic Minority Research Institute.

Littlejohn, J. (1998). *Federal control out of control: The Office for Civil Rights' hidden policies on bilingual education*. Sterling, VA: Center for Equal Opportunity (http://www.ceousa.org/READ/ocr2.html) (accessed November 23, 2004).

Loop, C. (2002). *Which tests are commonly used to determine English and/or Spanish language proficiency?* (AskNCELA No. 25). Washington, DC: National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs.

Mahoney, K. S., & MacSwan, J. (2005). Reexamining identification and reclassification of English language learners: A critical discussion of select state practices. *Bilingual Research Journal*, *29*(1), 31–42.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13–23.

Miller, R. A., & Zhou, X. (1997). *Survival analysis with long-term survivors*. New York: John Wiley & Sons.

NAEP (2007, September 18), NAEP Inclusion Policy (http://nces.ed.gov/nationsreportcard/about/inclusion.asp) (accessed November 12, 2007).

National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs (2002, December). *Glossary of terms related to the education of linguistically and culturally diverse students* (In Ask NCELA No. 10). Washington, DC: Author (http://www.ncela.gwu.edu/expert/glossary.htm) (accessed May 21, 2004).

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Parrish, T., Perez, M., Merickel, A., & Linquanti, R. (2006). *Effects of the implementation of Proposition 227 on the education of English learners, K-12: Findings from a five-year evaluation (final report)*. Palo

Alto and San Francisco: American Institutes for Research and WestEd.

Reutzel, D. R., & Cooter, R. B. (2007). *Strategies for reading assessment and instruction helping every child succeed*. Upper Saddle River, NJ: Pearson/Merrill Prentice Hall.

Rivera, C., Stansfield, C. W., Scialdone, L., & Sharkey, M. (2000). *An analysis of state policies for the inclusion and accommodation of English language learners in state assessment programs during 1998–1999*. Arlington, VA: The George Washington University, Center for Equity and Excellence in Education.

Rossell, C. H. (2000). *Different questions, different answers: A critique of the Hakuta, Butler, and Witt report, 'How long does it take English learners to attain proficiency?'* READ Perspectives, 7, The READ Institute (http://www.ceousa.org/READ/hakuta.html) (accessed November 23, 2004).

Rumberger, R., & Gándara, P. (2004). Seeking equity in the education of California's English learners. *Teachers College Record*, *106*, 2031–2055.

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.

Staley, L. E. (2005). *The effects of English language proficiency on students' performance on standardized tests of mathematics achievement*. Unpublished doctoral dissertation, University of California, Los Angeles.

Stefanakis, E. H. (1998). *Whose judgment counts? Assessing bilingual children, K-3*. Portsmouth, NH: Heinemann.

Tippeconnic, J. W., III, & Faircloth, S. C. (2002). *Using culturally and linguistically appropriate assessments to ensure that American Indian and Alaska Native students receive the special education programs and services they need* (EDO-RC-02-8). Charleston, WV: ERIC Clearinghouse on Rural Education and Small Schools.

United States General Accounting Office (2001). *Meeting the needs of students with limited English proficiency* (GAO-01-226). Washington, DC: Author.

Valdes, G., & Figueroa, R. A. (1994). *Bilingual and testing: A special case of bias*. Norwood, NJ: Ablex.

Zehler, A. M., Hopstock, P. J., Fleischman, H. L., & Greniuk, C. (1994). *An examination of assessment of limited English proficient students*. Arlington, VA: Development Associates, Special Issues Analysis Center.

TCRE	782	B	Dispatch: 6.9.06 Journal: TCRE	CE: Blackwell
Journal Name	Manuscript No.		Author Received: No. of pages: 22	PE: Rashmi/HVA

Psychometric Issues in the ELL Assessment and Special Education Eligibility

JAMAL ABEDI
University of California, Davis

Assessments in English that are constructed and normed for native English speakers may not provide valid inferences about the achievement of English language learners (ELLs). The linguistic complexity of the test items that are not related to the content of the assessment may increase the measurement error, thus reducing the reliability of the assessment. Language factors that are not relevant to the content being assessed may also be a source of construct-irrelevant variance and negatively impact the validity of the assessment. More important, the results of these tests used as the criteria for identification and classification of ELL students, particularly those at the lower end of the English proficiency spectrum, may be misleading. Caution must be exercised when the results of these tests are used for special education eligibility, particularly in placing ELL students with lower English language proficiency in the learning/reading disability category. This article discusses psychometric issues in the assessment of English language learners and examines the validity of classifying ELL students, with a focus on the possibility of misclassifying ELL students as students with learning disabilities.

UNCORRECTED PROOF

The policy of including of English language learners (ELLs) and students with disabilities (SD) is not only a necessary for reliable, valid, and fair assessments for all, but it is also a congressionally mandated policy. The recent No Child Left Behind Act (2001), which is the reauthorization of the Elementary and Secondary Education Act (the Improving America's Schools Act), and the amendments to the Individuals with Disabilities Education Act (IDEA) in 1997 require that all students be included in national and state assessments. The term *all* in the legislation means *every student*—including those with disabilities and limited English proficiency. The intent is to provide appropriate assessment based on the same high standards and ensure that all students are part of the indicators used for school accountability (Thurlow & Liu, 2001).

Many instructional decisions that will be made could have grave consequences for ELLs if their knowledge and skills are not appropriately assessed. Although the increasing level of attention to the inclusion and assessment of these students is encouraging, not enough work has been

Teachers College Record Volume 108, Number 11, November 2006, pp. 2282–2303
Copyright © by Teachers College, Columbia University
0161-4681

TCRE 782
EWUS TCRE 782.PDF 06-Sep-06 20:45 148003 Bytes 22 PAGES n operator=HVAnanth)

done to examine the issues and improve the quality of instruction and assessment for these students. For example, Ortiz (2002) indicated that "students learning English are likely to fail when they do not have access to effective bilingual education of English as a second language (ESL) program" (p. 41). Lack of access to effective education will also affect their assessment results. Research has clearly demonstrated that assessments designed and normed mainly for native English speakers may not as reliable and valid for ELL students.

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999) elaborated on this issue:

> For all test takers, any test that employs language is, in part, a measure of their language skills. This is of particular concern for test takers whose first language is not the language of the test. Test use with individuals who have not sufficiently acquired the language of the test may introduce construct-irrelevant components to the testing process. In such instances, test results may not reflect accurately the qualities and competencies intended to be measured. . . . Therefore it is important to consider language background in developing, selecting, and administering tests and in interpreting test performance. (p. 91)

This article will discuss the impact of linguistic factors on assessment and classification of ELL students. Among the major threats to the validity of classifying ELL students is the indistinct line between ELL students at the lower levels of English proficiency and students with learning disabilities.

We will first explain the psychometric issues in the assessment of ELLs and then discuss the possibility of misclassification—due to the use of inappropriate assessment tools—of ELL students as having a learning disability. Jenkins and O'Connor (2002) indicated that students with reading and/or learning disabilities are not proficient in reading and writing skills. ELL students, particularly those at the lower level of English proficiency spectrum, also suffer from such lack of proficiency in English. Although a comprehensive and valid diagnostic approach can distinguish students with reading/learning disabilities from ELL students, distinguishing between these two groups of students can be a daunting task.

PSYCHOMETRIC ISSUES IN THE ASSESSMENT OF ELL STUDENTS

Literature on the assessment of English language learners suggests that students' content-based assessment outcomes may be confounded by lan-

UNCORRECTED PROOF

TCRE 782.PDF 06-Sep-06 20:45 148003 Bytes 22 PAGES n operator=HVAnanth)

guage background variables. ELLs generally perform lower than non-ELLs on content-based assessments such as math, science, and social sciences—a strong indication that English language proficiency affects instruction and assessment. Research also shows that ELL students' assessment outcomes suffer from lower reliability and validity; that is, language factors may be a source of measurement error in the content-based assessment of ELL students and may impact the reliability of the test. Language factors may also be a source not relevant to the construct of such assessments (Messick, 1994) and may affect the test's construct validity.

In the assessment of ELL students, the results of studies suggested that unnecessary linguistic complexity is a nuisance variable that introduces a source of measurement error and is considered a construct-irrelevant factor in the assessment (Abedi & Leon, 1999; Abedi, Leon, & Mirocha, 2003). Such nuisance variables may seriously confound the outcome of assessment in content-based areas (Abedi, Lord, & Hofstetter, 1998; Cocking & Chipman, 1988; De Corte, Verschaffel, & DeWin, 1985; Kintsch & Greeno, 1985; Trenholme, Larsen, & Parker, 1978; Lepik, 1990; Mestre, 1988; Munro, 1979; Noonan, 1990; Orr, 1987; Rothman & Cohen, 1989; Spanos, Rhodes, Dale, & Crandall, 1988).

The results of a series of experimentally controlled studies by researchers at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) on the impact of language on test performance of ELLs have clearly demonstrated that (1) ELLs have difficulty with linguistically complex test items, and (2) reducing linguistic complexity of test items narrows the performance gap between ELL and non-ELL students in content-based areas such as math and science (see for example, Abedi & Lord, 2001; Abedi, Lord, & Plummer, 1997; Abedi, Lord, Hofstetter, & Baker, 2000; Abedi et al., 1998; Abedi, Lord, Kim-Boscardin, & Miyoshi, 2000). Summarized below are studies that demonstrate the impact of language on the assessment outcome of ELLs and suggest that (1) ELLs had more difficulty with test items that were more linguistically complex, and (2) modifying test items to reduce the level of their linguistic complexity reduced the performance gap between ELL and non-ELL students in content-based areas.

Analyses of test data from four locations nationwide (Abedi et al., 2003) found a large performance gap between ELL and non-ELL students in reading and writing, areas that have a substantial amount of language demand. The performance gap was lower for science and lowest for math problem solving, for which the test items were less linguistically challenging for ELL students. The performance gap virtually disappeared in math computation, for which the language demands of the test items were minimal.

By reducing the impact of language factors on content-based test performance, the validity and reliability of assessments can be improved and can result in fairer assessments for all students—including ELLs and stu-

UNCORRECTED PROOF

TCRE 782 US TCRE 782.PDF 06-Sep-06 20:45 148003 Bytes 22 PAGES n operator=HVAnanth)

dents with disabilities. To minimize the impact of language factors and consequently reduce the performance gap between ELL and other students, language modification of assessment tools may be a viable option. For example, when math test items were modified to reduce the level of linguistic complexity, over 80% of middle school students who were interviewed preferred the linguistically modified over the original English version of the test items (see Abedi et al., 1997). Abedi et al. (1998), in a study of 1,394 eighth-grade students in schools with high enrollments of Spanish speakers, showed that modification of language of the items contributed to improved performance on 49% of the items; the students generally scored higher on shorter problem statements.

Another study (Abedi & Lord, 2001) of 1,031 eighth-grade students in Southern California found small but significant differences in the scores of students in low- and average-level math classes taking the linguistically modified version of the test. Among the linguistic features that appeared to contribute to the differences were low-frequency vocabulary, conditional clauses, and passive-voice verb constructions (for a description of the linguistic features below and for a discussion of the nature of and rationale for the modifications, see Abedi et al., 1997).

Beattie, Grise, and Algozzine (1983) found positive results in modifying tests for students with a learning disability. Math, reading, and writing tests were modified in the following ways: hierarchical progression of difficulty; unjustified arrangement of sentences; vertical arrangement of bubbles; placement of passages in shaded boxes; examples set off from test items; and arrows and stop signs in the corner of pages to indicate continuing and ending pages.

Tindal, Anderson, Helwig, Miller, and Glasgow (2000) used "simplified language" as a test accommodation for students with a learning disability, which they argued could also be used for ELLs. Results indicated that simplifying the language did not affect the test. However, the authors noted that their simplification process was perhaps too limited, which suggested the need for future studies to expand the simplification process.

Another study consisting of 422 students in eighth-grade science classes (Abedi, Lord, Kim-Boscardin, & Miyoshi, 2000) compared performance on National Assessment of Educational Progress (NAEP) science items in three test formats: one booklet in original format (no accommodation); one booklet with English glosses and Spanish translations in the margins; and one booklet with a customized English dictionary at the end of the test booklet. The customized dictionary, which was used for the first time by Abedi and his colleagues, included only the non-content words that appeared in the test items. By helping students with their language needs, English learners scored highest on the customized dictionary accommodation (their mean scores for the customized dictionary was 10.18 on a 20-item test as compared with means of 8.36 and 8.51 for the other two accommodations).

UNCORRECTED PROOF

TCRE 782.PDF 06-Sep-06 20:45 148003 Bytes 22 PAGES n operator=HVAnanth)

In a study on the impact of accommodation on eighth-grade students in math, Abedi, Lord, Hofstetter, and Baker (2000) applied four different types of accommodation: linguistically modified English version of the test; standard NAEP items with glossary only; extra time only; and glossary plus extra time. Students were also tested using standard NAEP items with no accommodation. Among these accommodations, linguistic modification of test items was the only accommodation that reduced the performance gap between ELL students and non-ELL students. Because the non-ELL students in this study, who are among the low-performing student population, also benefited from linguistic modification of test items, this suggests that clarifying the language of assessment may be helpful not only to ELL students but to SD students as well.

The summary of research above suggests that reducing the linguistic complexity of assessment materials helps ELL students and low-performing native English speakers to provide a more valid picture of what they know and can do. All students can benefit from instructional materials that are easier to understand (i.e., material with unnecessary linguistic complexity). Similarly, all students can better understand assessments that clearly convey the message related to the concept being assessed.

LINGUISTIC FEATURES THAT MAY IMPACT COMPREHENSION

The results of CRESST research led to identification of linguistic features that have greater effects on ELL student performance. These features slow down students, make misinterpretation more likely, and add to students' cognitive load, thus interfering with concurrent tasks. Indexes of language difficulty include unfamiliar words, long phrases in questions, complex sentences, and conditional and adverbial clauses. Other linguistic features that may cause difficulty for readers include long noun phrases, relative clauses, prepositional phrases, abstract versus concrete presentation of problem, passive voice, and negation. Below is a brief description of some of these features, along with some illustrative examples. A few references are added for each of the features. For a detailed description of these features, see Abedi, Courtney, and Goldberg (2000).

UNFAMILIAR WORDS

Assessments containing unfamiliar words are more difficult for ELL students than those with more familiar words. Some words, word pairs, or groups of words still unfamiliar to ELLs might be used in a test item. They are unnecessary if they are not essential to the concept being tested.

Idioms are words, phrases, or sentences that cannot be understood literally. Many proverbs, slang phrases, phrasal verbs, and common sayings

UNCORRECTED PROOF

TCRE 782.PDF 06-Sep-06 20:45 148003 Bytes 22 PAGES n operator=HVAnanth)

cannot be decoded by ELLs. On the other hand, words that are high on a general frequency list for English are likely to be familiar to most readers because they are often encountered. Following are a few examples of unfamiliar words used in assessments (Abedi et al., 1997; Adams, 1990; Chall, Jacobs, & Baldwin, 1990; Dale & Chall, 1948; Gathercole & Baddeley, 1993; Klare, 1974;).

Circle the clumps of eggs in the illustration.
Patty expects that each tomato plant in her garden will bear 24 tomatoes.
In the last census, 80% of the households had one or more wage-earners.

LONG PHRASES IN QUESTIONS

Long questions are used less than short questions. Complex question types might have an opening phrase or clause that either replaces or postpones the question word (Adams, 1990).

At which of the following times should Ed feed the parking meter?
Of the following bar graphs, which represents the data?
According to the passage above, where do sea turtles lay their eggs?

COMPLEX SENTENCES

A complex sentence contains a main clause and one or more subordinating (dependent) clauses. Subordinating words include *because, when, after, although, if,* and *since* (more on *if* under Conditional Clauses; Botel & Granowsky, 1974; Hunt, 1965, 1977; Wang, 1970).

Because she wants to stay in touch, Peggy frequently _____.
When she came home, he _____ the letter.
Although the ship was _____, she was calm.

LOGICAL CONNECTORS: CONDITIONAL/ADVERBIAL CLAUSES

Conditional clauses and adverbial clauses are among the sources contributing to the linguistic complexity of assessments. Logical connectors are adverbial expressions that allow a listener or reader to infer connections between two structures. They include dependent words (subordinating conjunctions; see above). In mathematics, they often include conditional "if-then" statements. Some take the form of complex sentences (Celce-Murcia & Larsen-Freeman, 1983; Haiman, 1985; Shuard & Rothery, 1984; Spanos et al., 1988).

UNCORRECTED PROOF

/US TCRE 782.PDF 06-Sep-06 20:45 148003 Bytes 22 PAGES n operator=HVAnanth)

Adverbial clauses:

When the barber was finished with the haircut, he took the customer's money.
While he was listening to music, he did his homework.

Conditional clauses:

As long as you bring your own bedding, you can stay with us.
Given that *a* is a positive number, what is–*a?*
If one pint will fill 2 cups, how many cups can be filled from 8 pints? (vs. One pint will fill 2 cups. Eight pints will fill _____ cups).
In Jean's class, there are twice as many boys as girls. If there are 10 girls in the class, how many boys and girls are there in the class?

LONG NOUN PHRASES

Nouns sometimes work together to form one concept, such as *pie chart* or *bar graph*. Sometimes adjectives and nouns work together to create meaning: *high school diploma, income tax return*. To further complicate interpretation, strings of adjectives and nouns create subjects and objects: *freshwater pond, long-term investment, new word processing program* (Celce-Murcia & Larsen-Freeman, 1983; Halliday & Martin, 1993; Just & Carpenter, 1980; King & Just, 1991; MacDonald, 1993; Spanos et al., 1988).

A loaded trailer truck weighs 26,643 kilograms. When the trailer truck is . . .
Of the following number pairs, which is the dimension of a 100-square-foot room?
To become next year's tennis team captain, how many votes will Sandra need?

RELATIVE CLAUSES

A relative clause is an embedded clause that provides additional information about the subject or object it follows. Because relative clauses are less frequent in spoken English than in written English, some students may have had limited exposure to them. Words that lead a relative clause include *that, who*, and *which*. Note: Often *that* is omitted from a relative clause. When possible, relative clauses should be removed or recast (Pauley & Syder, 1983; Schachter, 1983).

UNCORRECTED PROOF

TCRE 782.PDF 06-Sep-06 20:45 148003 Bytes 22 PAGES n operator=HVAnanth)

A bag that contains 25 marbles . . . (vs. One bag has 25 marbles. A second . . .)
Joe found the student who had loaned him the book.

PREPOSITIONAL PHRASES

Prepositional phrases work as adjectives or adverbs to modify nouns, pronouns, verbs, adverbs, or adjectives. When they occur before question words, between the subject and the verb, or in strings, they can be especially confusing to ELLs (Orr, 1987; Slobin, 1968; Spanos et al., 1988).

Which of the following is the best approximation of the area of the shaded rectangle in the figure above if the shaded square represents one unit of area?

ABSTRACT (VS. CONCRETE) PRESENTATION OF PROBLEM

Respondents show better performance when assessment questions are presented in concrete rather than abstract terms. Information presented in narrative structures tends to be understood and remembered better than information presented in expository text (Cummins, Kintsch, Reusser, & Weimer, 1988; Lemke, 1986).

The weights of two objects were measured *vs*. The clerk weighed two suitcases.

PASSIVE VOICE

Assessments containing passive-voice construction are more difficult for ELL students to follow. In active voice, the subject is the one performing an action. In passive voice, the one receiving the action is in the subject position. Often the "actor" is not stated (Abedi et al, 1997; Celce-Murcia & Larsen-Freeman, 1983; Forster & Olbrei, 1973)

He was given a ticket *vs*. The officer gave him a ticket.
Girls' ears were pierced in infancy *vs*. Parents pierced infant girls' ears.
When comparisons were made, the amounts in each jar had been reduced.

NEGATION

Studies suggest that a sentence containing negations (e.g., *no*, *not*, *none*, *never*) are harder to comprehend than affirmative sentences. Several types of negative forms are confusing to ELLs (Mestre, 1988):

UNCORRECTED PROOF

TCRE 782.PDF 06-Sep-06 20:45 148003 Bytes 22 PAGES n operator=HVAnanth)

Proper double negative:

Not all the workers at the factory are not male.
It's not true that all the workers at the factory are not male.

POSSIBLE IMPACT OF LANGUAGE FACTORS ON RELIABILITY OF ASSESSMENT

The unnecessary linguistic complexity of test items can be a source of measurement error and can reduce the reliability of the tests. Because true-score variance (σ^2_T) in classical test theory is defined as the observed-score variance (σ^2_X) minus the error variance (σ^2_E), any increase in the size of error variance directly affects (reduces) the size of true score variance (Allen & Yen, 1979; Linn & Gronlund, 1995; Salvia & Ysseldyke, 1998) and consequently decreases the reliability of the assessment. In a perfectly reliable test, the error variance (σ^2_E) would be zero; therefore, the true-score variance (σ^2_T) would be equal to the observed-score variance.

However, in measurement involving human subjects, there is always an error component. Appropriate evaluation of the measurement error is important in any type of assessment, whether in the traditional multiple-choice approach or in performance-based assessments (Linn, 1995; see also AERA, APA, & NCME, 1999). Many different sources (e.g., occasion, task, test administration conditions) may contribute to measurement error in traditional assessment instruments for all students. It is important to note, however, that the unnecessary linguistic complexity of test items as a source of measurement error differentially impacts performance of different groups of students with different levels of English proficiency. The linguistic complexity factor affects the performance of ELL students because the common characteristic of these students is their needs in the area of English language. Thus, there is an interaction between students' ELL status and their underlying measurement model.

In the classical approach to estimating reliability of assessment tools, the level of contribution of different sources to measurement error may be indeterminable. Through the generalizability approach, one would be able to determine the extent of the variance that each individual source contributes (such as occasion, task, item, scorer, and language factors) to the overall measurement error (see Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991).

To estimate reliability of the standardized achievement tests and to investigate their measurement error across different subgroups of students (e.g., ELLs versus non-ELLs), we considered different approaches. Because parallel forms or test-retest data were not available in many districts and

UNCORRECTED PROOF

TCRE 782.PDF 06-Sep-06 20:45 148003 Bytes 22 PAGES n operator=HVAnanth)

states across the nation, we decided to use an internal consistency approach for estimating the reliability of some of the commonly used standardized achievement tests. As mentioned, the linguistic complexity of test items may introduce another dimension into the assessment, the dimension of language. If this is the case, and because the internal consistency approach is extremely sensitive to multidimensionality of test items (see, for example, Abedi, 1996; Cortina, 1993), then the estimated reliability for ELL students may be considerably lower than the reliability for native speakers of English. One may also argue that language factors may create a restriction of range on content-based outcome measures, and the restriction of range may reduce the internal consistency coefficient. In other words, unnecessary linguistic complexity makes assessment more difficult for ELL students, thereby reducing their performance level. Lower performance level then creates a restriction of range for the ELL performance distribution, and that in turn results in lower reliability for ELL students. These two explanations both relate to the impact of language factors on the assessment of ELLs, and both indicate the negative effects of linguistic complexity of content-based assessment on ELL student performance. To illustrate this phenomenon, we present results of some analyses performed by CRESST researchers on the reliability of test items.[1]

Table 1 presents internal consistency (alpha) coefficients for the Stanford 9 subscale scores for second- and ninth-grade students from one of our data sites. Data in Table 1 show two interesting trends: (1) as expected, the internal consistency coefficients for non-ELL students were higher than those for ELL students, and (2) the difference between ELL and non-ELL students was higher in the higher grades, in which content-based language is more complex.

For students in second grade, the alpha coefficient was .914 for non-ELL students as compared with .856 for ELL students. In math, the alpha for

Table 1. Internal Consistency Coefficients for ELL/Non-ELL Students in Grades 2 and 9

	Grade 2		Grade 9	
Content Area	Non-ELL	ELL	Non-ELL	ELL
Reading	.914 (n = 234,505)	.856 (n = 101,399)	.876 (n = 181,202)	.750 (n = 52,720)
Math	.894 (n = 249,000)	.881 (n = 118,740)	.898 (n = 183,262)	.802 (n = 54,815)
Science			.805 (n = 144,821)	.597 (n = 40,255)
Social science			.806 (n = 181,078)	.530 (n = 53,925)

UNCORRECTED PROOF

TCRE 782.PDF 06-Sep-06 20:45 148003 Bytes 22 PAGES n operator=HVAnanth)

non-ELL students was .894, as compared with .881 for ELL students. As these data show, although alpha coefficients for ELL students were lower, the gap between ELL and non-ELL students was not large in second grade. For students in ninth grade, however, there was a larger gap between ELL and non-ELL students. In reading, for non-ELL students, the alpha coefficient was .876, as compared with .750 for ELL students. In math, the alpha for ninth grade non-ELL students was .898, as compared with .802 for ELL students. In science, the alpha for non-ELLs was .805, as compared with .597 for ELLs. Finally, in social science, the alpha for non-ELLs was .806, as compared with .530 for ELLs.

As these data suggest, the difference between internal consistency coefficients for ELL and non-ELL subscale scores were substantially larger for students in higher grades (ninth grade) than for students in lower grades (second grade). These differences were statistically significant. The average alpha for students in second grade over all subject areas was .904 for non-ELL students, as compared with .869 for ELL students—a small difference of 4%. For students in ninth grade, however, the average alpha for non-ELL was .846, as compared with an average alpha of .670, a difference of 26%. Comparing the percent alpha difference of 26% for ninth-grade students with the 4% of second-grade students once again suggests that in a more linguistically complex environment, the difference between ELL and non-ELL students is more apparent.

VALIDITY

In content-based assessments such as in math and science, the linguistic complexity of test items may introduce another dimension or construct in addition to the construct that is the target of assessments. This may be the case particularly for ELL students. The linguistic complexity factors in content-based assessment may be considered a source of construct-irrelevant variance because it is not conceptually related to the content being measured (Messick, 1994):

> With respect to distortion of task performance, some aspects of the task may require skills or other attributes having nothing to do with the focal constructs in question, so that deficiencies in the construct-irrelevant skills might prevent some students from demonstrating the focal competencies. (p. 14)

The concept of "construct-irrelevant" applies to the situations in which a construct other than the construct targeted for assessment is involved. For example, when the linguistic structure of an assessment in a content area (e.g., math or science) is so complex that ELL students cannot adequately

UNCORRECTED PROOF

TCRE 782.PDF 06-Sep-06 20:45 148003 Bytes 22 PAGES n operator=HVAnanth)

understand the question, then the English language becomes another construct that is measured by the test but is not relevant to the content being assessed. In other words, language interferes with the targeted content in the assessment. Linguistic complexity of test items as a possible source of construct-irrelevant variance may be a threat to the validity of achievement tests because it could be a source of measurement error in estimating the reliability of the tests. The construct-irrelevant variance may change the structural relationships among test items, between subscale scores and the fit of the structural model. Because linguistic factors may have more influence on the assessment outcomes for ELL students, then the structural relationships of ELL assessment outcomes may be different with those of English-only students.

To examine the hypothesis of differences between ELL and non-ELL students on the structural relationship of test items, a series of structural equation models was created. Fit indices were compared across ELL and non-ELL groups. The results generally indicated that the relationships between individual items, items with the total test score, and items with the external criteria were stronger for non-ELL students than for ELL students.

Item parcels in each content area (e.g., reading, science, and math) were created. Each parcel was constructed to systematically contain items with varying degrees of item difficulty. Through this process, each parcel contained three categories of difficulty: easy, difficult, and moderately difficult items. The main reason for creating item parcels was to provide multiple measures for each of the content areas. For example, rather than having a single score of math as the sum of all items in a math test, items were divided into several groups, or parcels. Each parcel of items provided a measure of math; therefore, we obtained as many measures of math as the number of parcels (for a detailed description of item parcels and ways to create them, see Cattell & Burdsal, 1975). A reading latent variable was constructed based on these four parcels. Similarly, item parcels and latent variables for science and math were created from the 48 math items and 40 science items by the same process. The correlations between the reading, math, and science latent variables were estimated. Models were tested on randomly selected subsamples to demonstrate the cross-validation of the results.

Table 2 shows the results of the structural models for ninth-grade students from a large state. To examine the consistency of the results of these analyses over independent samples, students were randomly divided into two cross-validation sample, namely sample 1 and sample 2. The results obtained under these two independent samples were consistent. For example, the average factor loading across all four item parcels in each content area and across all three content areas was .795 for sample 1 and

UNCORRECTED PROOF

TWUS TCRE 782.PDF 06-Sep-06 20:45 148003 Bytes 22 PAGES n operator=HVAnanth)

Table 2. Grade 9 Students From Data Site, Stanford 9 Reading, Math, and Science Structural Modeling Results (df = 51)

	Non-ELL (N = 22,782)		ELL (N = 4,872)	
	Sample 1	Sample 2	Sample 1	Sample 2
Average factor loadings Across the four parcels				
Reading comprehension	.846	.846	.746	.749
Math	.830	.830	.711	.724
Science	.708	.707	.541	.539
Average factor correlation across the three content areas	.830	.827	.794	.755
Goodness of fit				
Chi square	488	446	152	158
NFI	.997	.998	.992	.992
NNFI	.997	.997	.993	.993
CFI	.998	.998	.995	.995

Note. There was significant invariance for all constraints tested with the multiple group model (Non-ELL/ELL).

UNCORRECTED PROOF

.794 for sample 2 for non-ELL students. For ELL students, the average factor loading was .666 for sample 1 and .671 for sample 2. Similarly, there is a high level of consistency on average factor correlations across the two independent samples. The average factor correlation for non-ELL was .830 for sample 1 and .827 for sample 2. For ELLs, the average factor correlation was .794 for sample 1 and .755 for sample 2. Once again, these data suggest high level of consistency across the cross-validation samples.

Table 2 also compares the structural relationships of test items across the categories of ELL. The data show major differences between ELL and non-ELL students. As data in Table 2 suggest, correlations of item parcels with the latent factors (factor loadings) were consistently lower for ELL students than they were for non-ELL students. For example, the average factor loadings across different content areas and multiple samples was .795 for non-ELLs as compared with .668 for ELLs, a substantial difference. Similarly, there was a large difference between ELL and non-ELLs on average factor correlations. For non-ELLs, the average factor correlation was .829, as compared with .774 for ELLs—once again, a large difference.

In term of fit indices, the structural models showed a good fit of the model to the data for both ELLs and non-ELLs. However, the trend of differences between ELL and non-ELLs was also seen here, even though the difference was small. Models for non-ELLs had slightly higher fit as compared with the models for ELLs.

TCRE 782.PDF 06-Sep-06 20:45 148003 Bytes 22 PAGES n operator=HVAnanth)

The hypotheses of invariance of factor loadings and factor correlations between the ELL and non-ELL groups were tested. Specifically, we tested the invariance of (1) the correlations between parcel scores and a reading latent variable; (2) correlations between parcel scores and a science latent variable; (3) correlations between parcel scores and a math latent variable; and (4) correlations between content-based latent variables across the ELL and non-ELL groups.

The null hypotheses for all these tests of invariance were rejected, suggesting that ELL and non-ELL students responded differently to the test items.

ISSUES CONCERNING ELL CLASSIFICATION AND SPECIAL EDUCATION ELIGIBILITY

Researchers have expressed concerns over the validity of classification for ELL students. Because of the lack of a commonly accepted operational definition of the term *ELL* or *LEP* (limited English proficiency) and because of validity issues in the criteria used for such classification, large discrepancies have been reported in the ELL classification practices across the nation (see, for example, Abedi, 2004, 2005; Abedi et al., 1997). Although problems in the classification of ELL students is very serious and affects both instruction and assessment for these students, a more serious problem is the possibility of ELL students at the lower level of English proficiency distribution being misidentified as students with disabilities because students' limitations in English may be interpreted as a sign of learning (or reading) disability.

Artiles, Rueda, Salazar, and Higareda (2005) found that ELL students with lower levels of proficiency in L1 and L2 (first and second language, respectively) showed the highest rate of identification in the special education categories. The authors also indicated that more ELL students tend to be placed in the "learning disability" category than in "language and speech impairment." Similarly, Artiles and Ortiz (2002) found a differential rate of overrepresentation of ELL students in special education programs. For example, based on their data, 26.5% of ELLs in Massachusetts, 25.3% in South Dakota, and 20.1% in New Mexico were placed in special education programs as compared with less than 1% of ELLs in Colorado, Maryland, and North Carolina placed in similar programs. Rueda, Artiles, Salazar, and Higareda (2002) reported that in a 5-year period—1993–1994 to 1998–1999—the placement rate of Latino ELLs increased by 345%, while their overall population in the district increased by only 12% during this period of time.

To examine this complex issue of classification of ELL students when related to eligibility for special education, we first discuss issues concerning

UNCORRECTED PROOF

TWUS TCRE 782.PDF 06-Sep-06 20:45 148003 Bytes 22 PAGES n operator=HVAnanth)

validity of classification for ELL students and then elaborate on the criteria for placing ELL students in learning disability category. It must be noted at this point, however, that issues concerning classification of students with learning disability are beyond the scope of this article. The aim of this section is to discuss some of the technical issues concerning ELL students who are placed in the learning disability category. (For a thorough discussion of classification issues for students with learning disabilities, see Bradley, Danielson, & Hallahan, 2002).

Based on Title IX No. 25 in the No Child Left Behind Act (2001), an LEP student is defined as someone who (1) is aged 3–21; (2) is enrolled or preparing to enroll in an elementary or secondary school; (3) was not born in the United States or whose native language is not English; (4) is a Native American, Alaskan Native, or a resident of outlying areas; (5) comes from an environment in which a language other than English has had a significant impact on an individual's English language proficiency; (6) is migratory and comes from an environment where English is not the dominant language; and (7) has difficulties in speaking, reading, writing, or understanding the English language that may deny the individual the ability to meet the state's proficient level of achievement, to successfully achieve in classrooms where English is the language of instruction, or to participate fully in society (No Child Left Behind Act).

The above definition is based on information related to students' language background and their level of English proficiency. Research has expressed concerns about these sources of information. Information on students' language background is obtained from the Home Language Survey (HLS), and data on the students' level of English proficiency are based on existing English language proficiency tests. Unfortunately, the validity of the HLS is often threatened because parents may give inconsistent information for a variety of reasons, including concerns over equity of opportunity for their children, citizenship issues, and the literacy of the parent (see Abedi, 2005). Research has also raised concerns about the validity of existing English language proficiency scores as a criterion for ELL classification. Reviewers of these tests found major differences between the content that these tests measure and the alignment of the content of these tests to English as a Second Language (ESL) standards (see Abedi, 2005; Zehler, Hopstock, Fleischman, & Greniuk, 1994).

Let us assume that students are correctly being classified across the categories of ELL (ELL/non-ELL). The next question would be to look into the validity of criteria used for special education eligibility for these students. There are many different forms of disabilities with different needs and different characteristics. Based on the data from the U.S. Department of Education, Office of Special Education and Rehabilitative Services, there are at least 12 categories of disabilities. Among these categories, however,

UNCORRECTED PROOF

TCRE 782.PDF 06-Sep-06 20:45 148003 Bytes 22 PAGES n operator=HVAnanth)

students with a learning disability are the largest group, constituting 46% of all students with disabilities (National Center for Education Statistics, 2000).

Jenkins and O'Connor (2002) summarized some of the techniques that have been used for identifying students with reading disabilities (RD). Students with reading and/or learning disabilities often leave elementary school with deficient reading and writing skills, which makes early identification and prevention important. The authors defined the foundation of reading as "the ability to read words; the ability to comprehend language; and the ability to access background and topical knowledge relevant to specific texts" (p. 100). A student with RD, they argued, has a weakness in one or more of these three foundation areas. However, during the early developmental stages of reading, "word-level reading skill" is the most salient characteristic, and difficulties in this area can signal an RD.

Jenkins and O'Connor (2002) provided instructions as what they consider "sensible actions" to identify children with reading disability based on research. Among the criteria suggested are assessment of the prerequisite skills of letter naming and phonemic awareness early in kindergarten; watching children as they attempt to write or spell words for clues into their understanding of the alphabetic principle; and recording progress in letter and phonemic knowledge in ways that encourage closer monitoring of children who appear most at risk. It is therefore clear from the discussion above that language factors are among the most important criteria for classifying a student as having a learning disability.

Students with learning disabilities and ELL students (particularly those at the lower levels of English proficiency distribution) may have more difficulty with test items that have unfamiliar words and/or a complex linguistic structure. Thus, language factors that affect the performance of ELLs may also influence the performance of students with a learning disability. These similarities between the language background characteristics and the level of English proficiency may make ELL students with lower level of English particularly vulnerable for misclassification as students with learning and/or reading disability.

Earlier in this article, we presented results of data analyses that showed larger performance gaps between ELL and non-ELLs in areas with greater levels of language demand. A similar trend can be seen for students with disabilities in general, and students with learning disabilities in particular. To demonstrate this trend, we present results of analyses of achievement test data for students identified as having disabilities/learning disability. These results show that these students also have difficulty in content areas with higher level of language demands.

A 4-year trend of student performance in reading and math was examined on the New York State Pupil Evaluation Program (PEP) test for the third and sixth grades from 1995 to 1998 (New York State Education De-

UNCORRECTED PROOF

TCRE 782.PDF 06-Sep-06 20:45 148003 Bytes 22 PAGES n operator=HVAnanth)

partment, 1999). In all 4 years and for both grade levels, the performance gap between the SD and non-SD students[2] was much larger on the reading assessment than on the math assessment. For example, in 1995, the gap between the percentage of third-grade SD and non-SD students scoring above the state reference point was 51.6 percentage points on the reading assessment and 35.0 points on the math assessment. In 1998, the gap between the percentage of SD and non-SD students scoring above the state reference point was 46.6 percentage points on the reading assessment and 27.2 points on the math assessment. Similar PEP test performance gaps between SD students and non-SD students were seen in sixth grade. It is interesting to note that on a separate state assessment (Regents Competency Test), sixth-grade performance gaps between SD and non-SD students for reading and math were much smaller. Many SD students were tested on the Regents Competency tests as compared with the PEP tests. This discrepancy highlights the effect that testing only a small proportion of the SD population can have on the interpretation of results.

As part of a recent CRESST study, we examined the 1998 reading and math Stanford 9 Test data for grades 3 and 11 for a state with a large student population. In each grade level, the gap in performance between the SD students and non-SD students was larger on the reading assessment than on the math assessment. For example, in grade 11, the gap between the mean normal curve equivalent (NCE) of SD/non-ELL and non-SD/non-ELL students was 23.2 on the reading assessment and 18.3 points on the math assessment. The gap between the mean NCE of SD/non-ELL students and non-SD/non-ELL students was 33.7 on the reading assessment and 23.1 points on the math assessment. The gaps were smaller in third grade, but again, the SD student population had more difficulty with the reading assessment than with the math assessment (Abedi et al., 2003).

In the same study, data from another state provided 1998 Stanford 9 test data for grades 3 and 10 in reading and math. In each grade level, the gap in performance between the SD students and non-SD/non-ELL students was larger on the reading assessment than on the math assessment. For example, in grade 10, the gap between the mean NCE of SD/non-ELL and non-SD/non-ELL students was 27.1 on the reading assessment and 21.2 points on the math assessment. The gap between the mean NCE of SD/non-ELL students and non-SD/non-ELL students was 39.8 on the reading assessment and 25.7 points on the math assessment. The gaps were smaller in third grade, but again, the SD population had more difficulty with the reading assessment than with the math assessment (Abedi et al., 2001).

An examination of New Jersey student performance in language arts, science, and math in 2001 on the Elementary and Grade School Proficiency Assessments (ESPA & GSPA) for fourth and eighth grades revealed that in each grade level, the gap in performance between the SD students and non-

UNCORRECTED PROOF

TCRE 782.PDF 06-Sep-06 20:45 148003 Bytes 22 PAGES n operator=HVAnanth)

SD students was larger on the language assessment than on the science and math assessments (New Jersey Department of Education, 2001). For example, in fourth grade, the gap between the percentage of students scoring in the partially proficient category of SD and non-SD/non-ELL students was 39.6 on the language arts assessment, compared with 20.1 on the science assessment and 33.2 points on the math assessment. In eighth grade, the gap between the percentage of students scoring in the partially proficient category of SD and non-SD/non-ELL students was 56.9 on the language arts assessment, compared with 42.5 on the science assessment and 52.7 points on the math assessment.

These findings once again clearly suggest that language factors not only influence the performance of ELL students, but they also affect the performance of students with disabilities, particularly those identified as having learning disability.

DISCUSSION

Federal and state legislation calls for equal educational opportunity and inclusion of all students in assessments. On the other hand, research on the assessment and accommodation of ELL students questions the fairness of assessments that are used for these students, particularly those assessments that are developed and normed for mainstream native English speakers.

Studies that were summarized in this article clearly show a large performance gap in content-based assessment outcomes between English language learners (ELLs) and native English speakers. However, this performance gap is not necessarily due to the lack of content knowledge; it may be due to students' lack of English proficiency. The confounding of language factors with the content knowledge has raised concerns about the validity and authenticity of the available high-stakes assessment and accountability systems for ELLs, particularly those at the lower level of English proficiency.

Assessment tools that have complex linguistic structures may provide poor achievement outcomes for ELLs and SDs. The results of such assessments may not be as reliable and valid for ELLs and SDs as for non-ELL/non-SD students. Consequently, decisions made based on the results of these assessments may be problematic for ELL students and other subgroups of students with language barriers. In this article, based on the findings of experimentally controlled studies, we illustrated that the reliability of commonly used standardized assessments in content-based areas may be negatively affected by the complex linguistic structure of test items, a construct that is not the target of assessment. We have also discussed the influence of linguistic complexity of test items as a source of construct-irrelevant variance in influencing the validity of assessment.

UNCORRECTED PROOF

TCRE 782.PDF 06-Sep-06 20:45 148003 Bytes 22 PAGES n operator=HVAnanth)

As we demonstrated in this article, there is a larger performance gap between ELL and non-ELLs in areas with greater levels of language demand. We also showed a similar trend for students with disabilities in general, and students with learning disabilities in particular. Therefore, language factors that affect the performance of ELLs may also influence the performance of students with a learning disability. These similarities between the language background characteristics and the level of English proficiency may make ELL students with lower levels of English particularly vulnerable to misclassification as students with learning and/or reading disability.

Thus, assessment results that are influenced by linguistic factors as construct-irrelevant may not be valid criteria in the classification of ELL students. This situation becomes even more complex when ELL students are being assessed for eligibility in special education. Unfortunately, as we demonstrated in this article, the likelihood of misclassification of low-performing ELL students as students with a learning disability is not negligible. Care must be taken to increase the validity and authenticity of criteria used for eligibility of ELL students for special education. Misclassification of ELL students, particularly misidentifying them as students with learning disabilities, may have very serious consequences for these students. They may be placed in an inappropriate educational system and subsequently receive inappropriate curriculum.

Based on the results of multiple studies, cited in this article, that focus on the impact of language factors on assessment of the special needs student population, we believe that if the education community truly wants no child left behind, serious attention must be given to the current assessment and classification system for English language learners and students with disabilities, particularly ELL students with lower levels of English proficiency.

UNCORRECTED PROOF

Notes

1 Data were obtained from four different locations in the nation. For further detail regarding these sites, please see Abedi, Leon, and Mirocha (2003).

2 In the studies mentioned in the rest of this subsection, the population referred to as "non-SD students" does not include English language learners (ELLs); thus, the comparison group is less likely struggling with language.

References

Abedi, J. (1996). The interrater/test reliability system (ITRS). *Multivariate Behavioral Research*, *31*, 409–417.

Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher*, *33*, 4–14.

TCRE 782.PDF 06-Sep-06 20:45 148003 Bytes 22 PAGES n operator=HVAnanth)

Abedi, J. (2005). Issues and consequences for English language learners. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data in accountability testing* (pp.). Malden, MA: Blackwell. Q1

Abedi, J., Courtney, M., & Goldberg, J. (2000). *Language modification of reading, science and math test items*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., & Leon, S. (1999). *Impact of students' language background on content-based performance: Analyses of extant data*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Leon, S., & Mirocha, J. (2001). *Examining ELL and non-ELL student performance differences and their relationship to background factors: Continued analyses of extant data*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Leon, S., & Mirocha, J. (2003). *Impact of student language background on content-based performance: Analyses of extant data* (CSE Tech. Rep. No. 603). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, *14*, 219–234.

Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance* (CSE Tech. Rep. No. 478). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

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.

Abedi, J., Lord, C., Kim-Boscardin, C., & Miyoshi, J. (2000). *The effects of accommodations on the assessment of LEP students in NAEP* (CSE Tech. Rep. No. 537). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Lord, C., & Plummer, J. (1997). *Language background as a variable in NAEP mathematics performance* (CSE Tech. Rep. No. 429). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

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.

Artiles, A. J., & Ortiz, A. (Eds.). (2002). *English language learners with special needs: Identification, placement, and instruction*. Washington, DC: Center for Applied Linguistics.

Artiles, A. J., Rueda, R., Salazar, J., & Higareda, I. (2005). Within-group diversity in minority disproportionate representation: English language learners in urban school districts. *Exceptional Children*, *7*(1), 283–300.

Beattie, S., Grise, P., & Algozzing, B. (1983). Effects of test modifications on the minimum competency of learning disabled students. *Learning Disabled Quarterly*, *6*, 75–77.

Bradley, R., Danielson, L., & Hallahan, D. P. (Eds.). (2002). *Identification of learning disabilities: Research to practice*. Hillsdale, NJ: Erlbaum.

Botel, M., & Granowsky, A. (1974). A formula for measuring syntactic complexity: A directional effort. *Elementary English*, *1*, 513–516.

Cattell, B. R., & Burdsal, A. C. (1975). The radial parcel double factoring design: A solution to the item-vs.-parcel controversy. *Multivariate Behavioral Research*, *10*, 165–179.

Celce-Murcia, M., & Larsen-Freeman, D. (1983). *The grammar book: An ESL/EFL teacher's book*. Rowley, MA: Newbury House.

Chall, J. S., Jacobs, V. S., & Baldwin, L. E. (1990). *The reading crisis: Why poor children fall behind*. Cambridge, MA: Harvard University Press.

UNCORRECTED PROOF

EMUS TCRE 782.PDF 06-Sep-06 20:45 148003 Bytes 22 PAGES n operator=HVAnanth)

Cocking, R. R., & Chipman, S. (1988). Conceptual issues related to mathematics achievement of language minority children. In R. R. Cocking & J. P. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics* (pp. 17–46). Hillsdale, NJ: Erlbaum.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*, 98–104.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology*, *20*, 405–438.

Dale, E., & Chall, J. S. (1948). A formula for predicting readability. *Educational Research Bulletin*, *27*, 11–20 *28*, 37–54.

De Corte, E., Verschaffel, L., & DeWin, L. (1985). Influence of rewording verbal problems on children's problem representations and solutions. *Journal of Educational Psychology*, 77, 460–470.

Forster, K. I., & Olbrei, I. (1973). Semantic heuristics and syntactic trial. *Cognition*, *2*, 319–347.

Gathercole, S. E., & Baddeley, A. D. (1993). *Working memory and language*. Hillsdale, NJ: Erlbaum.

Haiman, J. (1985). *Natural syntax: Iconicity and erosion*. New York: Cambridge University Press.

Halliday, M. A. K., & Martin, J. R. (1993). *Writing science: Literacy and discursive power*. Pittsburgh, PA: University of Pittsburgh Press.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels* (Research Rep. No. 3). Urbana, IL: National Council of Teachers of English.

Hunt, K. W. (1977). Early blooming and late blooming syntactic structures. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 69–90). Urbana, IL: National Council of Teachers of English.

Jenkins, J. R., & O'Connor, R. E. (2002). Chapter II: Early identification and intervention for young children with reading/learning disabilities. In R. Bradley, L. Danielson, & D. P. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 99–149). Hillsdale, NJ: Erlbaum.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixation to comprehension. *Psychological Review*, *87*, 329–354.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, *30*, 580–602.

Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review*, *92*, 109–129.

Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly*, *10*, 62–102.

Lemke, J. L. (1986). *Using language in classrooms*. Victoria, Australia: Deakin University Press.

Lepik, M. (1990). Algebraic word problems: Role of linguistic and structural variables. *Educational Studies in Mathematics*, *21*, 83–90.

Linn, R. L. (1995). *Assessment-based reform: Challenges to educational measurement*. Princeton, NJ: Educational Testing Service.

Linn, R. L., & Gronlund, N. E. (1995). *Measuring and assessment in teaching* (7th ed.). Englewood Cliffs, NJ: Prentice Hall.

MacDonald, M. C. (1993). The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language*, *32*, 692–715.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13–23.

Mestre, J. P. (1988). The role of language comprehension in mathematics and problem solving. In R. R. Cocking & J. P. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics* (pp. 200–220). Hillsdale, NJ: Erlbaum.

Munro, J. (1979). Language abilities and math performance. *Reading Teacher*, *32*, 900–915.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

UNCORRECTED PROOF

TCRE 782.PDF 06-Sep-06 20:45 148003 Bytes 22 PAGES n operator=HVAnanth)

National Center for Education Statistics. (2000). *Digest of education statistics, 2000*. Retrieved July 12, 2002, from http://nces.ed.gov/pubs2001/digest/dt053.html

New Jersey Department of Education. (2001). *New Jersey statewide assessment reports*. Retrieved July 11, 2002, from http://www.state.nj.us/njded/schools/achievement/2002/

New York State Education Department. (1999). *1999 pocketbook of goals and results for individuals with disabilities*. Retrieved July 11, 2002, from http://www.vesid.nysed.gov/pocketbook/1999/

Noonan, J. (1990). Readability problems presented by mathematics text. *Early Child Development and Care*, *54*, 57–81.

Orr, E. W. (1987). *Twice as less: Black English and the performance of black students in mathematics and science*. New York: W. W. Norton.

Ortiz, A. A. (2002). Prevention of school failure and early intervention for English language learners. In A. J. Artiles & A. A. Ortiz (Eds.), *English language learners with special education needs* (pp. 40–58). Washington, DC: Center for Applied Linguistics.

Pauley, A., & Syder, F. H. (1983). Natural selection in syntax: Notes on adaptive variation and change in vernacular and literary grammar. *Journal of Pragmatics*, *7*, 551–579.

Rothman, R. W., & Cohen, J. (1989). The language of math needs to be taught. *Academic Therapy*, *25*, 133–142.

Rueda, A., Artiles, A. J., Salazar, J., & Higareda, I. (2002). An analysis of special education as a response to the diminished academic achievement of Chicano/Latino students: An update. In R. R. Valenica (Ed.), *Chicano school failure and success: Past, present, and future* (2nd ed., pp. 310–332). London: Routledge/Falmer.

Salvia, J., & Ysseldyke, J. (1998). *Assessment*. Boston: Houghton Mifflin.

Schachter, P. (1983). *On syntactic categories*. Bloomington: Indiana University Linguistics Club.

Shavelson, R., & Webb, N. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Shuard, H., & Rothery, A. (Eds.). (1984). *Children reading mathematics*. London: J. Murray.

Slobin, D. I. (1968). Recall of full and truncated passive sentences in connected discourse. *Journal of Verbal Learning and Verbal Behavior*, *7*, 876–881.

Spanos, G., Rhodes, N. C., Dale, T. C., & Crandall, J. (1988). Linguistic features of mathematical problem solving: Insights and applications. In R. R. Cocking & J. P. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics* (pp. 221–240). Hillsdale, NJ: Erlbaum.

Thurlow, M., & Liu, K. (2001). *State and district assessments as an avenue to equity and excellence for English language learners with disabilities* (LEP Projects Report 2). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Tindal, G., Anderson, L., Helwig, R., Miller, S., & Glasgow, A. (2000). *Accommodating students with learning disabilities on math tests using language simplication*. Eugene: University of Oregon, Research, Consultation, and Teaching Program.

Trenholme, B., Larsen, S. C., & Parker, R. M. (1978). The effects of syntactic complexity upon arithmetic performance. *Learning Disabilities Quarterly*, *1*, 81–85.

Wang, M. D. (1970). The role of syntactic complexity as a determiner of comprehensibility. *Journal of Verbal Learning and Verbal Behavior*, *9*, 398–404.

Zehler, A. M., Hopstock, P. J., Fleischman, H. L., & Greniuk, C. (1994). *An examination of assessment of limited English proficient students*. Arlington, VA: Development Associates, Special Issues Analysis Center.

JAMAL ABEDI is a professor at the Graduate School of Education of the University of California, Davis, and a research partner at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). His interests include studies in the area of psychometrics focusing on the validity of assessment and accommodation for English language learners (ELL), and research on the opportunity to learn for ELLs.

UNCORRECTED PROOF

TCRE 782.PDF 06-Sep-06 20:45 148003 Bytes 22 PAGES n operator=HVAnanth)

Author Query Form

Journal	**TCRE**
Article	**782**

Dear Author,

During the copy-editing of your paper, the following queries arose. Please respond to these by marking up your proofs with the necessary changes/additions. Please write your answers clearly on the query sheet if there is insufficient space on the page proofs. If returning the proof by fax do not write too close to the paper's edge. Please remember that illegible mark-ups may delay publication.

Query No.	Description	Author Response
Q1	Please provide page range for Abedi, 2005 entry.	

[redacted]

From: Kimberly Sopko [mokoeduc8@cox.net]
Sent: Thursday, November 19, 2009 2:40 PM
To: Race To The Top Assessment Input
Subject: Race to the Top Assessment Program

Thank you for the opportunity to provide input to the Race to the Top AssessmentProgram. Below is a response to a highligted specific technical assessment question:

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)?

Due to ongoing budget shortfalls in education, it is imperative to examine all aspects of standards and assessments. Local education agencies and school principals struggle to design creative solutions to the challenges they face amid budget shortfalls. One creative solution suggested by a local principal is to loop several levels of elementary grades in order to keep class sizes reasonable while maintaining an appropriate staffing level that fits within a lower budget. For example, loop a grade 3 and grade 4 over two years and enable these students to receive all required knowledge for grades 3 and 4 over a two year period.

The greatest challenge to this looping is the current grade level standards and assessments that require an annual assessment of students' acquired knowledge and skills. Due to the nature of a looped class over a two year period, the majority of students would not be able to demonstrate acquired knowledge based on specific grade level standards on an annual basis. Therefore, it is imperative to incorporate FLEXIBILITY within an assessment program that will allow local education agencies and school principals to deliver assessments on a schedule to match their designed looping grade level programs. It would be possible for grade level standards to remain for each grade level, but an assessment schedule that allows for delivery of assessments upon the student's grade level loop completion that would occur every two years rather than annually is necessary.

The proposed framework for Design of Assessment Systems—General requirements indicates that *"the 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."* It is imperative to take this design one step further to allow for administration of summative component assessments over a two year period, rather than during a single school year.

There are many local education agencies and school principals who will develop creative solutions to ensure their students continue to receive a high quality education amidst the incredible shrinking budgets, but the federal and state government must afford them the flexibility to be creative. By considering options such as looping grade levels, government must also consider the option of looping standards and assessments as well.

Kimberly M. Sopko, Ed.D.
MOKO Educational Services

Burke, VA 22015
757-348-5051
mokoeduc8@cox.net

This email is intended for original recipient(s) only and is private and confidential.

Letter Report to the U.S. Department of Education on the Race to the Top Fund

Board on Testing and Assessment; National Research Council

ISBN: 0-309-14579-1, 19 pages, 8 1/2 x 11, (2009)

This free PDF was downloaded from:
http://www.nap.edu/catalog/12780.html

Visit the National Academies Press online, the authoritative source for all books from the National Academy of Sciences, the National Academy of Engineering, the Institute of Medicine, and the National Research Council:

- Download hundreds of free books in PDF
- Read thousands of books online, free
- Sign up to be notified when new books are published
- Purchase printed books
- Purchase PDFs
- Explore with our innovative research tools

Thank you for downloading this free PDF. If you have comments, questions or just want more information about the books published by the National Academies Press, you may contact our customer service department toll-free at 888-624-8373, visit us online, or send an email to comments@nap.edu.

This free book plus thousands more books are available at http://www.nap.edu.

Copyright © National Academy of Sciences. Permission is granted for this material to be shared for noncommercial, educational purposes, provided that this notice appears on the reproduced materials, the Web address of the online, full authoritative version is retained, and copies are not altered. To disseminate otherwise or to republish requires written permission from the National Academies Press.

THE NATIONAL ACADEMIES
Advisers to the Nation on Science, Engineering, and Medicine

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

Division of Behavioral and Social Sciences and Education
Board on Testing and Assessment

500 Fifth Street, NW
Washington, DC 20001
Phone: 202 334 2353
Fax: 202 334 1294
E-mail: bota1@nas.edu
www.nationalacademies.org

October 5, 2009

The Honorable Arne Duncan
Secretary of Education
U.S. Department of Education
400 Maryland Avenue, SW, Room 3W329
Washington, DC 20202

Dear Mr. Secretary:

This letter offers comments concerning the Department's Proposed Regulations on the Race to the Top (RTT) fund of the American Recovery and Reinvestment Act of 2009 (74 Fed. Reg. 37804, proposed July 29, 2009) from the Board on Testing and Assessment of the National Research Council. (See Attachment A for a list of members.) The comments reflect a consensus of the Board.

Under National Academies procedures, any letter report must be reviewed by an independent group of experts before it can be publicly released, which made it impossible to complete the letter within the public comment period of the *Federal Register* notice.[1] However, we hope that the Department will still find these comments helpful in revising the RTT plans.

The Board on Testing and Assessment stands ready to assist the federal government, Congress, and the states in addressing issues concerning the use of evidence to improve educational opportunities for the nation's young people.

Sincerely yours,

Edward Haertel

Edward H. Haertel, *Chair*
Board on Testing and Assessment

cc: Carmel Martin, Assistant Secretary for Planning, Evaluation, and Policy Development
Thelma Melendez, Assistant Secretary for Elementary and Secondary Education
Joanne Weiss, Senior Advisor to the Secretary and Director, Race to the Top
Marshall S. Smith, Senior Counselor to the Secretary, Director, International Affairs
John Q. Easton, Director, Institute of Education Sciences

[1] The reviewers for the report are acknowledged in Attachment B.

Copyright © National Academy of Sciences. All rights reserved.

COMMENTS ON THE DEPARTMENT OF EDUCATION'S PROPOSAL ON THE RACE TO THE TOP FUND

This letter offers comments concerning the Department's Proposed Regulations on the Race to the Top (RTT) fund of the American Recovery and Reinvestment Act (ARRA) of 2009 (74 Fed. Reg. 37804, proposed July 29, 2009) from the Board on Testing and Assessment (BOTA) of the National Research Council. (See Attachment A for a list of members.) The comments reflect a consensus of the Board.

The Board held an open session at its meeting on July 30, 2009, to discuss the importance of evaluating RTT spending. This meeting had been planned before the Department's posting in the *Federal Register*. In deciding to hold this session, BOTA hoped to ensure that the unusual opportunity offered by RTT to invest in educational innovations be fully exploited through evaluation of the supported programs. Although RTT represents a substantial amount of money in absolute terms, it is small in comparison with the total spending on education in the United States. It is only through careful evaluations from RTT-supported innovations that the investment in RTT is likely to be leveraged to support improvements that can affect the entire educational system. Without the learning that results from careful evaluation, any benefits of this one-time spending on innovation are likely to end when the funding ends.

Tests often play an important role in evaluating educational innovations, but an evaluation requires much more than tests alone. A rigorous evaluation plan typically involves implementation and outcome data that need to be collected throughout the course of a project. In addition, a rigorous evaluation plan may affect the way that participants are selected to be included in the project.

Originally, BOTA's July 30 meeting was planned to include several members of the Department, to allow a discussion of possible approaches for requiring and conducting evaluations of RTT innovations. However, as a result of the *Federal Register* filing on the preceding day, most of the Department's representatives were unable to attend BOTA's meeting. In addition, the availability of the Department's full proposal caused the Board to consider and discuss a number of other testing-related elements in the proposal. As a result, the Board concluded that it would be important and appropriate to communicate directly with the Department in the form of a letter, both to underline the importance of requiring evaluations for RTT-supported innovations and to address several other concerns about the use of tests suggested in the *Federal Register* notice.

The comments we offer in this letter are based on more than 15 years of work by BOTA on a wide range of topics concerning the use of testing and assessment. We draw from that body of work and from our collective knowledge about accepted practices in educational measurement to offer our comments here. (See Attachment C for a selected list of BOTA reports.)

We begin our comments with a general discussion about the role of testing and assessment in educational reform. This discussion underlines some important aspects of testing and assessment that are widely understood but not always considered in the design of educational policies that use tests. We then discuss six testing-related topics that are reflected in the Department's proposal in the *Federal Register*:

1. the need for states to include evaluations in their RTT program activities;
2. the use of NAEP to evaluate RTT outcomes;
3. the use of student growth data to evaluate teachers and principals;
4. the use of data to improve instruction;

Copyright © National Academy of Sciences. All rights reserved.

5. challenges in developing common assessments; and
6. challenges in creating longitudinal data systems.

We focus our comments on these six issues because we have concerns in each case about how they will be addressed and the resulting implications for the way that tests and assessments will be used in implementing education policy. Throughout, the Departments' specific proposed language is in italics, with the page numbers taken from the notice in the *Federal Register*.

RELIANCE ON TESTING AND ASSESSMENT IN EDUCATIONAL REFORM

The proposed regulations rely heavily on the use of testing and assessment both to drive and to measure education reform. BOTA strongly supports the Department's desire to incorporate objective measures of student performance to inform the process of educational reform. Standardized assessments have much to offer in this regard. BOTA has noted in *Lessons Learned* (p. 5):

> In many situations, standardized tests provide the most objective way to compare the performance of a large group of examinees across places and times.… Similarly, in K-12 education, statewide standardized tests are useful for comparing student achievement across classrooms, schools or districts, or across different times. When combined with other sources of information, these comparisons can help educators and policy makers decide how to target resources.

However, the *Lessons Learned* report (p. 5) also cautions that "A test score is an estimate rather than an exact measure of what a person knows and can do." The items on any test are a sample from some larger universe of knowledge and skills, and scores for individual students are affected by the particular questions included. A student may have done better or worse on a different sample of questions. In addition, guessing, motivation, momentary distractions, and other factors also introduce uncertainty into individual scores. When scores are averaged at the classroom, school, district, or state level, some of these sources of measurement error (e.g., guessing or momentary distractions) may average out, but other sources of error become much more salient. Average scores for groups of students are affected by exclusion and accommodation policies (e.g., for students with disabilities or English learners), retest policies for absentees, the timing of testing over the course of the school year, and by performance incentives that influence test takers' effort and motivation. Pupil grade retention policies may influence year-to-year changes in average scores for grade-level cohorts.

Moreover, test results are affected by the degree to which curriculum and instruction are aligned with the knowledge and skills measured by the test. Educators understandably try to align curriculum and instruction with knowledge and skills measured on high-stakes tests, and they similarly focus curriculum and instruction on the kinds of tasks and formats used by the test itself. For these reasons, as research has conclusively demonstrated, gains on high-stakes tests are typically larger than corresponding gains on concurrently administered "audit" tests, and sometimes they are much larger.[2] Improvements on the necessarily limited content of a high-states test may be offset by losses on other, equally valuable content that happens to be left untested.

[2] For references and further discussion of score inflation, see Koretz (2008, Chapter 10).

Copyright © National Academy of Sciences. All rights reserved.

These issues do not mean that test scores are unimportant or generally invalid. They do mean that test use should comport with relevant professional standards. A list of relevant professional and technical standards applicable to the use of educational assessments is included in Attachment D. Three core issues reflected in these documents are (1) the need to develop multiple measures of key outcomes, ideally using multiple assessment formats; (2) the need to validate these assessments for specific uses; and (3) the need to consider the populations involved and any associated validity and fairness issues. These documents also point to several technical issues (e.g., the reliability or precision of the measures, scaling and equating issues) and operational issues (e.g., data collection and processing) that can undermine the interpretability and usefulness of the test results if not handled appropriately. The documents provide guidelines on how to avoid these kinds of problems. In addition, a thorough account of the steps needed to improve the gathering and use of data by state assessment systems can be found in the BOTA report, *Systems for State Science Assessment* (2006).

We encourage the Department to pursue vigorously the use of multiple indicators of what students know and can do. A single test should not be relied on as the sole indicator of program effectiveness. This caveat applies as well to other targets of measurement, such as teacher quality and effectiveness and school progress in closing achievement gaps. Development of an appropriate system of multiple indicators involves thinking about the objectives of the system and the nature of the different information that different indicators can provide. Such a system should be constructed from a careful consideration of the complementary information that is provided by different measures.[3]

SPECIFIC AREAS OF CONCERN

Need for States to Include Evaluations in their RTT Program Activities

The Department specifically requests comment on whether states should be required to conduct evaluations of their support programs and indicates that the final notice for RTT applications will specify the nature of the evaluations that will be required:

> *The State and its participating LEAs must use funds under this program to participate in a national evaluation of the program, if the Department chooses to conduct one. In addition, the Department is seeking comment on whether a State should, instead of or in addition to a national evaluation, be required to conduct its own evaluation of its program activities using funds under this program. The Department will announce in the notice inviting applications the evaluation approach(es) that will be required. (Section II.D.a, p. 37808)*

We strongly affirm the importance of independent evaluation of all RTT initiatives. These evaluations should be conducted at all levels: national, state, and local. Such evaluations will help foster a culture of continuous improvement in the nation's educational systems. As part of the evaluation, all consequences—both positive and negative, and intended and unintended—should be carefully monitored and weighed.

At BOTA's July 30 meeting to discuss the Department's proposed regulations, there was strong support for rigorous evaluation requirements to help ensure both the success of the states'

[3] For a thoughtful description of the creation of such a system of multiple indicators, see Chester (2005).

Copyright © National Academy of Sciences. All rights reserved.

RTT policy and program implementations and to increase scientific understanding of effective school reform. Strong evaluation requirements can also help to establish a culture of evidence-based reform at the state and local levels, which will pay dividends long after the short-term RTT stimulus funding has ended.

As part of its discussion, BOTA considered the example of welfare reform in 1990s, which was informed by a history of state experimentation, in which states received waivers from federal requirements under the condition that they conduct evaluations of those experiments (Harvey, Camasso, and Jagannathan, 2000). The evaluation requirements from the federal government led to the scientific study of the impact of crucial features of welfare policy that were tried out in these state experiments. The RTT fund offers an analogous opportunity in education: a careful choice of evaluation requirements can result in substantial opportunities for research and learning about the educational innovations that are supported.

BOTA's discussion on evaluation focused on six key principles: (1) theory of action, (2) appropriateness of design, (3) documentation of implementation, (4) formative and summative components, (5) independent evaluators, and (6) valid outcome measures. These principles are commonly accepted as best practice in evaluation.[4] We briefly discuss each of these requirements below.

Theory of Action The design of each evaluation will benefit from a clear description by the applicant of the mechanism by which the reform initiative is expected to improve student learning outcomes. This "theory of action" will typically take the form of a connected set of propositions—a logical chain of reasoning that explains how the reform expenditures will lead to improved schooling practices. A theory of action "connects the dots," explaining in commonsense terms which program features are expected to produce which results in order to reach the final desired outcome. A theory of action gives shape to any evaluation plan. The evaluation is framed as an investigation of each link in the chain, focusing on those links for which there may be only limited prior evidence. If the evaluation demonstrates that the reform initiative failed to meet its objectives, the investigation of the individual links will facilitate the redesign of the reform initiative to make it effective.

For example, if the theory of action for an accountability program based on tests assumes that performance will improve because educators will use test results to refine their instruction in areas in which students perform poorly, then an evaluation of that accountability program should look at the ways that the tests are being used to influence teaching. If the test results are used to refocus the curriculum towards areas in which students perform poorly, that would confirm the result expected by the theory of action; in contrast, if the test results lead educators to shift their attention from teaching to test preparation, then that would not confirm the result expected by the theory of action and suggest the possibility of unintended consequences.

Appropriateness of Design There is no one best design for program or policy evaluation. Whatever design is chosen must be appropriate to the program or policy being evaluated. A strong design will typically involve the collection of both quantitative and qualitative information. The choice of a design should be guided by the theory of action of the intervention or activity, by best practices in the field, and by a clear sense of the purpose of the evaluation: that is, what questions are to be answered and how they are to be prioritized. An evaluation

[4] For further information, see *The Program Evaluation Standards* (Joint Committee on Standards for Educational Evaluation, 1994) listed in Attachment D.

Copyright © National Academy of Sciences. All rights reserved.

should be designed before the program is implemented so there is an opportunity to collect baseline data and for the evaluation plan to influence crucial decisions about the selection of participants for the program and the types of implementation and outcome information to be collected.

Documentation of Implementation The evaluation must include documentation of the extent to which the policy or program was in fact implemented as intended. Failure to implement a program adequately can be one explanation for a program's failure to produce the desired effect. Any significant problems encountered along the way should be documented so that future efforts can profit by lessons learned from those problems. Implementation evaluation can also help in spotting potential negative, unintended effects of reforms.

Formative and Summative Components Ideally, evaluation plans will provide for both short-term, "formative" feedback for program improvement, and longer-term, "summative" information to judge program impact. The formative component of an evaluation can be relatively informal, featuring quick turnaround, so that the local educators or state officials in charge of the program can adjust implementation and fine-tune policies as needed. Formative evaluation identifies ways to adjust or improve program implementation in real time. A summative evaluation is more formal, longer term, and typically requires more advance planning. This kind of evaluation does not influence the implementation of the policy or program; rather, it documents the extent of intended and unintended outcomes, using a rigorous research design and statistical methods.

For RTT initiatives, both formative and summative evaluation components should be included. Both components will benefit from documentation about implementation. When possible, summative evaluations should be designed in ways that provide generalizable knowledge to guide future education reform.

Independent Evaluators Evaluations should employ independent and well-qualified external evaluators. The field of evaluation is well developed, and much is known about sound evaluation practice that should be incorporated in an evaluation. Outside (external) evaluations will be more credible than evaluations conducted in-house by the states or LEAs themselves.

Valid Outcome Measures Evaluations must use student outcome measures that are valid for the intended use. Outcomes must reflect intended program or policy goals. An otherwise sound evaluation may be undermined by reliance on a test that is poorly aligned with the intervention. An outcome measure should not be chosen simply because it is inexpensive or readily available. Even a test's alignment to content standards, as typically implemented, may be an insufficient criterion for the choice of an outcome measure. Alignment of a test to a state's academic content standards typically means that the items on the test can each be matched to one or more specific content standards, in accordance with the test specification. But the standards are often much broader and more complex than any of the corresponding items. Taken together, the items representing a standard may be a pale reflection of the intent of that standard. Moreover, some standards are typically omitted altogether from the test specification. For these reasons, "alignment," as generally implemented, is not sufficient to ensure that a test is appropriate as an outcome measure for an evaluation.

Copyright © National Academy of Sciences. All rights reserved.

Letter Report to the U.S. Department of Education on the Race to the Top Fund
http://www.nap.edu/catalog/12780.html

Use of NAEP to Evaluate RTT Outcomes

Although BOTA is a strong advocate for the importance of NAEP in measuring U.S. educational progress, NAEP cannot play a primary role in evaluating RTT initiatives, a role that might be mistakenly inferred from the language in the Department's proposal.

> *We propose using the NAEP to monitor overall increases in student achievement and decreases in the achievement gap over the course of this grant because the NAEP provides a way to report consistently across Race to the Top grantees as well as within a State over time. . . .* (*Section I, footnote 1, p. 37805*)

It is necessary to be clear about the distinction between the requirements of an evaluation and the kind of "monitoring" that NAEP can provide. For the purposes of evaluating RTT initiatives, there are at least four critical limitations with regard to inferences that can be drawn from NAEP.

1. NAEP is intended to survey the knowledge of students across the nation with respect to a broad range of content and skills: it was not designed to be aligned to a specific curriculum. Because states differ in the extent to which their standards and curricula overlap with the standards assessed by NAEP, it is unlikely that NAEP will be able to fully reflect improvements taking place at the state level.[5]
2. Although NAEP can provide reliable information for states and certain large school districts, it cannot, as presently designed, fully reflect the effects of any interventions targeted at the local level or on a small portion of a state's students, such as are likely to be supported with RTT initiatives.
3. States are likely to undertake multiple initiatives under RTT, and NAEP results, even at the state level, cannot disaggregate the contributions of different RTT initiatives to state educational progress.
4. The specific grade levels included in NAEP (grades 4, 8, and 12) may not align with the targeted populations for some RTT interventions.

Consequently, NAEP will be of limited value in judging the success or failure of individual initiatives under RTT, even at the state level. The availability of NAEP does not in any way obviate the need to plan rigorous evaluations—at national, state, and local levels—that are appropriately designed to assess the implementation and outcomes of RTT initiatives.

These concerns do not diminish the importance of NAEP as an independent measure of educational progress at the state and national levels. NAEP has served and should continue to serve an essential "audit" function for evaluating states' reform efforts, using objective measurements that are comparable across states and over time. In order that it continue to serve that function, it is critically important that high-stakes decisions *not* be attached directly to NAEP results. What makes NAEP so valuable is precisely the fact that there are *no* incentives attached to it that might lead educators to "teach to the test." Consequently, NAEP is relatively immune to the pressures of score inflation that can distort any high-stakes measure. NAEP can

[5] Continued progress toward common standards across the states, as expected by the Department's proposal, could lead to a shift in NAEP standards to align more closely with those emerging state standards. However, that will not have occurred during the period of RTT initiatives.

Copyright © National Academy of Sciences. All rights reserved.

Letter Report to the U.S. Department of Education on the Race to the Top Fund
http://www.nap.edu/catalog/12780.html

only be useful for the monitoring function envisioned in the regulations to the extent that it is not used for any high-stakes decision.

Use of Student Growth Data to Evaluate Teachers and Principals

We applaud the Department's desire to move educational data collection systems forward by supporting and encouraging the construction of data systems that can link students and their teachers.

> *We propose that to be eligible under this program, a State must not have any legal, statutory, or regulatory barriers to linking student achievement or student growth data to teachers for the purpose of teacher and principal evaluation.* (*Section II.A., p. 37806*)

Such data systems are essential for conducting research related to the full range of potential approaches for evaluating educators and for developing pilot programs for evaluation approaches that will one day become operational. However, BOTA has significant concerns that the Department's proposal places too much emphasis on measures of growth in student achievement (1) that have not yet been adequately studied for the purposes of evaluating teachers and principals and (2) that face substantial practical barriers to being successfully deployed in an operational personnel system that is fair, reliable, and valid.

> *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)*

The most prominent way to measure student growth involves "value added" approaches that use results from earlier tests, as well as other information, to adjust new test scores for pre-existing differences across students. The objective of these statistical techniques is to produce a measure of the "value added" to a student's achievement by a teacher or a school in a given year.

The term "value-added model" (VAM) has been applied to a range of approaches, varying in their data requirements and statistical complexity. Although the idea has intuitive appeal, a great deal is unknown about the potential and the limitations of alternative statistical models for evaluating teachers' value-added contributions to student learning. BOTA agrees with other experts who have urged the need for caution and for further research prior to any large-scale, high-stakes reliance on these approaches (e.g., Braun, 2005; McCaffrey and Lockwood, 2008; McCaffrey et al., 2003).

BOTA and the National Academy of Education conducted a joint workshop (November 13-14, 2008) to obtain expert judgments and assessments of issues related to the use of value-added methodologies in education. The workshop topics included both the potential utility of VAM approaches and the limitations of VAM techniques, particularly in making high-stakes

Copyright © National Academy of Sciences. All rights reserved.

individual, institutional, or district accountability decisions. (A report of the workshop is forthcoming; the workshop papers are available at www7.nationalacademies.org/bota.) The considerable majority of experts at the workshop cautioned that although VAM approaches seem promising, particularly as an additional way to evaluate teachers, there is little scientific consensus about the many technical issues that have been raised about the techniques and their use.

Prominent testing expert Robert Linn concluded in his workshop paper: "As with any effort to isolate causal effects from observational data when random assignment is not feasible, there are reasons to question the ability of value-added methods to achieve the goal of determining the value added by a particular teacher, school, or educational program" (Linn, 2008, p. 3). Teachers are not assigned randomly to schools, and students are not assigned randomly to teachers. Without a way to account for important unobservable differences across students, VAM techniques fail to control fully for those differences and are therefore unable to provide objective comparisons between teachers who work with different populations. As a result, value-added scores that are attributed to a teacher or principal may be affected by other factors, such as student motivation and parental support.

VAM also raises important technical issues about test scores that are not raised by other uses of those scores. In particular, the statistical procedures assume that a one-unit difference in a test score means the same amount of learning—and the same amount of teaching—for low-performing, average, and high-performing students. If this is not the case, then the value-added scores for teachers who work with different types of students will not be comparable. One common version of this problem occurs for students whose achievement levels are too high or too low to be measured by the available tests. For such students, the tests show "ceiling" or "floor" effects and cannot be used to provide a valid measure of growth. It is not possible to calculate valid value-added measures for teachers with students who have achievement levels that are too high or too low to be measured by the available tests.

In addition to these unresolved issues, there are a number of important practical difficulties in using value-added measures in an operational, high-stakes program to evaluate teachers and principals in a way that is fair, reliable, and valid. Those difficulties include the following:

1. Estimates of value added by a teacher can vary greatly from year to year, with many teachers moving between high and low performance categories in successive years (McCaffrey, Sass, and Lockwood, 2008).
2. Estimates of value added by a teacher may vary depending on the method used to calculate the value added, which may make it difficult to defend the choice of a particular method (e.g., Briggs, Weeks, and Wiley, 2008).
3. VAM cannot be used to evaluate educators for untested grades and subjects.
4. Most data bases used to support value-added analyses still face fundamental challenges related to their ability to correctly link students with teachers by subject.
5. Students often receive instruction from multiple teachers, making it difficult to attribute learning gains to a specific teacher, even if the data bases were to correctly record the contributions of all teachers.
6. There are considerable limitations to the transparency of VAM approaches for educators, parents and policy makers, among others, given the sophisticated statistical methods they employ.

Copyright © National Academy of Sciences. All rights reserved.

Many of these difficulties could be addressed in time—with further research and development of VAM statistical approaches, expansion of testing programs into more grades and subjects, improvement of data bases, and careful development of personnel evaluation systems that use multiple measures. However, it is unlikely that any state at this time could make a proposal for using VAM approaches in an operational program for teacher or principal evaluation that adequately addresses all of these concerns.

The use of test data for teacher and educator evaluation requires the same types of cautions that are stressed when test data are used to evaluate students: "Tests are one objective and efficient way to measure what people know and can do, and they can help make comparisons across large groups of people. However, test scores are not perfect measures: they should be considered with other sources of information when making important decisions about individuals" (*Lessons Learned*, p. 15). This caution is even more important when applied to complex statistics—like value-added analyses—derived from tests.

In sum, value-added methodologies should be used only after careful consideration of their appropriateness for the data that are available, and if used, should be subjected to rigorous evaluation. At present, the best use of VAM techniques is in closely studied pilot projects. Even in pilot projects, VAM estimates of teacher effectiveness should not be used as the sole or primary basis for making operational decisions because the extent to which the measures reflect the contribution of teachers themselves, rather than other factors, is not understood. Even in pilot projects, VAM estimates of teacher effectiveness should not be used to make operational decisions for teachers with students who have achievement levels that are too high or too low to be measured by the available tests because the estimates for such teachers will be essentially meaningless. Even in pilot projects, VAM estimates of teacher effectiveness that are based on data for a single class of students should not used to make operational decisions because such estimates are far too unstable to be considered fair or reliable.

Use of Data to Improve Instruction

We applaud the Department's support of "instructional improvement systems" that include the use of assessments, as well as student work, to help teachers focus instruction and evaluate the success of their instructional efforts. It is appropriate that the Department has included the use of such systems in its criteria for evaluating state proposals for the use of RTT funds.

> *The extent to which the State, in collaboration with its participating LEAs, has a high-quality plan to-- (i) Increase the use of instructional improvement systems (as defined in this notice) that provide teachers, principals, and administrators with the information they need to inform and improve their instructional practices, decision-making, and overall effectiveness. (Section III.B.(B)(3), p. 37809)*

The choice of appropriate assessments for use in instructional improvement systems is critical. Because of the extensive focus on large-scale, high-stakes, summative tests, policy makers and educators sometimes mistakenly believe that such tests are appropriate to use to provide rapid feedback to guide instruction. This is not the case.

Tests that mimic the structure of large-scale, high-stakes, summative tests, which lightly sample broad domains of content taught over an extended period of time, are unlikely to provide the kind of fine-grained, diagnostic information that teachers need to guide their day-to-day instructional decisions. In addition, an attempt to use such tests to guide instruction encourages a

Copyright © National Academy of Sciences. All rights reserved.

narrow focus on the skills used in a particular test—"teaching to the test"—that can severely restrict instruction. Some topics and types of performance are more difficult to assess with large-scale, high-stakes, summative tests, including the kind of extended reasoning and problem-solving tasks that show that a student is able to apply concepts from a domain in a meaningful way. The use of high-stakes tests already leads to concerns about narrowing the curriculum towards the knowledge and skills that are easy to assess on such tests; it is critical that the choice of assessments for use in instructional improvement systems not reinforce the same kind of narrowing.

In applying its criteria to evaluate state proposals for RTT funds, BOTA urges the Department to clarify that assessments that simply reproduce the formats of large-scale, high-stakes, summative tests are not sufficient for instructional improvement systems. The multiple-choice format in particular lends itself more easily to measuring declarative knowledge than complex "higher-order" cognitive skills. Instructional improvement systems that rely heavily on such item formats may reinforce a tendency to narrow instruction to reflect little more than tested content and formats.

In addition, BOTA urges a great deal of caution about the nature of assessments that could meet the Department's definition for inclusion in a "rapid-time" turnaround system.

> *Rapid-time, in reference to reporting and availability of school- and LEA-level data, means that data is available quickly enough to inform current lessons, instruction, and related supports; in most cases, this will be within 72 hours of an assessment or data gathering in classrooms, schools, and LEAs. (Section IV, p. 37811)*

If the Department is referring to informal, classroom assessment methods that can be scored and interpreted by the teacher, a 72-hour turnaround is a reasonable goal. It is important to provide teachers with a better understanding about the types of assessment they can use informally in their classrooms.

However, if the Department is referring to more formal assessment methods, then a 72-hour turnaround is very difficult to attain. Even for multiple-choice items, it would be hard to provide a 72-hour turnaround. Assessment of complex reasoning and problem-solving skills typically demands assessment formats that require students to generate their own extended responses rather than selecting a word or phrase from a short list of options. Automated scoring of such "constructed responses" is an active field of psychometric research: to date, however, except for automated scoring of written essays, the "state of the art" extends only to small demonstration projects not large-scale applications. Likewise, the logistics of human scoring for large-scale assessments would make a 72-hour turnaround extremely costly. It is premature to specify such a required period of time within which assessment results must be returned to teachers. When seeking to provide quick turnaround of information, it is essential that the quality of that information not be sacrificed in the process. The first priority must be that the right information is produced and that the information meets professional standards for technical adequacy so that the information can accurately guide decision making.

Copyright © National Academy of Sciences. All rights reserved.

Challenges in Developing Common Assessments

We applaud the Department's plan to invest in state consortia that will work toward developing high quality assessments. It is appropriate that the Department has included the commitment toward improving the quality of state assessment in its criteria for evaluating state proposals for the use of RTT funds.

> *Whether the State has demonstrated a commitment to improving the quality of its assessments by participating in a consortium of States that is working toward jointly developing and implementing common, high-quality assessments (as defined in this notice) aligned with the consortium's common set of K-12 standards (as defined in this notice) that are internationally benchmarked and that build toward college and career readiness by the time of high school graduation, and the extent to which this consortium includes a significant number of States. (Section III.A.(A)(2), p. 37808)*

The opportunity of working toward common assessments of common standards raises the possibility of significant economies of scale in the development of high-quality assessments.

Although BOTA supports the value of a joint development effort toward common assessments, we want to stress that the requirements are quite high for producing common assessments that would truly allow comparisons in student achievement across states in the same way that NAEP currently does. BOTA's previous work on potential approaches to developing common standards and assessments (*Uncommon Measures*, 1999; *Embedding Questions*, 1999) concluded that this aspiration is very difficult to accomplish in practice. The fundamental problem relates to dissimilarities across states in their standards, instruction and curriculum, and uses of test scores, as well as the assessments themselves: these dissimilarities ultimately make assessment results incomparable.

If states indeed adopt fully common standards and develop common assessments, these concerns would be reduced, but even seemingly small deviations from fully common standards and assessments will introduce important limitations in the degree of comparability of outcomes. For instance, to be fully comparable, the assessments would need to be administered under standardized conditions across the country. This means that time limits, the length of testing sessions, the instructions given to test takers, test security provisions, and the use of manipulatives and calculators would need to be the same everywhere. The test administration would need to be scheduled at an appropriate time during the school year such that all the students who will take the test have been exposed to and have an opportunity to learn the content covered on the test. The stakes attached to the assessment results would also need to be constant across the country to ensure that students are equally motivated to perform well. Including state-specific content on the assessments—that is, content that differs across states—introduces the issue of context effects. These state-specific items would need to be placed on the assessment in a location where they would not influence how students perform on the common items. See *Embedding Questions* (1999, pp. 20-36) for a fuller discussion.

In addition to the aspiration of creating common assessments, the Department's proposal also notes the objective of creating assessments that are "internationally benchmarked." There are several different ways this phrase might be interpreted. However, for assessment results that could be directly compared to their international counterparts, we note that the difficulties that arise in comparing test results from different states apply even more strongly for comparing test

Copyright © National Academy of Sciences. All rights reserved.

results from different countries. For making comparisons internationally, the problems with differing standards, assessments, instruction, curricula, and testing regimes are magnified. In addition, international test comparisons raise difficult problems related to language translation. Because of these challenges, the Department should think carefully about the kind of "international benchmarking" that it wants to encourage states to pursue.

The aspiration for higher quality in standards and assessments is one that BOTA members share, and over the coming months, BOTA will be conducting a series of workshops on improving state assessment systems, with particular attention to the opportunities offered by the current interest in moving toward common assessments and by the ARRA funding set aside for this purpose.

Challenges in Creating Longitudinal Data Systems

We applaud the Department's desire to support statewide longitudinal data systems. It is appropriate that the Department has included the commitment toward development such a system in its plan for evaluating state proposals for the use of RTT funds.

> *The extent to which the State has a high-quality plan to ensure that data from the State's statewide longitudinal data system are accessible to, and used to inform and engage, as appropriate, key stakeholders (e.g., parents, students, teachers, principals, LEA leaders, community members, unions, researchers, and policymakers); that the data support decision-makers in the continuous improvement of instruction, operations, management, and resource allocation. (Section III,B,(B)(2), p. 37809)*

The development and application of such longitudinal data systems are considerably more difficult than many people realize. State proposals for RTT funds should indicate a familiarity with the full complexity of the challenges and describe appropriate plans to begin to address them. The workshop on VAM approaches discussed above highlighted the challenges associated with creating data systems that produce understandable and useful information for all stakeholders, including the challenge of accurately connecting students to their teachers. An ongoing set of evaluation activities, as described above, should inform states' efforts in this regard. Another upcoming joint report from BOTA and the National Academy of Education on calculating dropout and graduation rates will also offer advice on building longitudinal data systems and making use of the data to inform decision making. Issues such as dealing with transfer students and students who drop out, some of whom re-enroll and drop out repeatedly, pose challenges in building data systems that accurately code and track students' status. We hope to address the qualities for best practices in state assessment systems in the workshop series we will be conducting over the coming year.

CONCLUSION

In closing, we return to the beginning of this letter, with the importance of rigorously evaluating the innovations supported by RTT funds. Careful evaluation of this spending should not be seen as optional; it is likely to be the only way that this substantial investment in educational innovation can have a lasting impact on the U.S. education system. BOTA urges the Department to carefully craft a set of requirements for rigorous evaluation of all initiatives support by RTT funds.

Copyright © National Academy of Sciences. All rights reserved.

Attachment A	Membership of the Board on Testing and Assessment
Attachment B	Reviewer Acknowledgments
Attachment C	Selected Reports of the Board on Testing and Assessment
Attachment D	Key Professional and Technical Standards Documents Relevant to the Use of Educational Assessments
Attachment E	References

Copyright © National Academy of Sciences. All rights reserved.

Letter Report to the U.S. Department of Education on the Race to the Top Fund
http://www.nap.edu/catalog/12780.html

Attachment A: Membership of the Board on Testing and Assessment

Edward H. Haertel (*Chair*), School of Education, Stanford University
Lyle Bachman, Department of Applied Linguistics, University of California, Los Angeles
Stephen Dunbar, College of Education, University of Iowa
David J. Francis, Department of Psychology, University of Houston
Arthur Goldberger[6], Department of Economics, Emeritus, University of Wisconsin–Madison
Michael Hout, Graduate Group in Sociology and Demography, University of California, Berkeley
Michael Kane, Research and Development Division, Educational Testing Service
Kevin Lang, Department of Economics, Boston University
Michael Nettles, Policy Evaluation and Research Center, Educational Testing Service
Diana Pullin, Lynch School of Education, Boston College
Brian Stecher, Education Program, RAND
Mark R. Wilson, Graduate School of Education, University of California, Berkeley
Rebecca Zwick, Gevirtz Graduate School of Education, University of California, Santa Barbara

Stuart W. Elliott, Board Director
Judith Anderson Koenig, Senior Program Officer

[6] Was unable to participate in BOTA's Committee to Respond to the Department of Education RTT Proposal.

Copyright © National Academy of Sciences. All rights reserved.

Letter Report to the U.S. Department of Education on the Race to the Top Fund
http://www.nap.edu/catalog/12780.html

Attachment B: Reviewer Acknowledgments

This report has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the National Research Council's Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We wish to thank the following individuals for their review of this report: Jonathan G. Dings, Office of Planning and Assessment, Boulder Valley School District; Mark Dynarski, Center for Improving Research Evidence, Mathematica Policy Research, Inc.; Margaret E. Goertz, Graduate School of Education, University of Pennsylvania; Jane Hannaway, Education Policy Center, The Urban Institute; Scott F. Marion, Office of the Associate Director, National Center for the Improvement of Educational Assessment; Lorraine McDonnell, Department of Political Science, University of California, Santa Barbara; and Lauress (Laurie) L. Wise, Human Resources Research Organization (HumRRO), Monterey, CA.

Although the reviewers listed above have provided many constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations, nor did they see the final draft of the report before its release. The review of this report was overseen by and Robert L. Linn, Professor Emeritus, Department of Education, University of Colorado, and Stephen E. Fienberg, Department of Statistics, Carnegie Mellon University. Appointed by the National Research Council, they were responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the authoring committee and the institution.

Copyright © National Academy of Sciences. All rights reserved.

Attachment C: Selected Reports of the Board on Testing and Assessment

NOTE: All of the reports listed here were or will be published by the National Academies Press.

Assessing Accomplished Teaching: Advanced-Level Certification Programs (2008)
Assessment in Support of Instruction and Learning: Bridging the Gap Between Large-Scale and Classroom Assessment (2003)
Educating One and All: Students with Disabilities and Standards-Based Reform (1997)
Embedding Questions: The Pursuit of a Common Measure in Uncommon Tests (1999)
Evaluation of the Voluntary National Tests: Year 1 and Year 2 (1999)
Getting Value Out of Value-Added: Report of a Workshop (forthcoming, 2009)
Grading the Nation's Report Card: Evaluating NAEP and Transforming the Assessment of Educational Progress (1999)
High Stakes: Testing for Tracking, Promotion, and Graduation (1999)
Incentives and Test-Based Accountability in Education (forthcoming, 2010)
Keeping Score for All: The Effects of Inclusion and Accommodation Policies on Large-Scale Educational Assessment (2004)
Knowing What Students Know: The Science and Design of Educational Assessment (2001)
Lessons Learned About Testing: Ten Years of Work at the National Research Council (no date)
Measuring High School Graduation and Dropout Rates: Next Steps for Research and Policy (forthcoming, 2010)
Measuring Literacy: Performance Levels for Adults (2005)
Myths and Tradeoffs: The Role of Tests in Undergraduate Admissions (1999)
NAEP Reporting Practices: Investigating District-Level and Market-Basked Reporting (2001)
Systems for State Science Assessment (2005)
Testing, Teaching, and Learning: A Guide for States and School Districts (1999)
Testing Teacher Candidates: The Role of Licensure Tests in Improving Teacher Quality (2001)
Uncommon Measures: Equivalence and Linkage Among Educational Tests (1999)
Understanding Dropouts: Statistics, Strategies, and High-Stakes Testing (2001)

Copyright © National Academy of Sciences. All rights reserved.

Attachment D: Key Professional and Technical Standards Documents Relevant to the Use of Educational Assessments

The Personnel Evaluation Standards: How to Assess Systems for Evaluating Educators (1988). Joint Committee on Standards for American Educational Research Educational Evaluation. Thousand Oaks, CA: Corwin Press.

Position Statement of the American Educational Research Association Concerning High-Stakes Testing in PreK-12 (2000). American Educational Research Association, Washington, DC

The Program Evaluation Standards: How to Assess Systems for Evaluating Educators. 2nd Edition (1994). Joint Committee on Standards for Educational Evaluation. Thousand Oaks, CA: Corwin Press.

Standards for Educational Accountability Systems (2002). Baker, E., Linn, R.L., Herman, J. and Koretz, D. CRESST Policy Brief 5. Available through the Educational Resources Information Center (ERIC).

Standards for Educational and Psychological Testing (1999). American Educational Research Association, American Psychological Association, National Council on Measurement in Education (AERA/APA/NCME), Washington, DC.

The Student Evaluation Standards: How to Improve Evaluations of Students (2003). Joint Committee on Standards for Educational Evaluation. Thousand Oaks, CA: Corwin Press.

Copyright © National Academy of Sciences. All rights reserved.

Attachment E: References

Braun, H. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: Educational Testing Service.

Briggs, D.C., Weeks, J.P., and Wiley, E. (2008). *The sensitivity of value-added modeling to the creation of a vertical score scale*. Paper presented at the National Conference on Value-Added Modeling, University of Wisconsin–Madison, April 22-24.

Chester, M. (2005). Making valid and consistent inferences about school effectiveness from multiple measures. *Educational Measurement: Issues and Practice, 24*(4), 40-52.

Harvey, C., Camasso, M.J., and Jagannathan, R. (2000). Evaluating welfare reform waivers under section 1115. *Journal of Economic Perspectives, 14*:4, 165-188.

Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Linn, R.L. (2008). *Measurement issues associated with value-added models*. Paper prepared for the workshop of the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, National Research Council, Washington, DC, November 13-14. Available: http://www7.nationalacademies.org/bota/1VAM_Workshop_Agenda.html [September 2009].

McCaffrey, D., and Lockwood, J.R. (2008). *Value-added models: Analytic issues*. Paper prepared for the workshop of the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, National Research Council, Washington, DC, November 13-14. Available: http://www7.nationalacademies.org/bota/1VAM_Workshop_Agenda.html [September 2009].

McCaffrey, D., Lockwood, J.R., Koretz, D.M., and Hamilton, L.S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND Corporation.

McCaffrey, D., Sass., T.R., and Lockwood, J.R. (2008). *The intertemporal stability of teacher effect estimates*. Paper presented at the National Conference on Value-Added Modeling, University of Wisconsin–Madison, April 22-24.

Copyright © National Academy of Sciences. All rights reserved.

Assessments for Students with Disabilities[1]

Martha L. Thurlow
University of Minnesota

In the early 1990s, I jumped head first into thinking about statewide assessment systems that included students with disabilities. This was after a considerable time studying instructional interventions and dropout prevention strategies for students with disabilities of all ages – an endeavor that was minimally successful.

In the early 1990s, most states included 10% or fewer of their students with disabilities in state assessments. Participation and accommodation policies were either non-existent or limiting. NAEP had less than 50% of students with disabilities participating in its assessments, and accommodations were not allowed. Only two states had developed or were developing an inclusive assessment system, one that assessed students with the most significant cognitive disabilities as well as those with other disabilities.

I am not going to review the tremendous strides that we have made in including students with disabilities, and in better assessing them. Today, nearly all states have more than 95% of their students with disabilities in assessments, supported in all states by written accommodation policies.

---

[1] Opinions presented in this paper are those of the author, and do not necessarily reflect those of any of her funding sources, including the U.S. Department of Education or Offices within it.

I think that it is generally recognized what a difference the inclusion of all students with disabilities has made in improving their education, in increasing our understanding of who these students are, and in improving the design of our assessment systems. I know for sure that including students with disabilities in large-scale assessments has had more of an effect than any of my research and development efforts to improve instruction and educational outcomes for students with disabilities.

This does not mean that the inclusion of students with disabilities in assessments has been easy. Because of a history of restricted access to the general education curriculum, these students are not performing well. The assessments, in general, are showing us the reality of being a student with a disability in U.S. schools – these students are often the lowest of the low – in getting access to the curriculum and then, in their performance. Of course, this is not true of all students with disabilities – their performance covers the range of performance of students in schools. I recall districts and states being startled in the mid 1990s when they found that adding students with disabilities to their assessment systems actually increased overall performance levels. **It is irresponsible to assume that because a student has a disability, that student is a low performer who cannot learn.**

So, when I think about the opportunity that is provided by the Race to the Top funds to support the development and administration of the next generation of assessments, assessments that are valid and reliable, and that enable students to demonstrate their knowledge and skills in core academic areas, I reflect on all that we have learned during

the past decade. We have learned a lot – and the Race to the Top funds can ensure that we build on that learning rather than step back from it. Too often, we have been retrofitting our assessments for students with disabilities. We should not allow that to happen with these funds!

I have identified several requirements that I think are fundamental as consortia of states develop the next generation of assessments. I have considered the diversity of students with disabilities in proposing these, and have also thought about the potential of technology to be inclusive of more students.

Design Assessments with the Assumption that Students with Disabilities are General Education Students First

First, the construction of an assessment system that provides valid and reliable results for students with disabilities must start by recognizing that students with disabilities are general education students first. This statement should apply regardless of the disability, or the diversity of the students. Jacqui will talk about (has talked about) students with significant cognitive disabilities. I include them here, but will not focus specifically on them. But, I do remind all that given the nature of the disabilities that students have, which do vary a lot, we still should expect 80% of all students with disabilities to meet the same achievement standards as students without disabilities.

Starting the development of assessments from the recognition that students with disabilities are general education students first has several implications. The major

implication is that **assessments should be better designed from the beginning with all students in mind**, but should not be different for most students with disabilities. Whether we call these universally designed assessments or accessible assessments, it is critical that during the development process we think of all students, clearly define what each assessment is intended to measure, and how that content can be measured for all students.

The research base for developing accountability assessments that are more appropriate for all students has dramatically increased in the past several years. To summarize this research, the National Center on Educational Outcomes has developed principles for assessments used for accountability, identifying 5 major principles:

1) All students are included in assessments in ways that hold schools accountable for their learning.
2) Assessments allow all students to show their knowledge and skills on the same challenging content.
3) High quality decision making determines how students participate.
4) Public reporting includes the assessment results of all students.
5) Accountability determinations are affected in the same way by all students.
6) Continuous improvement, monitoring, and training ensure the quality of the overall system.

Each of these is supported by specific characteristics of assessment systems that are appropriate for all students, including students with disabilities. All together, they provide an important framework for any future assessment system.

These principles are just one example of a whirlwind of research and development activity to support better assessments for every student, including students with disabilities.

A second fundamental requirement is that, as part of the assessment development process, allowable accommodations should be defined. To the extent that content targets are intertwined, such as decoding and understanding text, approaches should be developed that separate those. Too often, students with disabilities are penalized because their disability is in an area that is assessed.

For example, the ability to decode is a separate skill from the ability to understand text and draw inferences from it. Students may be poor decoders but fine comprehenders. We cannot tell that from many of our assessments.

There is more and more research on the effects of accommodations. One of the critical findings of that research is that the definition of the content standards – what we are really trying to measure – is of critical importance to accommodations. With explicit definitions of content standards – ideally ones on which all agree – we can more easily develop assessments on those standards, and we can figure out how to increase the accessibility of the assessments and the accommodations that may still be needed by some students. **Clear, grade-level content standards** are critical to the development

of assessments appropriate for the greatest number of students, including students with disabilities.

The National Accessible Reading Assessment Projects, a group of three projects focused on the development of accessible reading assessments for students with disabilities have developed a set of accessibility principles. One principle states: “Reading assessments are grounded in a definition of reading that is composed of clearly specified constructs, informed by scholarship, and attuned to accessibility concerns.” Ensuring that common standards have addressed accessibility concerns does NOT mean lowering the standards, but it does mean, for example, providing a way for students who cannot hear to demonstrate their “listening” skills.

The National Accessible Reading Assessment projects have identified several critical principles, with supporting implementation guidelines, to help ensure that reading assessments are appropriate for all students, including students with disabilities. Again, these principles should be considered as a framework for future assessments.

A third fundamental requirement is that the assessment system must include all students – there can be no exceptions. We know from past history and research that when any group of students is excluded from the assessment system, there is always the next group of students on the precipice of being eliminated.

In the early 1990s, research demonstrated that when students were excluded from the assessment system, they were denied quality instruction that they should have received. And, soon students who originally were not in the group that could be excluded were pushed into that group, then also excluded from assessments and quality instruction. The group of excluded students became larger and larger – not just those students with disabilities but also students who previously were not identified as having disabilities. This point cannot be emphasized too much – having an "exception" for some students who are not included in assessments and in high expectations for instruction will result in increasing numbers of students being identified regardless of whether they were among the intended group in the first place!

Develop Computer-based Assessments that Include Advantages for Students with Disabilities and Avoid the Pitfalls of Low Expectations

Computer-based assessments show promise for increasing the accessibility of assessments. They also make it easier to fall back into some pitfalls that have been demonstrated to create problems for the assessment of students with disabilities.

There are many promising aspects of computer-based assessments for students with disabilities. They can be developed in a way that embeds what are called "accommodations" when the test is paper based. Michael Russell at Boston College and his colleagues have identified many of the possibilities that accessible computer-based assessments provide:

- Allow users to navigate and interact with the functional elements of the test delivery system using a standard mouse, keyboard, touch screen, Intellikeys, switch mechanism, sip-and-puff device, eagle-eyes, and other assistive communication devices
- Allow all text to be read aloud using a human voice or synthesized voice or signed.
- Provide spoken descriptions of all graphics, drawings, tables, functions, formulas, and other non-text-based elements of an item
- Provide multiple magnification tools that allow users to alter the level of magnification at any time during testing
- Provide an auditory calming tool that allows all students to select from among a list of pre-approved sound files and to have the sound file play softly in the background as the user works on the test
- Allow users to alter the color contrast of the display or mask portions of the screen
- Allow text to be presented on an electronic Braille display attached to the user's computer

A computer-based system could record each use of an incorporated feature or accommodation to document use for individual items as well as overall.

There are tremendous possibilities for dramatically increasing the accessibility of assessments in a computer-based assessment system based on grade-level content

standards. These assessments also have the potential to aid teachers as they determine how to move students to grade-level achievement.

Computer-based systems should be developed to be as transparent as possible about the content on which students are assessed and the ways in which the content is assessed. They should not revert to normative assessments, even in the name of being able to measure growth. We know from the Title I evaluation systems prior to 1994 that schools can show that students make progress, but without being anywhere close to where they should be. The system failed to reveal where schools were succeeding and where they were not. Computer-based systems should not revert to an out-of-level testing approach. Any adaptive computer-based assessments must be on **grade-level**, and must be transparent enough to know when a student in inaccurately measured because of poor basic skills but good higher level skills.

To conclude, it is puzzling to realize that it took assessments in an accountability system with teeth to push improved achievement for students with disabilities, and for that matter, for children of poverty and minority status. Standards-based assessments for accountability have accomplished more in raising expectations and outcomes than three decades of IEPs, one child at a time. We cannot, as a nation, afford to discontinue higher expectations and outcomes for our students with disabilities. We cannot affort to develop assessment systems that track students inappropriately into differential access

to the general education curriculum nor, as they do in many states, result in early decisions about the diploma options available for the student with a disability.
The Race to the Top funds provide a unique opportunity for states to work intensively to develop innovative and inclusive assessments for students with disabilities.

So, although there are ways to improve assessments to let all students show what they know and can do, we have evidence that without a commitment to the same standards, these tests will simply demonstrate that students with disabilities are not getting what is due them – specialized instruction and access to the curriculum so that they can achieve the standards and goals set for all students.

Regardless of the specific nature of the assessment systems – classroom, performance based, technology-based, or a blend of approaches, it is critical that we think of all students from the outset of the development process. Race to the Top funding should be contingent on comprehensive and cohesive plans for increasing accessibility for students with disabilities.

Including Students with Disabilities in Educational Accountability Systems[1]

Martha L. Thurlow, PhD, Director
National Center on Educational Outcomes
University of Minnesota

Students with disabilities are part of the U.S educational system. Dramatic benefits to students with disabilities have resulted from the inclusion of these students in the federal educational accountability system focused on results. The reported challenges related to including students with disabilities are often rooted in false assumptions. The accountability systems of the past decade have shown a light on practices that consign students with disabilities to a separate and less rigorous curriculum based solely on disability status. These accountability systems have also shown a light on practices in those schools that have provided leadership, staff support and training, and laser-like focus on the standards for all students. These schools are an "existence proof" that students with disabilities can learn to high levels in the general – NOT a "special" – curriculum when given necessary supports such as appropriate accommodations and assistive technology, and the specialized instruction they need, Current turmoil in the field about the low performance of students with disabilities is a symptom of an accountability system working – it has increased attention to a group of students historically subjected to low expectations and minimal access to the general education curriculum, and has opened the door to examination and exploration of ways to increase the academic outcomes of student with disabilities.

Accountability systems starting from the 1994 Elementary and Secondary Education Act (Improving America's Schools Act) have done more to bring attention to students with disabilities, their need for access to the curriculum, for assistive technology, for individualized services, supports, and specialized instruction, and for better assessments, than previous laws and requirements. Despite the fact that it may be challenging for schools to rethink traditional but ineffective practices and separate curricula for students with disabilities, federal educational accountability has been good for students with disabilities. With the evolution of the federal accountability system, it has ensured that these students *finally* are being given access to the curriculum that Individuals with Disabilities Education Act (IDEA) required that they be given access to since 1997.

Stories in the media do not necessarily paint the same picture that I will portray. The theory of action underlying the accountability system under ESEA assumes that the data from standards-based assessment systems will generate needed retraining, redesign, and refocusing of schools that are not meeting accountability targets. In many cases, the resources that should be used in improving instruction and access to the curriculum have been used to lobby for the status quo. Frequently the stories in the news reflect the perspectives of those who do not want to be held responsible for students with disabilities, whom they still view as the most difficult to teach of their students, and in many cases, whom they view as those students who cannot learn. Research also has shown that students with disabilities have not always benefited from educational reforms, with some districts resorting to increased referrals to special education as a way to deal with struggling students (Furney, Hasazi, Clark/Keefe, & Hartnett (2003), and many simply not considering students with disabilities in the same way as other students in the context of reform (Gagnon, McLaughlin, Rhim, & Davis, 2002). Evidence of continued emphasis

[1] Opinions presented in this paper are those of the author, and do not necessarily reflect those of any of her funding sources, including the U.S. Department of Education or Offices within it.

on a separate curriculum that prioritizes functional or adaptive skills rather than academics continues to emerge in research (Carter & Hughes, 2006; Dymond, Renzaglia, Gilson, & Slagor, 2007; Mayrowetz, 2009). Increasing numbers of studies indicate that instructional and curricular deficits are a common factor in the low performance of students with disabilities (Hess, McDivitt, & Fincher, 2008; Kettler et al., 2009; Lazarus & Thurlow, in press; Perie, 2009; Thurlow, 2008).

Failure to provide students with a strong standards-based approach to education is not the case everywhere. There is evidence that where educators believe that students can learn, the mandate of a standards-based accountability system has been taken seriously and students have made tremendous improvements in performance, with the achievement gaps narrowing to an impressive extent (Center on Education Policy, 2008; 2009; Chenoweth, 2009)**.**

A standards-based vision of accountability for all has been especially challenging for the special education subgroup–-to the point that districts and states have devised ways not to be accountable for some of these children. There is one very small group of students – those with the most significant cognitive disabilities who are less than 1% of the total student population – where there is general agreement on appropriateness of flexibility in achievement standards. Additional flexibility has been provided for another group of students with disabilities, students sometimes referred to as the “gap” students. The “2%” regulations permit a modified achievement standard against which up to 2% of all students are held to be considered proficient. The regulations suggest that these students are those who may need more time than a school year to reach proficiency on grade-level achievement standards. Analyses of state data show that 30-50% of students who are achieving at a slower or lower rate are students WITHOUT disabilities, but the regulation applies only to students with disabilities. It is a relatively new flexibility provision; its implementation has not progressed very far and there has been little attention yet to the consequences of this flexibility. However, states and districts continue to clamor for more “relief” from accountability requirements for more and more students with disabilities. This continuing pressure on policymakers to hold more and more students to lowered expectations reflects real challenges facing educators with the inclusion of students with disabilities in accountability systems, but that does not mean that there are not solutions to these challenges. It is NOT an appropriate response to back off of holding schools to high expectations for these students.

Before describing the benefits and challenges of standards-based accountability systems, there are two important basic pieces of information that should be revisited: first, a quick reminder of why state assessments are appropriate for students with disabilities; second, a refresher on who students with disabilities are so that we all are thinking about the same students.

Some Basics

Why State Assessments are Appropriate. A primary function of accountability assessments is to give a picture of achievement, and to serve as a warning flag – something that alerts us to review low achievement results, *so that something can be done about it*. We need to know when students are not performing well. This is particularly important for students with disabilities, who in many schools have not had meaningful access to the general education curriculum. In other words, many students with disabilities have not received the accommodations, services, supports, and specialized instruction they need to succeed in the same challenging curriculum as their peers, as required by IDEA. State assessments can provide a window to shed light on whether meaningful access and high quality instruction has been provided, but assessments can improve outcomes only *if* limited access or poor instruction

change as a result of that view. This does not mean that assessments cannot be improved, and be even better indicators of achievement for students with disabilities, and other students as well. But, it does mean that we need standards-based measures that document the grade-level achievement of students with disabilities. We also must ensure that classroom-based assessments provide strong, standards-based data to avoid lowered expectations within the year, and to avoid a false sense of growth in student learning that is based on a separate and less rigorous curriculum and accompanying assessments.

Who Are Students with Disabilities? It is also important to consider who students with disabilities are. There are misperceptions among the public, policymakers, and even educators about the makeup of this subgroup. Figure 1 summarizes data on the numbers of students with disabilities in each category of disability. *Most* students with disabilities (75% altogether) have learning disabilities, speech/language impairments, and emotional/behavioral disabilities. I believe that there should be no question that these students, along with those who have physical, visual, hearing, and other health impairments (another 4-5%), totaling about 80% of students with disabilities, can learn the grade-level content in the general education curriculum when given appropriate accommodations, services, supports, and specialized instruction. They do so by going around the effects of their disabilities, often via accommodations, and thus achieve proficiency on the grade-level content standards.

People might question whether students with *intellectual impairments* (less than *20% of all students with disabilities*) can be expected to achieve proficiency on grade-level standards, but in fact in many cases they can, when they receive high quality instruction in the grade-level content, appropriate services and supports, and appropriate accommodations. An example of such a student—one who does have an intellectual disability but who with adequate school supports and services, and with accommodations during instruction and during assessment, was able to obtain access to the content—is Katie Bartlett, a young woman with Down syndrome, who passed the high school competency exam in Massachusetts (see Appendix A). Katie was held to the same standards as other students and met them.

But, the low expectations for students with disabilities are more pervasive than being just for those students with intellectual disabilities. The example of Lance Rogers, a young man with attention deficit hyperactivity disorder exemplifies that low expectations can hit students with all types of disabilities, with potentially devastating effects (see Appendix B).

Expecting less than proficiency on grade level content from students like Katie Bartlett and Lance Rogers, and indeed expecting less from any student with a disability, reflects limited understanding of who these children are. Students with disabilities can meet standards. Special education eligibility should not be an excuse to expect less from a child, nor to provide little to a child. Existing assessment and accountability requirements have helped us recognize and shed light on the outrageous problem of low expectations for these children, reflected in statements like this: “If special education students could perform well on these tests, they wouldn’t be in special education.”

What Have Been the Benefits of Standards-based Educational Accountability for Students with Disabilities?

Data are emerging that show increases in the achievement of students with disabilities, and other effects as well. For schools to produce the increases in achievement, there has to first be the belief system that students with disabilities can learn and, second, a commitment to their

inclusion in the accountability system. These two conditions, of course, relate to the challenges that continue to surround the inclusion of students with disabilities in accountability systems.

Statements made in the news media by educators and administrators reveal that the belief system that students with disabilities can learn is not yet pervasive. Other indicators of questionable beliefs about the special education subgroup are (a) initial decisions of state policymakers to require a larger number of students in the disability subgroup than in other subgroups for results to be reported or included in accountability calculations; and (b) the pattern of omission from key policy decisions, such as certain applications of growth models that overlooked students with the most significant cognitive disabilities.

Despite the existence of these indicators of beliefs about the special education subgroup, there are pockets of believers who have resisted loopholes in accountability for the disability subgroup. These schools are showing that accountability data are important for students with disabilities, and that they can serve exactly the purpose for which the inclusion of all students in the accountability system was intended.

Transparent Data. State directors of special education have responded to National Center on Educational Outcomes surveys about the improvements that they are seeing in the performance of their students, attributing the improvements to use of assessment data to inform decision making, emphasis on inclusion and access to the curriculum, increased access to standards-based instruction on the grade-level content, alignment of Individual Education Programs (IEPs) with grade-level standards, increased inclusion of students with disabilities in general education classrooms, increased use of research-based “best practices,” and improved professional development (Altman, Lazarus, Thurlow, Quenemoen, Cuthbert, & Cormier, 2008).

Increases in Participation and Performance. Analyses of publicly reported assessment data since 2000-2001 show improvements in the transparency of data for students with disabilities, both for participation and for performance (Thurlow, Quenemoen, Altman, & Cuthbert, 2008). For example, NCEO’s identification of states with clear participation reporting to the public for students with disabilities showed only 5 states in 2000-2001, but 22 states in 2006-2007. (Albus, Thurlow, & Bremer, 2009) These data also showed large increases in participation percentages across time for most states. Data on performance showed similar changes—more states with clear transparent reporting, and increases in performance across years (see examples of one state’s data in Figures 1 and 2). Data from the National Assessment of Educational Progress (NAEP) support these data—overall there has been an increase in the performance of students with disabilities, and in some states the increases are dramatic for all students and for students with disabilities (see Figure 3).

Improved Provision of Accommodations. Students with disabilities are receiving more appropriate accommodations (such as large print editions, extra breaks, sign language interpreters, calculators, individual administrations) during testing. We are still improving our knowledge base about the kinds of changes that are appropriate during testing and that do not invalidate scores. States are still determining their policies in line with what they are testing, and increasingly engaging in training and monitoring to ensure appropriate implementation of accommodations. All the attention on accommodations has lifted awareness in the field so that students who need them are more likely today to be receiving needed accommodations—during testing and during instruction. Prior to the inclusion of students with disabilities in accountability, high estimates indicated that the use of testing accommodations was about 53% for elementary schools and 44% for middle and high schools (Thompson & Thurlow, 1999). In 2005, the percentages were 65% for elementary schools, 64% for middle schools, and 62% for high

schools (Thurlow, Moen, & Altman, 2006) We do not know, as yet, what the percentages *should* be, but having them closer to the same across school levels is a good sign.

Changes in Instruction. Studies are being conducted to determine what produces good performance among students with disabilities. One of these studies, conducted in Massachusetts, identified low income urban schools where special education students were performing better than expected. The researchers then investigated what was happening in those schools for the special education students. They found 11 common factors that characterized those schools where students with disabilities were performing well: (1) a pervasive emphasis on the curriculum and alignment with the standards, (2) effective systems to support curriculum alignment, (3) emphasis on inclusion and access to the curriculum, (4) culture and practices that support high standards and student achievement, (5) well-disciplined academic and social environment, (6) use of student data to inform decision making, (7) unified practice supported by targeted professional development, (8) access to resources to support key initiatives, (9) effective staff recruitment, retention, and deployment, (10) flexible leaders and staff that work effectively in a dynamic environment, and (11) effective leadership (Donahue Institute, 2004).

One group of students for whom accountability requirements have had tremendous impact is students with *significant cognitive disabilities,* individuals who for the most part have been in the educational system only since the passage of Public Law 94-142, the Education of All Handicapped Children Act. Prior to this, some of these children would have been in institutions, or in care-giving facilities. In December 2003 the 1% rule permitted states to develop alternate assessments that allow proficiency to be defined differently from the regular assessment. States could set alternate achievement standards on the results of these alternate assessments in order to define proficiency for students with the most significant cognitive disabilities. These assessments are identified as assessments for students who are unable to take regular assessments. In those states where the right students were targeted, these alternate assessments have resulted in a sea change to the instruction that students with the most significant cognitive disabilities are receiving. Assessment has changed their school experiences. They are now receiving an education—access to the general curriculum—rather than a separate curriculum, or even simply receiving care in the schools. NCEO has data from state assessments documenting the shift to the general curriculum, and qualitative data on the perceptions of educators on the benefit of this shift (Moore-Lamminen & Olsen, 2005). Impact data are being collected nationally by the National Alternate Assessment Center. Along with these changes have come organized resistance to the changes. We have, for the first time, succeeded in looking behind the closed resource room door, and although many wonderful teachers and practices have been uncovered, these are not universal. Keeping the resource room door “open” with the lights of accountability shining is a necessity at this critical time to bring along those who are afraid of or do not know how to step up to our new understanding of what these students can do if taught.

Attempts are being made to collect information on overall impacts of accountability requirements. One of these is time spent in general education environments, which is a proxy for access to the general curriculum. The *26th Annual Report to Congress* reported that the percentage of time spent in the general education classroom had increased for students with disabilities through the 2002 data available in the report. NCEO surveys of state directors of special education suggest that standards-based accountability continues to expand this access to the general curriculum for students with disabilities (Altman et al., 2009).

Better Assessments. There have been several benefits to assessment systems that have resulted from requirements to include all students, particularly those with disabilities.

Assessment developers who typically have been able to exclude anyone who did not fit the norm have had to revisit assumptions and revise assessment models. This has been good for the assessment world. States, in particular, have focused on how to make their assessments more accessible—through better accommodation policies and by applying universal design principles to assessments—not by changing the construct that is to be measured or reducing the grade-level content that is being assessed (Johnstone, Thompson, Miller, & Thurlow, 2008; Thurlow, in press). Rather, developers have focused on ensuring that tests really measure what they are intended to measure—not extraneous factors, such as whether the student can figure out what the test developer means by a question or whether a picture has important clues about the answer to a question (Dolan, Burling, Harms, Beck, Hanna, Jude, Murray, Rose, & Way, 2009; Thurlow, Laitusis, Dillon, Cook, Moen, Abedi, & O'Brien, 2009; Thurlow, Quenemoen, Lazarus, Moen, Johnstone, Liu, Christensen, Albus, & Altman, 2008). Identifying ways to improve assessments for students with disabilities has, in fact, resulted in improving assessments for all students.

What Have Been the Challenges of Standards-Based Educational Accountability?

Despite the many benefits of standards-based educational accountability, there has been an undercurrent of challenges surrounding the disability subgroup. These have manifested themselves in a number of ways. The most obvious was the decision made early on in many states that assessment results would be publicly reported and included in accountability for students in the disability subgroup only when a larger number was included relative to those in other subgroups. It is not uncommon to find that reporting and accountability occurred in a school, for example, for all subgroups except the disability subgroup when there were 30 students in a group; for students with disabilities, the number had be 45 students. Despite detailed arguments about error rates and confidentiality, it seemed rather to be a policy decision—a way to avoid responsibility for a group perceived to be the lowest performing subgroup. Stories also soon spread about districts having students disappear from the denominator by moving them around, either from school to school, or from one grade to another, or by having their scores count for participation but not for performance. We heard about the "EGG" Game being played by principals—the "enrolled grade game" – holding 9th grade students back one year, then promoting them to 11th grade to ensure they did not take the 10th grade test, which was the accountability test. Not all states and districts are doing these things, and some states have policies and monitoring to prevent it. When they do these "games," it hides the very thing that we need to know about—the performance of low performing students—very often, students with disabilities.

The overall initial poor performance of students in the disability subgroup may be the reason that states, districts, and schools have reacted out of fear and have tried to identify loopholes for their accountability systems. It also may be the reason that they and others continue to seek other loopholes and additional flexibility in a system that is designed to do exactly what it should do—hold the system accountable for students who are not reaching grade-level achievement standards.

Where Should Educational Accountability Go From Here?

I can remember the time before federal laws required accountability for all students, when there were no requirements for students with disabilities to be included in the assessment systems along with other students. I recall stories of these students being sent on field trips to the zoo, or the parents of these students being told kindly by the teacher or the school principal

that their children could stay home from school the day that the state test was given. I remember watching the referral rates of students to special education creep up not so slowly as more and more students were exempted from school testing in those states that had accountability systems in place and did not want to be accountable for all of their students. Their solution was to get rid of the students for whom they did not want to be accountable.

The solution is NOT to get rid of students who are tough to teach. It was not the correct solution then, and it is not the correct solution now. We have enough emerging evidence that students with disabilities can learn when given access to the curriculum and rich content instruction by qualified teachers that we can no longer justify coming up with a different accountability system for them. There are some principles that need to guide our approach to accountability for all—to ensure that students with disabilities thrive along with other students.

1. We need to recognize that all students with disabilities are general education students first—they need access to the general curriculum, qualified teachers, and high expectations for their learning. With accommodations, services, and supports, we should expect 85-90% of these students to achieve grade level mastery on achievement tests that are administered in today's schools. The remaining students have significant cognitive disabilities, and they achieve proficiency through the appropriate option of alternate achievement standards. Thus, we need to treat students with disabilities the same as other students. Except for the 1%, and the recognition of the need for accommodations, services, and supports as provided for by the IEP, the same approaches should exist for students with disabilities as exist for general education students.

2. We need to focus attention on the lowest performing students, not try to hide their performance or get them out of the system. Experiences with the 1% students have shown that there have been dramatic benefits to this population through the requirement that they be assessed, and for their performance be reported separately. A strong research effort has developed around this type of assessment, and there are now systematic methods to ensure greater access to the general curriculum for these students. These are succeeding both in ensuring access to the curriculum and in increasing student performance to levels educators and parents thought impossible. There are many other students with disabilities who are performing at very low levels—students with and without intellectual disabilities. We need to focus on these students, the lowest performing students—to open up reporting and provide information on what is happening instructionally, and in the provision of accommodations and services. Based on previous research done in some states, I am confident that by looking at the lowest performing students on the regular assessment, for example the lowest 20%, a state will find not only students with disabilities, but low performing general education students as well (Hess et al., 2008). Then, the state can focus more appropriately on the issues that exist for all of those students—access to the curriculum, improved instruction, and yes, accommodations, and special education services as well if appropriate.

3. Adjustments to the accountability system itself should be made with care and should apply to all students, not to one subgroup. Critical elements of accountability for the disability subgroup are the 1% flexibility, the recognition that tests should be universally designed (which really benefits all students), the requirement that accommodations should be provided to those students needing them, and separate reporting by subgroup. Other adjustments to accountability systems should be made for all students, not just one subgroup, with consideration of intended and unintended consequences for students overall and for student subgroups. Any accountability system that includes measures of growth must be related to absolute standards. It would be unacceptable a growth system to be based on individualized growth goals set by an IEP team. An IEP-based approach would have unintended negative

consequences for the disability subgroup (and others groups as well); students with disabilities for years have shown growth against IEP goals, but have inappropriately been held to inappropriate goals or estimates of needed improvement. Considerations for growth models that include all students with disabilities have been identified (see Appendix C – *Using Growth for Accountability: Considerations for Students with Disabilities*).

The critical difference that standards-based accountability has made for students with disabilities is the link to grade-level content standards, a link that will bring them up to the level that they need to meet to be successful to move forward in their school work and to become

References

Albus, D., Thurlow, M., & Bremer, C. (2009). *Achieving transparency in the public reporting of 2006-2007 assessment results* (Technical Report 53). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Altman, J. R., Lazarus, S. L., Thurlow, M. L., Quenemoen, R. F., Cuthbert, M., & Cormier, D. C. (2008). *2007survey of states: Activities, changes, and challenges for special education*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Carter, E.W. & Hughes, C. (2006). Including high school students with severe disabilities in general education classes. *Research and Practice for Persons with Severe Disabilities, 31*, 174-185.

Center on Education Policy. (2008). *Has students achievement increased since 2002? State test score trends through 2006-07.* Washington, DC: Author.

Center on Education Policy. (2009). *Is there a plateau effect in test scores?* Washington, DC: Author.

Chenoweth, K. (2009). *How it's being done: Urgent lessons from unexpected schools.* Cambridge, MA: Harvard Education Press.

Dolan, R.P., Burling, K.S., Harms, M., Beck, R., Hanna, E., Jude, J., Murray, E.A., Rose, D.H., & Way, W. (2009). *Universal design for computer-based testing guidelines.* Iowa City, IA: Pearson.

Donahue Institute. (2004). *A study of MCAS achievement and promising practices in urban special education: Data analysis and site selection methodology.* Amherst, MA: University of Massachusetts.

Dymond, S., Renzaglia, A. Gilson, C. & Slagor, M. (2007). Defining access to the general curriculum for high school students with significant cognitive disabilities. *Research and Practice for Persons with Severe Disabilities*, 32(1), 1-15.

Furney, K. S., Hasazi, S. B., Clark/Keefe, K., & Hartnett, J. (2003). A longitudinal analysis of shifting policy landscapes in special and general education reform. *Exceptional Children, 70*(1), 81-94.

Gagnon, J. C., McLaughlin, M. J., Rhim, L. M., & Davis, G. A. (2002). Standards-driven reform policies at the local level: Report on a survey of local special education directors in large districts. *Journal of Special Education Leadership, 15*(1), 3-9.

Hess. K., McDivitt, P. & Fincher, M. (2008, June). *Who are those 2% students and how do we design items that provide greater access for them? Results from a pilot study with Georgia students.* Paper presented at the 2008 CCSSO National Conference on Student Assessment, Orlando, FL. Retrieved from http://www.nciea.org/publications/CCSSO_KHPMMF08.pdf

Johnstone, C.J., Thompson, S.J., Miller, N.A., & Thurlow, M.L. (2008). Universal design and multi-method approaches to item review. *Educational Measurement: Issues and Practice*, 27 (1), 25-36.

Kettler, R., Russell, M., Camacho, C., Thurlow, M., Geller, L., Godin, K., McDivitt, P., Hess, K., & Bechard, S. (2009). *Improving reading measurement for alternate assessment: Suggestions for designing research on item and test alterations.* A white paper based on discussion which occurred at the Invitational Research Symposium on Alternate Assessment based on Modified Achievement Standards in Reading.

Lazarus, S.S., & Thurlow, M.L. (in press). The changing landscape of alternate assessments based on modified academic achievement standards (AA-MAS): An analysis of early adopters of AA-MAS. *Peabody Journal of Education*.

Mayrowetz, D. (2009). Instructional practice in the context of converging policies: Teaching mathematics in inclusive elementary classrooms in the standards reform era. *Educational Policy, 23*(4), 554-588.

Moore-Lamminen, L., & Olsen, K. (2005). *Alternate assessment: Why bother?* Lexington, KY: Mid-South Regional Resource Center.

Perie, M. (Ed.) (2009). *Considerations for the alternate assessment based on modified achievement Standards (AA-MAS): Understanding the eligible population and applying that knowledge to their instruction and assessment.* Washington, DC: U.S. Department of Education.

Thompson, S. J., Johnstone, C. J., Thurlow, M. L., & Altman, J. R. (2005). *2005 State special education outcomes: Steps forward in a decade of change*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S.J., & Thurlow, M.L. (1999). *1999 State special education outcomes: A report on state activities at the end of the century*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M.L. (2008). Assessment and instructional implications of the alternate assessment based on modified academic achievement standards (AA-MAS). *Journal of Disability Policy Studies, 19*(3), 132-139.

Thurlow, M.L. (in press). Steps toward creating fully accessible reading assessments. *Applied Measurement in Education.*

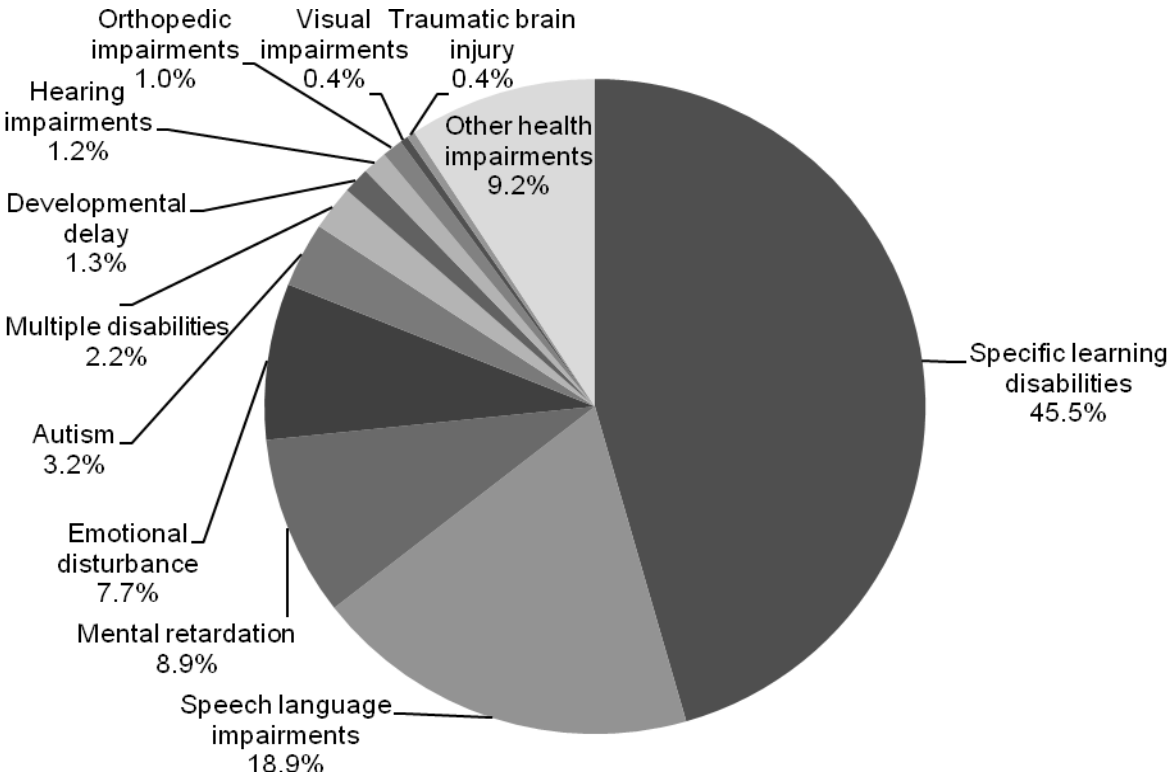
Thurlow, M. L., Laitusis, C. C., Dillon, D. R., Cook, L. L., Moen, R. E., Abedi, J., & O'Brien, D. G. (2009). *Accessibility principles for reading assessments.* Minneapolis, MN: National Accessible Reading Assessment Projects.

Thurlow, M.L., Moen, R., & Altman, J. (2006). *Annual performance reports: 2003-2004 state assessment data (Available on the NCEO Website at http://www.education.umn.edu/nceo/OnlinePubs/APRsummary2006.pdf).* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M., Quenemoen, R., Altman, J., & Cuthbert, M. (2008). *Trends in the participation and performance of students with disabilities* (Technical Report 50). Minneapolis, MN: University of Minnesota, National center on Educational Outcomes.

Thurlow, M.L., Quenemoen, R.F., Lazarus, S.S., Moen, R.E., Johnstone, C.J., Liu, K.K., Christensen, L.L., Albus, D.A., & Altman, J. (2008). *A principled approach to accountability assessments for students with disabilities* (Synthesis Report 70). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Figure 1. Distribution of Disability Categories



Note: Percentages in this figure are based on a total number of 6.5 million students receiving special education services (**www.ideadata.org**, 2008).

Figures 2 and 3. Performance Data Showing Increases for Special Education Subgroup

Figure 2. Percent of Students who Earned a Competency Determination - Class of 2003

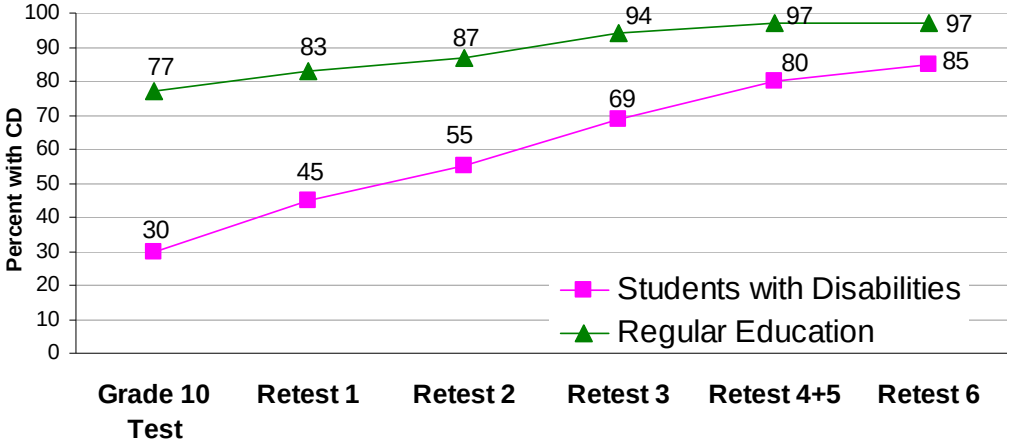


Figure 3. Percent of Massachusetts Special Education Students Attaining the Competency Determination (through the May 2004 MCAS)

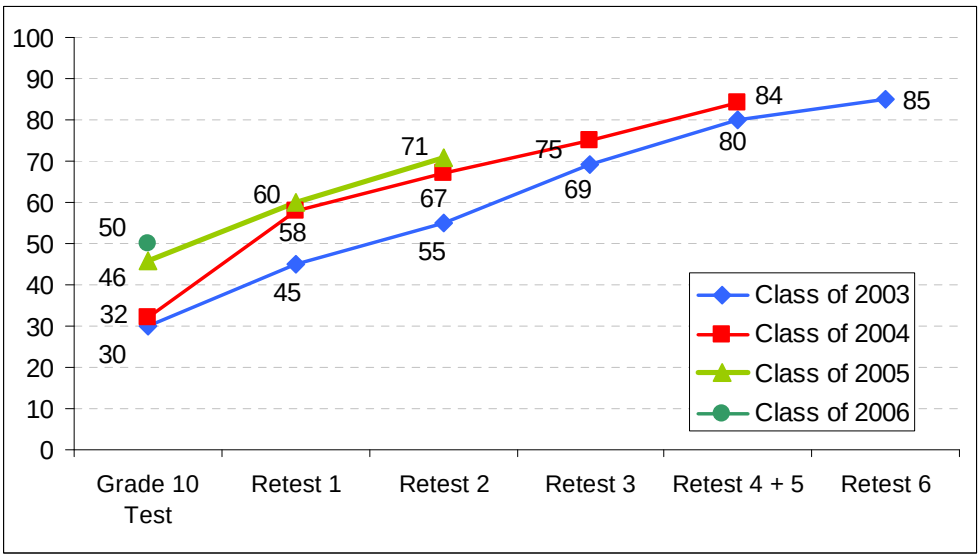
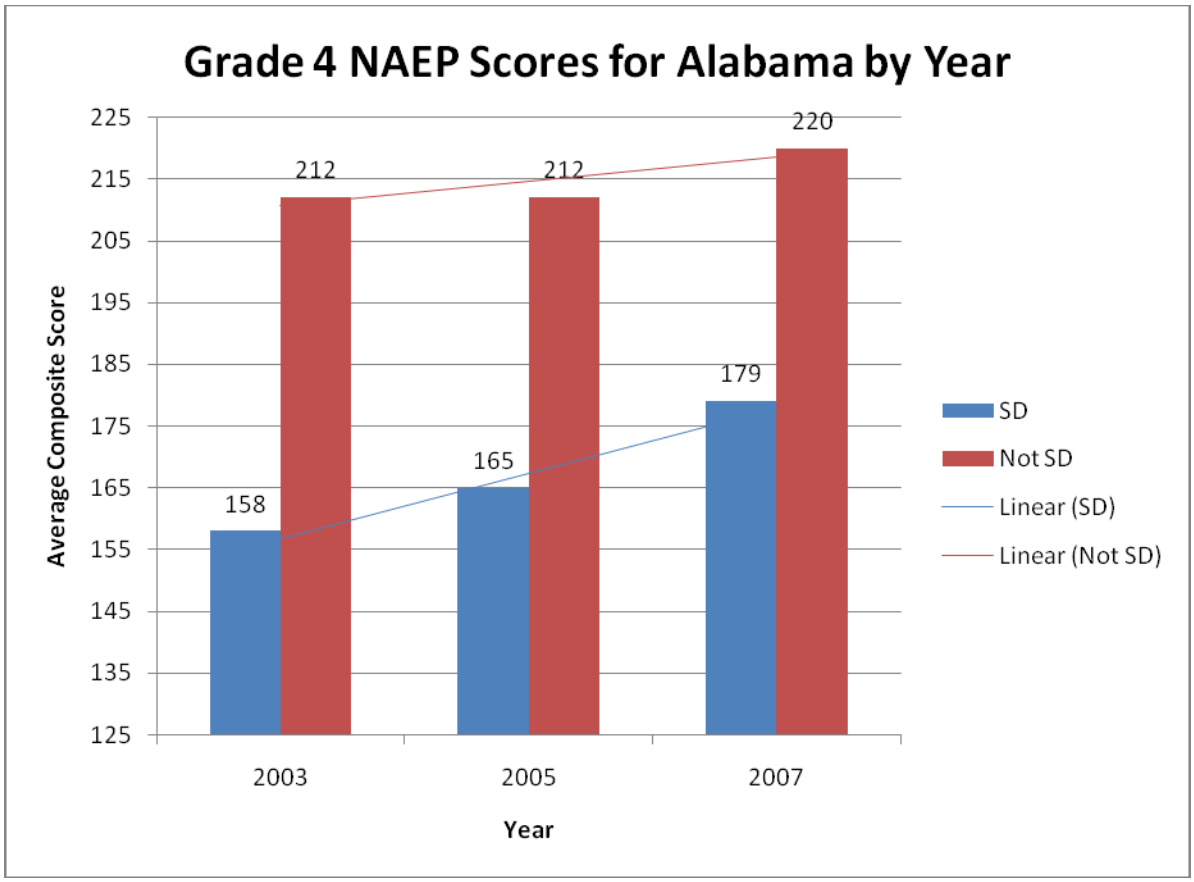


Figure 4. NAEP Performance Data Showing Increases for All Students and Students with Disabilities in Alabama



Appendix A
Story of Katie Bartlett from Boston Globe

Education test holds a special value

Teen's MCAS success jolts notions on Down syndrome

By Marcella Bombardieri, Globe Staff, 12/22/2002

ARLINGTON - Katie Bartlett has spent all of her 17 years exceeding the expectations the world placed on her when she was born with Down syndrome. The 11th-grader earns A's and B's in college-prep classes. She competes on the Arlington High School swim team, sings in the chorus, and devotes her spare time to writing poetry, keeping a diary, and drawing a colorful array of girls' fashions.

Still, no one was quite sure what would happen when Bartlett took the MCAS exam, now a requirement for a high school diploma in Massachusetts.

This is what happened: She passed.

On her first try, no less, with only a few accommodations - like the use of a calculator - that are required by law for students with disabilities. She passed the regular paper-and-pencil test with the same questions posed to everyone else.

"Thank you! Thank you!" Bartlett responded recently when she learned her scores. "I'm so glad I don't have to take it again."

Some advocates see Bartlett's success as not just a personal triumph, but a victory for all special-needs students. They believe it proves that 30 years after the birth of modern special education, society's expectations are still far too low for children with disabilities.

Many education specialists who disapprove of a high-stakes standardized test still see a value in MCAS, because it highlights which classrooms are giving special-ed students a first-rate education and which are not. A few even hold the belief that, in the long run, the Massachusetts Comprehensive Assessment System will prove a boon for the education of disabled students, whether they pass it or not.

The MCAS requirement "has been the most effective advocacy tool that I've seen come across the horizon, ever," said Rich Robison, executive director of the Federation for Children with Special Needs, even though his 18-year-old son with Down syndrome and autism has not yet passed. "It forces schools to say, `Oh, yeah, we can't just ignore them. Their scores are going to count. We do have to teach them something.'"

Bartlett's mother, Jan Tobin, is uncomfortable focusing on this one accomplishment, which is "not the best that Katie Bartlett has done or will do."

Just as important to those who know Bartlett is how she has overcome shyness so she now can sing in public with gusto, dance in her jazz class, and swim in front of a crowd. At a recent holiday concert, she belted out a medley of songs - one in Hebrew - to great applause. She has inspired classmates with her hard work and high spirits.

"I don't like B's much. I like getting all A's," said Bartlett, who is poised and well-spoken, if not verbose. She's short for her age, and this day wore a headband in her long brown hair and salmon-pink polish on her nails.

Bartlett has been a star since she was a young child, when she appeared in two oft-replayed episodes on Sesame Street.

"Our first goal for her is that she be happy," said Tobin. "Our second goal is that she have friends, and the third goal is that she learns what she needs to live in the world."

Yet the test communicates something that the subjective accomplishments cannot. Without it, "some would still suspect she was passed along because they liked her," said Gerald Fain, chairman of the special education department at Boston University's School of Education. "This really dignifies her. No more patronizing. It shows anyone who cares to take note that she has achieved her high school diploma."

Only 55 percent of this year's high school seniors with disabilities have passed the test. But Fain says that doesn't mean the test is bad or the students incapable. While some students' disabilities make it impossible for them to pass, most would earn passing marks if they had the right instruction and the proper accommodations, he said.

"MCAS is necessary in order to correct social injustice," Fain said.

The reason: Many school systems have not met their obligations to properly educate special needs students, he and others said. A US Department of Education review of Massachusetts in 2000 found "a lack of opportunity for children with disabilities to be involved and progress in the general curriculum" in some districts. The state is still working to correct those problems to the satisfaction of the federal government. Meanwhile, two disabled students are among seven students suing Massachusetts over the MCAS.

Fain doesn't believe that a failing MCAS score should prevent students from graduating, at least not until the state has better met its obligations to properly educate all students.

Tobin and her husband, Larry Bartlett, have been pushing for Katie to get the best education possible ever since she started an "early intervention" program as an infant. When they looked at preschools, their favorite was an ordinary, mainstream preschool. That class worked out so well that they chose an ordinary kindergarten, and went on from there.

As satisfied as they have been with the Arlington schools, there were times along the way when Bartlett's parents had to convince educators that inclusion was appropriate for their daughter.

"When it was time for first grade, for middle school, and for high school, people said, `It's time to think about special classes,'" Tobin said. "`It's time now, it's time now,' - and it's not! What we need is more academics, not less."

Paving the way for Bartlett's achievements required a lot of effort. She works diligently, her parents stay closely involved, and the school provides intensive support. She has a full-time aide who goes with her to all academic classes - currently English, second-year algebra, and US history - and works with her in two study periods each day. She also takes health and gym. Although her course load is slightly lighter than most students', she's earning enough credits to graduate on time.

The aide, Carol Konz, often takes class notes so Bartlett can concentrate on listening, and reviews the material with her. Konz sometimes modifies an exam to reduce the number of questions or simplify an essay question. During tests - but not the MCAS - Konz provides cues to help her retrieve information in her memory, but doesn't give her the answer. Tobin and Konz developed tricks over the years, like singing about the parts of a human cell to the tune of "Yankee Doodle."

Arlington High School classes are tracked into several levels, and Bartlett takes college-prep classes except in math, where she's in the "basic" class. She gets "a slightly accommodated grade, but a fair grade," said Marilyn Bisbicos, director of special education in Arlington. "It's not a gift."

While Bartlett is more capable than many people with Down syndrome, she's not far out of the average range for those with the chromosomal disorder. In eighth grade, she read at a sixth-grade level, but her parents believe her reading comprehension has advanced since then.

One of her advantages is that she has had few of the health complications, such as heart problems, that can accompany Down syndrome. She is also blessed with an excellent memory, to the point that she often got 100s on spelling tests and still uses the Spanish she learned in earlier grades.

Bartlett and her parents approached MCAS with more or less the same attitude they've always had - it can't hurt to try. In addition to the preparation in her classes, she did some extra work with Konz and took a math tutorial designed for students who'd already failed.

During the test, Konz was not allowed to offer any assistance other than reading aloud the passages, which included some by Shakespeare, George Eliot, and Truman Capote. Bartlett took the test in a private room, and she was allowed to use a calculator. She also spent an entire day on each portion, in contrast to a morning for most students. However, the MCAS is untimed, so taking longer than others is allowed.

By law, disabled students are allowed to use the kinds of modifications stipulated in their education plans, to level the playing field.

The test "was really hard for me," Bartlett said. "I really had to stay focused. Shakespeare was the hardest part."

She doesn't know exactly what she wants to do with her life, but she's aiming high. "I want to go to college," she said. "That's my main goal. I like to learn about different things."

Bartlett will probably live at home while attending a community college for two years. After that, a four-year college might be the next step, said Tobin, who has two younger sons.

"I have the same expectations for Katie that I do for my boys," she said. "I never doubted Katie could do anything."

Her success has surprised and inspired plenty of other people.

"She has achieved way beyond our expectations," said Bisbicos. "Maybe that means we should have higher expectations for everyone."

This story ran on page A1 of the Boston Globe on 12/22/2002.
© Copyright 2002 Globe Newspaper Company.

Appendix B
Story of Lance Rogers from Voice of San Diego

Deterred from Diplomas for Better or Worse

By EMILY ALPERT

Thursday, Oct. 2, 2008 | Just after starting high school, Lance Rogers was told he wouldn't earn an ordinary diploma. He struggled with attention deficit hyperactivity disorder and other disabilities, and had trouble focusing in big classes at Point Loma High School.

Instead he took special education classes that were smaller and easier, but wouldn't help him earn a degree. His mother Ruth Rogers hoped he would flourish there, even if he was "non-diploma bound." It is a label given to thousands of San Diego Unified students with disabilities who focus on skills that will help them live independently instead of prepping for college or beyond, studying shopping lists and sales tax instead of calculus or Cervantes.

But Lance Rogers grew depressed and bored in those classes. He can't remember what he learned -- only that he was often asked to draw pictures or maps -- and ultimately ditched school.

"I was downhearted," said Lance Rogers, now 16 years old. "I didn't do my work, because what was the point of doing it? I didn't get any credit. So I didn't go to school."

Yet when the Rogers family moved to Texas, their son thrived in a school with a mixture of small classes and counseling. His grades rose from Ds to Bs. And when the family returned to San Diego, teachers at another school said Lance Rogers was perfectly capable of earning a diploma.

"I was blown away," Ruth Rogers said. "I was shocked that he was in the classroom, doing what he's supposed to be doing."

Stories like his raise questions about whether San Diego Unified has wrongly tagged some students with disabilities, deeming them unlikely to earn an ordinary diploma. Newly released data show that roughly 13 percent of disabled children in San Diego Unified have been marked as "non-diploma bound," some as early as first grade. Some have mental retardation or severe autism; others live with learning disabilities or orthopedic impairments.

Many families have no problem with the label, saying that a diploma is a distant dream for their children. Focusing on basic routines makes more sense to them than coaching their children to pass the high school exit exam. Others complain about premature or sloppy labeling that denies their children the opportunity to achieve by tracking them into non-academic classes. Harvard University researcher Thomas Hehir called the frequency of the practice in San Diego Unified "disconcerting" and "very questionable," and linked the phenomenon to stagnating test scores among students with disabilities.

"You decide early on that kids are not capable," said Hehir, who consults San Diego Unified on

its special education programs and wrote an extensive report on its problems last year. "... If you don't assume they can read, you don't teach them to read."

Special education has been an Achilles heel for San Diego Unified when it comes to standardized testing. It has missed targets under No Child Left Behind due in part to underachievement among students with disabilities. Their scores have scarcely budged while their peers without disabilities surged higher and higher, widening the gap between the able and the less able in classrooms.

No educator means to shortchange children with disabilities, but an overburdened and underfunded system causes mistakes when diagnosing and placing children in classes, said parent Joyce Clark, chairwoman of a San Diego Unified committee on special education. Clark said some children are funneled into easier classes instead of making ordinary classes accessible through technology or other aids.

"Teachers are wonderful but they get weary of trying to address all the needs they are asked to do," Clark said. "And somehow they just fall through the cracks."

Educators and families decide jointly whether a student is headed for a diploma while creating individualized plans for their schooling. Such plans are legally binding and outline what services and supports a student is supposed to receive. Instead of being groomed for college, children who are non-diploma bound are taught skills to live more independently, said Arun Ramanathan, chief student services officer. Instead of a diploma, they usually earn a "letter of recognition" for meeting goals set by their educators and families.

That makes sense for a small fraction of older students with significant cognitive impairments, such as severe mental retardation, Hehir said. But he and Ramanathan believe the label has been overused, branding too many children too easily and too early.

Data released in September by the school district show that more than 2,200 of the roughly 16,400 children with disabilities in San Diego Unified are non-diploma bound. Dozens of middle schoolers with learning disabilities have been deemed unlikely to get diplomas, along with dozens of even younger children with speech impairments. Such early labeling is unnecessary under federal law, which doesn't require educators and families to plan what a student could do after high school until he or she turns 16.

"It's an atrocity if we're deciding in 3rd or 4th grade that [students] won't have access to a regular diploma," said Laura Kaloi, public policy director of the National Center for Learning Disabilities. "You're predetermining the outcomes for students who should have every opportunity to learn and succeed."

Some parents are confused by the process and may agree to divert their children from a diploma despite their misgivings or uncertainty, Clark said. Ruth Rogers said she knew little about what "non-diploma" meant and just wanted smaller classes for her son. But others agree that a diploma is out of reach for their child, and decide to focus on everyday skills instead.

More than one-fourth of non-diploma bound students are diagnosed with mental retardation; another one-sixth are categorized as autistic. One such student is Alex McRee, a gentle senior at

Mission Bay High School who lives with severe autism and a seizure disorder. He does not speak or write, though he sometimes flashes an impish smile when he misbehaves; he likes to tear paper and play with string. His mother jokes that if there were no electric shredders, he would have a perfect job.

Putting McRee in an algebra class or other academic classes seems unthinkable to Janice Yuwiler, whose hopes for her son are more modest: the ability to communicate when he needs to use the restroom or open a granola bar by himself.

"Maybe I'm shortchanging him," Yuwiler said. "Maybe he would have absorbed information through his pores (in a chemistry class). But it seemed like he'd just drift off to sleep."

Teachers call the classes that McRee and other severely disabled students take "functional" instead of academic. They teach basic skills and do not count toward an ordinary diploma. Such classes are common at the Whittier Center, a tiny, specialized school in San Diego Unified for students who cannot be accommodated at ordinary public schools. At Whittier, children with traumatic brain injuries, severe mental retardation and a wide range of other disabilities learn to search newspapers for jobs, calculate sales tax when shopping, and write simple letters in small groups. Even schoolyard games are a lesson.

"It's important for them to be able to take turns," said Principal Cathie Whitley, watching a cluster of students bouncing a rubber ball against a wall. "To lose."

As teenagers they explain their goals for after high school in PowerPoint presentations for their families and Whittier teachers. One talkative boy with autism wrote that he wants to go to culinary school, be financially independent and share an apartment with a roommate. He wants to get a letter of recognition when high school ends, then earn a GED afterward. Whitley said more and more Whittier students are taking that path.

"Maybe they won't earn a diploma in four years," Whitley said. "But increasingly we're looking at our students and saying, maybe it is realistic for them to earn a diploma. Even if it takes longer."

That road is much harder if a student has been learning life skills exclusively and has no usable credits. Trying to earn a diploma after being labeled "non-diploma bound" means catching up on algebra and gearing up for the high school exit exam after spending months, perhaps years, taking classes that don't count towards a diploma.

One San Diego educator who regularly coaches students with disabilities to finish their diplomas, Jill Prier, is currently aiding a student who had taken functional classes for most of her school career after being diagnosed with mental retardation, and "all of a sudden wanted a diploma."

"She's carrying on a job at a clothing store. She doesn't present like a developmentally delayed kid," Prier said. "But she reads at a third grade level. ... She had no credits that were of value. It's going to take her at least three years."

Prier has already helped more than 20 adult students with disabilities graduate last spring with diplomas. Graduation rates for students with disabilities have grown nationwide over the past

decade, with nearly 57 percent of special education students earning ordinary diplomas in 2006, according to a 2008 report by the National Center for Education Statistics.

Fifteen percent earned a certificate or other document that falls short of a diploma, which gives them limited opportunities for jobs and higher education. Most colleges and universities only accept diplomas and many employers require them. Individual states differ dramatically in the percentage of disabled students who get diplomas versus certificates, throwing doubt on whether the numbers reflect the genuine abilities of the students.

As graduation rates have grown, brain research has shown the risks of underestimating children with disabilities, even those with severe conditions that prevent them from speaking, said Anne M. Donnellan, director of the Autism Institute at University of San Diego's School of Leadership and Education Sciences. Donnellan has seen a number of nonverbal students such as Peyton Goddard overcome diagnoses of mental retardation and graduate from college.

Hehir likewise noted that diagnoses are sometimes wrong and students should be given the benefit of the doubt. For instance, conventional wisdom that students with Down Syndrome couldn't learn to read has been shattered as many prove themselves capable of reading and writing as well.

"I'm not interested in predicting what people can do," Donnellan said. "We've made some terrible mistakes with that."

Lance Rogers believes he was a victim of those mistakes. Now a sophomore at the Marcy School, a San Diego Unified center that combines classes and counseling, Rogers said he's taking algebra, chemistry and history to earn the diploma he once was blocked from.

By pursuing a diploma, "I did something they didn't think I was going to accomplish," he said. "They didn't say it like that. But that's what it comes down to."

Please contact Emily Alpert directly at emily.alpert@voiceofsandiego.org with your thoughts, ideas, personal stories or tips. Or set the tone of the debate with a letter to the editor.

voice of sandiego.org™ an independent nonprofit | November 6, 2008

Appendix C
Using Growth for Accountability: Considerations for Students with Disabilities

(See Draft NCEO Policy Directions – next 5 pages)

NCEO POLICY DIRECTIONS

Number 21 November 2009

Using Growth for Accountability: Considerations for Students with Disabilities *(DRAFT)*

Background

The academic progress of students over time is an important indicator or educational success, and something for which schools should be held accountable. Yet, people may mean different things when they use the word *growth* in their suggestions for educational accountability systems. In addition, recommendations may be made for growth-based accountability models that have not considered all students, particularly students with disabilities.

Using a student's academic growth for accountability is generally viewed as an addition to or a replacement for a status model of accountability. **Status** models compare the performance of a cohort of students in a certain grade one year with the performance of another cohort of students in the same grade the preceding year. This comparison is used to determine whether the educational system has shown improvement in student achievement in the specific grade from one year to the next. There are both limitations and strengths in this approach to measuring growth (see Table 1).

Growth models typically track the achievement of individual students from year to year. Tracking student performance over time requires that data systems be set up in specific ways, with specific capabilities. Growth models also can take many forms. Although there are many strengths that measuring growth brings to an educational accountability system, there are also potential limitations (see Table 1).

Growth first emerged in federal accountability systems when the Elementary and Secondary Education Act (ESEA) identified states to participate in a growth model pilot program. This program placed constraints on the nature of growth models, such as the requirement that the model be based on state standards and that it result in 100% of all students and students in each subgroup reaching proficiency in the year 2014.

In this report, we identify some considerations for states and districts as they propose growth models. These considerations focus on the appropriate inclusion of all students, particularly students with disabilities regardless of the assessment they take.

Considerations in Assessing Students with Disabilities

Most students with disabilities participate in the same assessments as other students, although a percentage of

NATIONAL CENTER ON EDUCATIONAL OUTCOMES
The College of Education & Human Development
UNIVERSITY OF MINNESOTA

Table 1. Strengths and Limitations of Status and Growth Models for Accountability

Strengths	Limitations
Status Models	
• Provides snapshot of the performance of a group of students (e.g., school, subgroup) • Compares performance to a designated target • Relatively transparent and easily understood	• Does not recognize improvements in student performance unless target is met • Uses same performance target for all students regardless of initial achievement levels • More frequently viewed as unfair because one group of students can be responsible for the perception that a school is not performing well
Growth Models	
• Provides picture of changes in achievement over a defined period of time • Compares performance to previous performance • Simple models are relatively transparent and easy to understand	• Difficult to define how much change should occur over time for individual students • Easier to mask performance of students not improving when all are aggregated • Complex models are very difficult to understand

students with disabilities may take the assessment with accommodations. It is generally recognized today that the purpose of accommodations is to obtain a valid measure of what the student knows and is able to do, and therefore scores from accommodated assessments can be treated in the same way as other scores.

Some students with disabilities participate in the assessment system by taking alternate assessments. These may be assessments that are based on grade-level achievement standards, assessments based on modified achievement standards, or assessments based on alternate achievement standards. Each of these types of assessments may create challenges for growth models, specifically for including scores from them along with the scores from other assessments.

Historically there have been low expectations for students with disabilities, and evidence of low expectations for these students continues today. One result is that students with disabilities often have been denied access to grade-level content. With the standards-based reform movement, along with access to grade-level content and inclusion in accountability systems, the performance of students with disabilities on grade-level assessments often has exceeded expectations. Understanding the tendency toward low expectations for students with disabilities and the critical importance of inclusion in assessments and in the accountability system to push access to grade-level content, are essential in considerations for growth measures and models that include students with disabilities.

Ways to Measure Growth

The idea of tracking changes in student performance over time sounds simple. But, there are a variety of approaches and measures that may be used to track "growth." Each involves specific assumptions and each has implications for the ways in which student performance is included (see Table 2).

Several of the assumptions and implications of growth models are important to consider when thinking about growth models that include students with disabilities. For example, most growth models do not include results from tests with different scales (only the Transition Matrix model does). This is important for students with disabilities because a significant percentage of them may participate in alternate assessments. States and districts considering growth models must address the limitations for students with disabilities before deciding on an approach to growth models.

Other assumptions of growth models raise general issues, but can have specific implications for students with disabilities as well. For example, most growth models do not include students with missing scores (only the Multi-level model does include

missing scores). The Multi-level approach estimates the scores of students who are missing, as long as at least two scores are available. Unfortunately, estimates of scores may be less reliable if a group of students (such as students with disabilities) has more missing scores than other groups. The greater mobility and poorer attendance of students with disabilities has been documented, and thus is an important factor in considering growth models that are appropriate for all students.

Implications of Disability Characteristics and Assessments for Measuring Growth

Conversations about growth can be derailed when individuals are talking about different approaches to measuring growth. This easily occurs when those most familiar with special education begin to talk with those most familiar with large-scale state and district assessments. For example, growth in special education often refers to changes in performance on curriculum-based or norm-referenced measures most often designed to assess students on low-level basic skills, regardless of

Table 2. Description and Implications of Approaches to Growth Models

Difference Gain Scores
"A student's score at a starting point is subtracted from the same student's score at an ending point. The difference or gain is the measure of an individual's growth. The difference scores can be aggregated to the school or district level to obtain a group growth measure. Growth relative to performance standards can be measured by determining the difference between a student's current score and the score that would meet standard in a set number of years (usually one to three)." (p. 7) Answers the question: "Is the gain for a group higher or lower than average?" (p. 12)
Residual Gain Scores
"…students' current scores are adjusted by their prior scores using simple linear regression….The difference between predicted and actual scores is the residual gain score and it is an indication of the student's growth compared with others in the group. Residual gains near zero indicate average growth, positive scores indicate greater than average growth and negative scores indicate less than average growth." (p. 8) Answers the question: "How much growth was produced by a group?" (p. 12)
Linear Equating
"A student's growth is defined as the student's score in Year 2 minus the student's predicted score for Year 2. A student's predicted score for Year 2 is the score in the distribution of Year 2 that corresponds to the student's Year 1 score….If the student's score is above the expected score, the student is considered to have grown. If the student's score is below the expected (predicted) score, the student is considered to have regressed. Expected growth is defined as maintaining location in the distribution from year to year." (p. 8) Answers the question: "Did students stay at the same percentile?" (p. 12)
Transition Matrix
"This method tracks students' growth at the performance standard level. A transition matrix is set up with the performance levels (e.g., Does not meet, Meets, Exceeds) for a given year as rows and the performance levels for a later year as columns. Each cell indicates the number or percent of students that moved from year 1 to year 2 levels….Transition matrices can be combined to show the progress of student across all tested years." (p. 8) Answers the question: "Are students in a group making adequate progress across performance levels?" (p. 12)
Multi-Level
"This method simultaneously estimates student-level and group-level (e.g., school or district) growth….the results of this method appear to be more complex because the methods are still unfamiliar to many people." (pp. 8-9) Answers the question: "How much of a group's growth is the result of group-level effects?" (p. 12)

[a] CCSSO (2005) identified 4 growth models – gain scores, multiple-time point growth, conditional gain score, and value added. The approaches shown here are based on CCSSO (2008), and quoted materials are directly from the 2008 document.

the student's grade level. Although these measures may be important for diagnostic purposes, they may not be the best measures for inclusion in growth models. Their basic-skills focus can inaccurately measure the performance of students who have mastered grade level higher-order skills (such as problem solving), without yet mastering lower level skills (such as computation).

Recommendations

Students with disabilities have benefited from a focus on grade-level content skills, regardless of the assessment in which they participate. The recommendations provided here retain a focus on the benefits of accountability for students with disabilities, while considering potential positive and negative consequences of different approaches to a growth model for accountability.

- **Retain a standards-based approach.** Standards-based education has a foundation in content standards of what students should know and be able to do, as well as performance standards that indicate how well students demonstrate their knowledge and skills. A focus only on growth can lose sight of the content standards and establish "growth standards" (such as a year's worth of growth) that have no connection to any particular content.

- **Maintain grade-level focus.** Approaches to measuring growth for accountability can retain their connection to content standards but still risk falling into below grade-level testing (and instruction). Below grade-level approaches can lower learning expectations in addition to poorly reflecting progress in grade-level content.

- **Include all students.** To include all students in growth models, we need to identify ways to include students with disabilities who participate in assessments other than the general assessment (e.g., the alternate assessments available to students with disabilities). Similarly, a growth model should treat students the same regardless of where they are in the performance continuum. If any students with disabilities are excluded from a selected growth model, they are likely to suffer the negative consequences of reduced access to standards-based instruction, limited focus on grade-level content, and lowered expectations.

- **Ensure that all students are visible in the results.** Growth approaches must retain transparency for the performance of all students groups. The progress of students with disabilities must not be masked by the performance of other students. In other words, the performance of students who are making good progress must not be allowed to compensate for (and thus mask) the performance of students who are not making good progress. This also means that the performance of students with only one year of data cannot be masked or ignored for accountability purposes.

Policymakers and their measurement, special education, and curriculum partners should analyze suggested options for using growth for accountability, such as those suggested by CCSSO, against the above recommendations. The results of this analysis should be articulated publicly and clearly, so that any option that is selected can be defended in light of these recommendations. Unless the selected option has been studied carefully and defended in light of these recommendations, there is a risk of unintended and negative consequences for students with disabilities.

Resources

CCSSO. (2005). ***Policymakers' guide to growth models for school accountability.*** Washington, DC: Author.

CCSSO. (2008). ***Implementer's guide to growth models.*** Washington, DC: Author.

National Center for Education Statistics. ***"Student absenteeism," in The Condition of Education, 2006*** (24-2006), Table 24-2. Washington, DC: U.S. Department of Education. (See http://nces.ed.gov/programs/coe/2006/section3/table.asp?tableID=480.)

Wagner, M., Newman, L., Cameto, R., Levine, P., & Marder, C. (2003). ***Going to School: Instructional Contexts, Programs, and participation of secondary school students with disabilities. A report from the National Longitudinal Transition Study-2 (NLTS2).*** Menlo Park, CA: SRI International. (See www.nlts2.org/reports/2003_12/nlts2_report_2003_12_complete.pdf.v.)▲

About NCEO

The National Center on Educational Outcomes (NCEO) was established in 1990 to provide national leadership in the identification of outcomes and indicators to monitor educational results for all students, including students with disabilities. NCEO addresses the participation of students with disabilities in national and state assessments, standards-setting efforts, and graduation requirements.

The Center is supported primarily through a Cooperative Agreement (#H326G050007) with the Research to Practice Division, Office of Special Education Programs, U.S. Department of Education. Additional support for targeted projects, including those on LEP students, is provided by other federal and state agencies. The Center is affiliated with the Institute on Community Integration in the College of Education and Human Development, University of Minnesota. Opinions expressed herein do not necessarily reflect those of the U.S. Department of Education or Offices within it.

NCEO Staff

Martha L. Thurlow, Director
Deb A. Albus
Jason R. Altman
Manuel T. Barrera
Laurene L. Christensen
Christopher J. Johnstone
Jane L. Krentz
Sheryl S. Lazarus
Kristi K. Liu
Ross E. Moen
Michael L. Moore
Rachel F. Quenemoen
Christopher Rogers
Dorene L. Scott
Miong Vang
Yi-Chen Wu

The University of Minnesota shall provide equal access to and opportunity in its programs, facilities, and employment without regard to race, color, creed, religion, national origin, gender, age, marital status, disability, public assistance status, veteran status, sexual orientation, gender identity, or gender expression.

NCEO *Policy Directions* is a series of reports that address national policy issues related to students with disabilities. This report was prepared by Martha Thurlow, Sheryl Lazarus, Rachel Quenemoen, and Ross Moen. It is available in alternative formats upon request.

NCEO
207 Pattee Hall
150 Pillsbury Dr. SE
University of Minnesota
Minneapolis, MN 55455
Phone: 612-626-1530
Fax: 612-624-0879

Visit our Web site at:
http://www.nceo.info

Assessment Accommodation Misperceptions

Often Used Concepts

Accommodation = a change in materials or procedures that **does not change** the targeted content.
Modification = a change in materials or procedures that **does change** the targeted content.

Misperception #1 – All students with disabilities need test accommodations.

Assumptions Underlying Misperception: People sometimes assume that all students with disabilities need to have testing materials or procedures changed, and the need is directly related to the student's category of disability and severity of disability.	**Research Response:** Accommodations are provided to meet the specific needs of certain students with disabilities, as a way to provide access to the test and to obtain a valid indication of their knowledge and skills. Not all students with disabilities will need accommodations for a test. The need for accommodations is not determined by having a disability, by a student's category of disability, nor by the severity of the student's disability. It is determined by the extent to which the conditions of the assessment interact with the student's specific disability-related characteristics that inhibit the student's performance on the test.
Recommendation: Provide training about how the needs and characteristics of students with disabilities interact with the academic content they are learning. This training should include demonstrations, modeling, and other examples that show the interaction of student needs and characteristics and the content they are learning.	

Misperception #2 – Only students who have IEPs (Individualized Education Programs) need assessment accommodations.

Assumptions Underlying Misperception: People sometimes assume that only students who receive special education services should be allowed to use assessment accommodations.	**Research Response** – Assessment accommodations may be used by students who have Section 504 accommodation plans and English language learners, as well as by students who receive special education services. In addition, state policies may recognize that other students should be allowed to use accommodations, specifically those for whom accommodations would increase their opportunity to be able show knowledge and skills without undermining the intent of the assessment. Need for access, without undermining what an assessment is intended to measure, should be the major determinant of whether accommodations are needed by a student.
Recommendation: Create assessments with as few barriers as possible for all students through universal design of assessment approaches, and then allow any student with a demonstrated learning-related need to use accommodations (not modifications) during testing.	

Misperception #3 – If a student uses an accommodation during instruction, then the student is entitled to use the same change in materials and procedures during testing.

Assumptions Underlying Misperception: People sometimes assume that changes in materials and procedures that a student uses during assessments are supposed to be the same as instructional accommodations, and the IEP gives students with disabilities the right to receive the same changes during both instruction and assessment.	**Research Response:** Assessments have a specific purpose – to measure a student's knowledge and skills in a way that produces valid results. Provision of accommodations to students who need them is an important part of obtaining valid results. However, if an accommodation is used that compromises what an assessment is designed to measure, it becomes a modification (a change in materials or procedures that changes what the assessment intends to measure). Even though it may be appropriate for instruction, it is no longer appropriate for use during the assessment. A student's IEP allows the student to use instructional accommodations that provide access to the general curriculum and assessment accommodations that improve the validity of assessment results, but not modifications that compromise the purpose of the assessment.
Recommendation: Decisions about accommodations should be made separately for instruction and assessments. Although there should be alignment between instructional and assessment accommodations, the specific changes in materials or procedures that are used during the two may not be exactly the same.	

Misperception #4 – Providing students with assessment accommodations results in an unfair advantage for those students.

Assumptions Underlying Misperception: People sometimes assume that assessment accommodations alter what the test measures, giving students who use them an unfair advantage over other students.	**Research Response:** Accommodations are intended to produce a more accurate measure of a student's knowledge and skills. Research and logical justifications are used to provide evidence that this is the case. If it is not (i.e., if what the assessment is intended to measure is altered), the change is referred to as a "modification" or an accommodation that produces invalid results. Assessment accommodations and modifications are not the same. Modifications represent a change in what is measured and scores from assessments taken with modifications cannot be interpreted in the same way as scores from assessments with and without accommodations.
Recommendation: Carefully identify what the test is intended to measure, then begin the design of the assessments from a universal design perspective (i.e., one that considers all students from the beginning in the design of the assessment). When necessary, allow accommodations that do not change what the assessment is intended to measure.	

Misperception #5 – Students who perform poorly are more likely to need accommodations during testing.

Assumptions Underlying Misperception: People sometimes assume that the purpose of accommodations is to increase a student's score on the assessment.	**Research Response:** The purpose of assessment accommodations is to provide students access to the assessment so that they can show their knowledge and skills most accurately. Better measurement and valid assessment results do not necessarily mean higher scores for the student.
Recommendation: Provide training to ensure that policymakers and educators understand the purpose of assessment accommodations. Offer training in accessible formats, and include test developers, state assessment personnel, educators, and parents in the training effort.	

Misperception #6 – Any change in assessment materials or procedures that lacks an experimental research base to support it should be considered a modification.

Assumptions Underlying Misperception: People sometimes assume that assessment accommodation policies must allow only those changes to assessment procedures or materials that have been shown by experimental research to produce valid scores.	**Research Response:** Although experimental research is desirable as a basis for accommodation policies, there are many changes in testing materials or procedures for which it is extremely difficult to undertake an experimental study. All accommodation policies should be based on a logical justification that considers what the assessment intends to measure and whether a specific change compromises that intent, regardless of whether an experimental study has been conducted.
Recommendation: When formulating accommodation policies, review available research, identify the limitations of the research, examine existing data on accommodations use and unintended uses, and develop sound justification for allowing or not allowing a specific accommodation.	

Misperception #7 – A specific set of assessment accommodations exists that all states should include in their policies.

Assumptions Underlying Misperception: People sometimes assume that there is one appropriate and valid set of test accommodations.	**Research Response:** It would be nice if it were that easy. Appropriate accommodations for an assessment depend on numerous factors, including an individual student's learning-related characteristics and needs, and what the assessment intends to measure. Because state content standards and assessments vary from one state to another, there is no single set of appropriate and valid accommodations.
Recommendation: Clarify the purpose of each assessment and what it is intended to measure, and base accommodation policies on these.	

Misperception #8 – Students with disabilities consistently receive the assessment accommodations identified in their IEPs or Section 504 Plans during testing.

Assumptions Underlying Misperception: People sometimes assume that providing accommodations is a natural part of providing educational services to students with IEPs.	**Research Response:** There are many challenges that schools face in the appropriate provision of accommodations. Research has shown that educators sometimes believe that providing accommodations is too logistically challenging and expensive to be practical. It also has shown that many students do not receive the accommodations that are listed in their IEPs, and that they also may be provided accommodations that are not listed in their IEPs because of administrative convenience.
Recommendation: Ensure that educators, including administrators, are trained on the importance of making the provision of accommodations a natural part of the educational process.	

Misperception #9 – After accommodations needed by a student are identified, the accommodations should remain the same for the student's entire educational career.

Assumptions Underlying Misperception: People sometimes assume that students never change in their need for specific accommodations, nor is there a need for accommodations to vary for different content areas.	**Research Response:** Over time, students' accommodation needs may change as the impact of their disability changes over time. The purpose of different assessments and what they intend to measure also may be different and may change across grade levels, resulting in changes in what is appropriate or allowed for the student to use as an assessment accommodation during the student's educational career.
Recommendation: Policymakers should carefully consider changes in accommodation policies across grades. IEP teams should consider the changing needs of the student over time, in light of the assessments the student will take, so that appropriate accommodations are identified for the time covered by the IEP.	

Misperception #10 – There is really no good way to figure out what specific accommodations an individual student needs.

Assumptions Underlying Misperception: People sometimes assume that it is too difficult to figure out which accommodations a student needs, and then to check these against assessment accommodation policies.	**Research Response:** Making good accommodation decisions for an individual student requires careful consideration of the student's learning-related characteristics, the purpose of the assessment, what it intends to measure, and the task demands of the assessment. It also requires that educators check out accommodations in the classroom to ensure that they are effective for the student. In addition, it is helpful to gather input from parents and students on the usefulness of accommodations and the student's willingness to use accommodations. Doing these things will help avoid non-beneficial over-accommodation practices.
Recommendation: Provide training on accommodation decision making for educators, parents, and students.	

Myth Busters!

Researching Out-of-Level Testing

~ 10 Recommendations for States That MUST Test Out of Level ~

Myth #1 -- More students participate in statewide testing when tested out of level.

Assumption -- Testing students with disabilities out of level is a good way to include students who cannot participate fully in the regular assessment - even with accommodations.	**Research Response: States struggle with how to include out-of-level test scores in accountability indices. Students participate in state tests, but are then excluded from accountability results.**
Recommendation: Create universally designed assessments.	

Myth #2 – Out-of-level testing is good for instructional decision making.

Assumption – Testing students with disabilities at their instructional level provides more usable information for instructional planning.	**Research Response: Special educators report that they do not use the results from out-of-level tests. Some test companies only provide raw scores for out-of-level tests with no other test results. Often times, states' test results are not available until six months after testing when the information is no longer relevant for instructional planning.**
Recommendation: Work with test contractors to get meaningful and timely test results.	

Myth #3 -- Students receive tests with items that they can answer.

Assumption -- Out-of-level tests are challenging standards-based assessments.	**Research Response: One state indicated that too many students were tested out of level at too low of a test level. Some students with disabilities report that out-of-level tests are too easy.**
Recommendation: Analyze out-of-level test participation and performance to ensure that unreasonable numbers of students are not tested below grade level.	

Myth #4 -- Out-of-level test scores are more valid.

Assumption -- Out-of-level tests measure what students actually know, which eliminates guessing during testing.	**Research Response: Some teachers and students with disabilities report that out-of-level tests are not taken seriously when the test items are not developmentally appropriate.**
Recommendation: Limit the number of levels below grade level at which an out-of-level test can be administered to no more than one level below grade level.	

Myth #5 – Out-of-level testing is better for students' self esteem.

Assumption -- Taking a test that is too hard for students with disabilities is frustrating, embarrassing, and negatively affects their level of self confidence.	**Research Response: Special educators report that teachers, parents, and students do not think about the future consequences of taking state tests out of level. A student with disabilities who is tested out of level may never be able to meet the requirements for high school graduation - a consequence that permanently impacts students' dreams, aspirations, and self worth.**
Recommendation: Provide a written document for parents and teachers that states clearly how out-of-level testing could affect students' academic futures. Have parents and students sign it.	

Myth #6 -- IEP teams work together to select students for out-of-level tests.

Assumption -- Most states have written assessment policies that require IEP teams to select students with disabilities for out-of-level tests according to specific criteria.	**Research Response: Special education teachers report that they often make the decision to test a student out of level independently before the IEP team meeting convenes.**
Recommendation: Provide a decision-matrix that contains state criteria for guiding an IEP team through the decision to test a student out of level.	

Myth #7 – Test scores improve when students are tested out of level.

Assumption – Out-of-level tests yield better student test performance because students are taking tests that are matched to their ability level.	**Research Response: Students who took the <u>same</u> norm-referenced test in level and out of level performed worse on the out-of-level test. One states' data showed a similar pattern when out-of-level and in-level test scores were compared between <u>two</u> norm-referenced tests.**
Recommendation: Do not assume that out-of-level tests will improve students' test performance.	

Myth #8 – Test companies have adequate out-of-level testing instruments.

Assumption – All test companies publish standardized tests that can be administered and interpreted as is.	**Research Response: Test companies vary in the specificity of information provided for making decisions about interpreting and using out-of-level test scores.**
Recommendation: Work with test companies to gather the necessary technical assistance to guide the interpretation and use of out-of-level test scores.	

Myth #9 – Equating out-of-level test scores yields accurate on-level test scores.

Assumption – Test companies provide accurate equating functions for test score conversion.	**Research Response: Test company representatives are unable to provide a statistically derived estimate of the amount of error that was introduced through the equating process developed for their instrument.**
Recommendation: Interpret out-of-level scores that are converted to on-grade test scores cautiously so that the unknown error is acknowledged.	

Myth #10 – Out-of-level test scores are easy to locate in states' reports.

Assumption – Out-of-level test scores are reported at the district and state level in the same manner as scores from the regular assessment are reported.	**Research Response: Some states do not analyze out-of-level test scores at all. Other states aggregate out of-level test scores with other types of test scores.**
Recommendation: Report out-of-level test scores in a clearly identifiable format.	