Race to the Top Assessment Program
Written Input
Submitted January 18-20, 2010

Author	Title	Date Submitted
Tim Crockett	Race to the Top Assessment Considerations	1/19/2010
Kathleen Leos	Race to the Top Assessment-Public/Expert Input Meeting	1/19/2010
Denis Udall	What a Federal RFP for New Assessments Should Require and Encourage	1/19/2010
Larry Berger	Race to the Top Assessment-Procurement Comments	1/20/2010
Linda Darling-Hammond	Race to the Top Assessment Comments and Three Attachments	1/20/2010
Nicholas C. Donohue	Race to the Top Assessment Competition Public and Expert Input Meetings High School	1/20/2010
David Foster	Input for USED on General and Technical Assessment	1/20/2010
John C. Foster	Race to the Top Assessment Meeting: Technical Assessment	1/20/2010
Susan Gendron	Project and Consortium Management	1/20/2010
Nancy Green	Statement of NAGC before the U.S. Department of Education Assessment Meeting	1/20/2010
Myrtice J. Irish	A Submission to the Assessment Session	1/20/2010
Barbara Kapinus	Testimony of Barbara Kapinus, Senior Policy Analyst, Education Policy and Practice Department, National Education Association	1/20/2010
Christine Y. Mason	No Title	1/20/2010
John F. Olson, Ed Roeber, and Barry Topol	The Cost of New Higher Quality Assessments: A Comprehensive Analysis of the Potential Costs for Future State Assessments	1/20/2010
Richard J. Patz	CTB McGraw-Hill Input on General and Technical Assessment Topics	1/20/2010
Lauren B. Resnick	In Response to Request for Public and Expert Input, Topic 3.C. General and Technical Assessment	1/20/2010
Bill Ring	Comments on General Assessment-A Parent Perspective	1/20/2010
John S. Twing	Lessons Learned and the Road Ahead for the American Diploma Project (ADP) Assessment Consortium	1/20/2010
Rosa Villegas	Race to the Top (RttT) Hearing	1/20/2010

Race to the Top Assessment Considerations

Comments by Measured Progress, Inc.

Tim Crockett, Senior Vice President Client Services

January 19, 2010

Currently there are a few assessment consortiums in the United States, but only the New England Common Assessment Program (NECAP) has addressed all NCLB content in grades three through eight. Measured Progress is proud to have been the contractor providing leadership from the assessment vendor side, supporting the state DOE leaders in NECAP since its inception. In 2002, as a result of the passage of NCLB, New Hampshire, Rhode Island, Vermont and later Maine combined resources to build the NECAP consortium. The states' commissioners and deputy commissioners recognized that their existing tests would not meet NCLB requirements. However, by forming a consortium the states could realize efficiencies and develop better quality tests for less money. This also meant that the states could afford to include constructed-response items on their tests, which was important in maintaining the quality of the assessments.

Our experience working with the NECAP states has taught us a great deal about forming, managing, and maintaining state assessment consortia. We have found that there are key components that contribute to the success of any consortium, including the following:

- Commitment—it is a lengthy and time-consuming process. States must agree on many things from types of items on the tests, to administration windows, to cost-sharing formulas.
- Shared values—it is helpful to have similar educational philosophies and ideas about what constitutes high-quality assessments.
- Geographical proximity—states within close proximity can meet more easily face-to-face. Geographic commonality allows for the use of "regionally flavored" content (e.g. reading passages on cold weather as opposed to hurricanes) which can boost student engagement and motivation during testing.
- Strong leadership—support from the highest levels helps solidify relationships and strengthen the collaboration.
- External management—an independent facilitator offers support in planning meetings, serves as mediator, and ensures that one vision is conveyed to the testing contractor.
- Identical tests—ensuring that the core test that is administered in all states is identical and deviations from this can compromise cost benefits.
- Cost-sharing formulas—it is important to determine how the states will pay for and share the various costs associated with developing

assessments. States should share fixed costs (such as development) and divide variable costs (such as shipping and scoring) according to number of pupils tested in each state.

Project and Consortium Management Questions

1. How would you recommend organizing a consortium to achieve success in developing and implementing the proposed assessment system?

Merriam-Webster defines a consortium as an agreement, combination, or group … formed to undertake an enterprise that would exceed the resources of any one member.[1] There are several reasons states may desire to enter into a consortium, all of which will influence the success such a venture. As described by Charles A. DePascale of the National Center for the Improvement of Educational Assessment (NCIEA) those reasons include

- the belief that a consortium of states working together is more efficient and cost effective than individual states developing their own assessments;
- the conviction that a consortium of states working together will produce higher quality assessments than any individual state working on its own;
- the belief that a consortium of states working together increases the equity in assessment resources available across states; and
- the desire for national, cross-state comparisons on a common assessment produced by a consortium of states.

In the Race to the Top assessment information published in the Federal Register, the Department of Education states that the competitive assessment program will "support one or more consortia of States that are working toward jointly developing and implementing common, high-quality assessments aligned with a consortium's common set of kindergarten-through-grade-12 (K–12) standards that are internationally benchmarked and that build toward college and career readiness by the time of high school completion."[2]

It is quite possible that state assessment consortia, as developed under the Race to the Top assessment program, will be developed to fulfill one, two, or all of the conditions listed above. However, it is also possible for consortia to be formed that will, in the end, meet none of these criteria.

Multi-state assessments can work well for states desiring to reduce costs while maintaining the quality of their statewide accountability program.

---

[1] Retrieved January 7, 2010 from **http://www.merriam-webster.com/dictionary/consortium**

[2] Department of Education Race to the Top Fund; 74 Federal Register 249 (30 Dec. 2009) pp. 69081 – 60984.

When states can come together and share a similar educational and assessment philosophy, it is possible for multi-state consortia to grow and flourish. Consortia can be formed to create common end-of-course, English language learner, and alternate assessments. Initial collaboration on smaller testing programs with higher per-pupil costs may pave the way for larger-scale cooperative efforts. States can learn how well they can work together on a modest initiative before they collaborate on entire testing programs. States looking to try out new testing programs—for example, assessments that measure 21st-Century Skills or student progress over time, may consider working with other states to build these programs from the ground up.

Collaboration is the crucial component that must be present in any consortium. States must agree to work together—side by side and as equals—with no state exerting more power or influence than the others. Consortia states must have common goals and educational philosophies, and these must be determined before any consortium is formed.

How Did NECAP Begin?

Measured Progress invited all six New England states to an invitational meeting in the fall of 2001. Congress had recently passed the No Child Left Behind Act, requiring states to test all students in grades three through eight and one grade in high school. Looking for ways to help states comply with NCLB without taxing their budgets, Measured Progress introduced the idea of forming a multi-state assessment. The states of Maine, New Hampshire, Rhode Island, and Vermont agreed to explore collaboration.

In spring of 2003, Vermont issued a Request for Proposal to develop content standards in reading and mathematics, and Measured Progress was awarded the contract. Vermont published its standards, opening the door for the other states to do preliminary work on establishing common Grade Level Expectations (GLEs). The states collectively worked with the National Center for the Improvement of Educational Assessment (the Center) to complete GLE work in the summer of 2003. By the fall of 2003, the states had common standards in reading, writing, and mathematics, and issued a Request for Proposal to develop a common assessment for students in grades three through eight for all states. In December of that same year, Measured Progress was awarded the contract and began working on a pilot test for grades three through eight.

The NECAP pilot assessment for grades three through eight in reading and writing was administered in the fall/winter of 2004–2005. Measured Progress staff scored the assessment, and the results were used to build test forms for the first NECAP operational assessment, administered in October 2005. States released results from the operational assessment to schools and districts in early 2006, after standard setting.

The NECAP program next expanded to test students in grade 11. Measured Progress and the states agreed on a contract amendment and administered the

NECAP grade 11 pilot in November 2006. The grade 11 test became operational in October of 2007.

As the requirements of NCLB next required science testing at three grade levels, in spring of 2006, the states came together again to issue a Request for Proposal to test students on common science standards. Measured Progress was awarded the contract and administered the first operational NECAP science test in spring 2008. In 2009 Maine became the fourth state to join the NECAP consortium.

Hopefully this historical perspective demonstrates that NECAP was developed over several years and provides a glimpse into the extensive work that has been accomplished in phases over that time period. Each of the tasks mentioned involved careful planning, with the states and Measured Progress learning together and thriving from the experience of building a highly regarded, shared assessment program.

a. What governance model do you suggest and why? What leadership model do you suggest and why?

Based on our experience as contractor for NECAP, the most appropriate governance model is one in which member states are partners, each having an equal role in the management of the consortium. A Memorandum of Understanding could be executed to define the relationship and to outline

- goals and purposes of the consortium;
- rules regarding decision making;
- leadership structure;
- distribution of costs (fixed and variable);
- cost sharing formulas; and
- other roles and responsibilities.

There are many reasons NECAP has been successful. Among these are that the members have worked harmoniously and have been able to reach consensus on all important aspects of the program. Each NECAP state also maintains a contract with the National Center for the Improvement of Educational Assessment (NCIEA), separate from their Measured Progress contracts, to provide support to the program.

b. What recommendations do you have on the decision-making process within a consortium?

Ideally, decisions should be made based on group consensus. This approach has worked very well for NECAP for many reasons. Because of the proximity of the states to each other, as well as to the Measured Progress offices, it is possible to conduct face-to-face meetings—often on a relatively short notice. Participants are often able to problem-solve and come to resolution more easily when they are sitting together in the same room, especially at the beginning of the relationship.

In addition, the NECAP states clearly articulated program goals at the start. Maintaining well-understood and -documented goals ensures that participants have the same goals and that they work collaboratively to achieve them. The third-party program manager should work with the states to ensure that decisions are made in a timely manner and that they are in the best interest of the consortium.

c. What recommendations do you have for States that are organizing consortia regarding:
- **how to differentiate roles, responsibilities, and workloads within a consortium?**
- **roles for third parties (e.g., conveners, project managers, assessment developers/partners, intermediaries)?**

Ideally, all state consortia will have staff members who have the same roles, responsibilities, and workloads. However, budget constraints within state departments of education may mean that not every department is staffed in the same manner or some departments will have staff with expertise in one area and not another. States should work together to identify their strengths and weaknesses and find ways to use their available resources to the benefit of the program.

As mentioned earlier, we have found the use of NCIEA, the third-party manager, to be essential to the smooth operation of the NECAP program. The NCIEA program manager attends NECAP meetings, organizes and oversees the Technical Advisory Committee, provides technical guidance to the states, and acts as liaison among the states to provide essential, yet impartial advice.

It is important to note that the third-party program manager was hired by the consortium, independently of Measured Progress, and is under separate contract with each NECAP state.

During formation of each consortium, members should be provided with a set of guidelines and guiding principles including the following:

- If one state leaves the consortium, the program will terminate unless there is agreement for the other states to pick up the cost for that state (redistribute costs).
- No state has more power or say than the others.
- All states have equal ownership of all aspects of the program.
- The program should be identical in all states to realize cost savings.

d. What advice do you have on the characteristics that all consortium members must have in common in order for a consortium to be successful, and what characteristics can vary across member States, e.g., philosophical approaches to assessment, standards, scope and sequence, etc.?

Collaborating with another state or states to create a common assessment consortium is no small task. Below are some of the key points we feel consortium states must have in common.

- Similar educational and assessment philosophy
- Approval of the assessment consortium concept by political leadership at the highest levels of state government
- Strong assessment directors and staff who will play significant roles
- Agreement on common assessment attributes:
  - Standards (content and achievement)
  - Test design
  - Quality
- Desire for cost savings and agreement on a cost formula. Programs must be identical in all states to realize optimum cost savings.
- Patience. Tasks take longer and states must plan accordingly.
- Good will, trust, and a spirit of compromise

Several favorable conditions were present with the NECAP states which enabled Measured Progress to assist in developing the assessment program. The states shared similar educational and assessment philosophies and had relatively small student populations (although the largest was twice the size of the smallest). As mentioned earlier, geographic proximity, though not necessarily a requirement, has also been a helpful factor in the success of NECAP. The common need to save costs and meet NCLB requirements while continuing to provide a high-quality assessment program was met when the states joined the NECAP consortium.

2. **What would you recommend that a consortium be asked to demonstrate in its application to show that it has the capacity, structure, and potential to implement its proposed plan? What are the critical success indicators six, 12 and 18 months into the life of a consortium? What signals are predictive of ultimate success or failure?**

First and foremost states must demonstrate that they have common goals and a willingness to collaborate to achieve them. A clearly articulated Memorandum of Understanding (MOU) should be developed to outline the goals and purposes of the consortium, as well as business details including

- cost sharing formulas;
- governance:
  - leadership
  - roles and responsibilities of each member; and
- organizational structure.

The obvious indicators of success will be that after six, 12, or more months consortium members are achieving their goals and member states are not leaving the consortium. The early months of the consortium are when it is absolutely critical to have a third-party manager, particularly in consortia where members are not well acquainted with each other. This program manager will be a trusted advisor to all parties who will act in the best interests of the consortium.

3. What could go wrong in the development and management of a consortium and what can States do to mitigate these factors up front? In what ways could the Department structure the competition to help mitigate these factors?

While it is understood that the Department believes consortia-led assessment programs can make a major contribution to the future of assessments, it would be unrealistic to expect all consortia to be universally successful. The expectations that will define success should be made clear from the outset.

Currently, state assessment programs run the gamut, from all multiple-choice formats to models with a combination of multiple-choice, constructed-response, and even performance assessments. Some states have set the bar very high for student performance, while others have set the bar lower. The level of funding for assessment varies from state to state, allowing some states to have content specialists, measurement experts, and program managers on staff while others have few of these resources.

Perhaps the development of a consortium could be seen as a *merger* of state assessment programs. A set of common goals must be developed and the extent of cooperation among the states must be determined. As in any merger, it will be important to evaluate the attributes and challenges that each state brings to the consortium and to allocate resources appropriately and fairly. All of these issues can be resolved with a detailed MOU, which should be a required for each consortium developed. The Department of Education might consider the degree to which each state is thoroughly involved in the partnership and outlines a procurement process that seeks the involvement of experienced contractors to provide program support as criteria for award.

Procurement Questions

1. How do differences in State procurement rules affect how you would design and manage a consortium? How will State procurement regulations, processes and time frames likely affect how a consortium carries out the development, piloting, and implementation of common assessments? (You may use examples from outside the education sector, if relevant.)

It is clear that individual states' procurement rules would need to be considered in the design and management of any consortium. Currently, each NECAP state has a separate contract with Measured Progress. In this way,

each state's unique procurement laws are addressed, and this model works well. Despite having separate contracts, each NECAP state agreed to compromise on certain practices in order to meet the objectives of the collaboration.

However, if viewed on a larger scale with the distinct possibility of much larger consortiums (both in the number of states and larger student populations) being formed, it could become cumbersome to both vendors and the states to maintain separate contracts. Therefore, there are numerous factors that should be considered and addressed by the consortium in order to anticipate and accommodate the impact of individual procurement rules.

- Members should examine each state's procurement rules to identify rules that are common to all members of the consortium.
- States should agree on a group process for determining the project scope of work.
- States should agree on a cost-sharing formula in advance that specifies how fixed and variable costs are shared, and how this formula would be affected should one state pull out of the consortium. There must be a process in place to ensure accurate student counts for each state that will affect the cost-sharing model.
- States should agree on all project milestones and timelines.
- States should agree on a process for issuing RFPs that will take common requirements and rules into account and allow for each state to include a minimal amount of unique requirements (i.e. language translations of test materials, reporting).
- States should implement a proposal evaluation team process that includes members from each state who are empowered to make a vendor selection(s) for their state.

A clear, coherent RFP process is essential for ensuring a competitive, fair process for states/consortiums and vendors. There are two possible models, each with its own advantages and disadvantages.

The first, used currently in NECAP, involves the issuance of an RFP from each member state. RFPs issued by each state would be significantly similar to the others, but state-specific requirements would be addressed. This would allow the RFP process itself to address the individual state's procurement rules and regulations. While this process has been successful with NECAP it could become quite cumbersome with larger consortiums.

The second involves the issuance of one RFP by the consortium as a whole. With this solution, differences in state rules could be addressed with addenda for each state included with the RFP. Of course depending on the circumstances, this approach could also prove to be very cumbersome.

2. States have expressed interest in acquiring information about, and input and ideas from, potential assessment partners/vendors in

advance of completing their applications. What actions, if any, would you advise the Department to take to help facilitate this?

There are a number of actions the Department may take to advise states as they prepare their response to the summative assessment RFR. Competitive solicitations are at the core of establishing partnerships with states. Fundamental to these solicitations is the clear articulation of requirements and program goals. As a preliminary activity, prior to the Department release of the RFP, the Department could encourage states to release an RFI seeking input regarding vendor products and capabilities. It will be very important to ensure that the RFIs are competitively released to all qualified vendors.

Another approach the Department could take would be to release an RFI to assessment vendors as part of the solicitation to states, requesting specific information and ideas. Responses to the RFIs could be submitted to the Department and shared with states upon their request.

A slightly different approach for the Department to consider would be to allow assessment partners/vendors to respond directly to aspects of the Department's RFP as they relate to services, integrated solutions and products. Providing this as a requirement or as an option might change the state-specific RFP requirements.

The Department could also develop a survey to be administered to state and district assessment directors asking for input regarding their current assessment systems. Assessment directors would provide information about vendors, processes, procedures, tools, and technologies. The results of the survey could be made available to state and district assessment offices and could be used to obtain additional information prior to issuing RFPs.

Finally, the Department could sponsor regional conferences and/or exhibits where vendors would be given the opportunity to display and describe their products and processes and states can interact with vendors.

Whichever method the Department selects, it is crucial that all qualified vendors, both large and small, receive consistent, timely notification and opportunity to respond in order to maintain the fairness of the process.

3. **States expressed a desire to run competitive (as opposed to sole sourced) processes for selecting partners/vendors. How would you advise the Department to structure the application to enable States to do so? What other ideas would you offer in designing a process that is flexible enough to accommodate other challenges that States might encounter over time (e.g., challenges related to partner/vendor selection or contract change management)?**

A competitive bid process is essential to ensuring that qualified vendors bid and that fairness to both the consortia and vendors is maintained. Over time, if consortia grow too large, smaller vendors would eventually not be able to compete, and this could force multiple vendors to form partnerships in order

to provide the required services. This in turn could result in anti-trust issues and restrain a competitive environment. The Department could issue application rules requiring each consortium to identify a proposed scope of work. The Department should prohibit states from identifying a preferred vendor during the application process. Instead, the Department could allocate monies only to consortia that agree to a competitive bid process (i.e. issuance of an RFP). During the proposal evaluation process, unanimous agreement among states in a specific consortium should be required when selecting a vendor. By the same token, all members would need to agree to any contract scope changes once a contract is in place.

The Department could also issue guidelines restricting the total size allowable for each consortium, preferably by number of students tested rather than by number of states. It is our belief that smaller consortia will ensure more control at the state level and minimize potential for disagreement among members. By limiting consortia size, the Department can also ensure that no vendor in the industry is unfairly disadvantaged solely by their size.

RACE TO THE TOP ASSESSMENT – PUBLIC/EXPERT INPUT MEETING January 13, 2010

Background and comments submitted by: The Global Institute for Language & Literacy Development, LLC.
Kathleen Leos, President

Good Afternoon,
I am Kathleen Leos, former Assistant Deputy Secretary for the U.S. Dept of Education, No Child Left Behind, Title III Office of English Language Acquisition (OELA), and currently serve as the President for the Global Institute for Language and Literacy Development (GILD). Thank you for giving us an opportunity to address this vitally important topic and for considering the input we are providing you.

Background:

- Since FY 2002, states have worked hard to develop the first generation of English Language Proficiency (ELP) standards and aligned ELP assessments for ELLs. However, in the first attempt to actualize the legislation, not all states fully developed or aligned the state ELP standards to the academic content standards which left a significant linguistic (language) gap and hence countless ELLs unable to access or master the necessary grade level academic knowledge in each content area.
- A new approach is needed in developing the second generation of standards and assessments for English language learners based on the original legal premise outlined in the 1974 Supreme Court Case, Lau v Nichols that mandates that "language may not be a barrier to academic content knowledge for English language learners".

First Step - A Paradigm Shift & New English Language Proficiency Standards:
To appropriately include ELLs in state standardized assessments, it is important to begin with the Standards. Forty-seven states have signed an agreement to participate in the development of National Common Core Content Standards in English Language Arts and Mathematics. In order to fulfill the requirements set forth in Lau, ELLs must be provided "Access to Academic Content" and "Opportunities to Demonstrate Mastery of Academic Content" on grade Level, based on their level of English language proficiency. Therefore, it is an optimal time to support state efforts to simultaneously develop National Common English Language Proficiency Standards that are ALIGNED to the National Common Core Content Standards; thus providing a foundation for ensuring compliance with the law, and a framework for narrowing the achievement gap between ELLs and non-ELLs.

A focus on language development and language acquisition needs to form the basis for the second generation of English language proficiency standards. Experts with deep background knowledge in the science of language development and an understanding of integrating the essential components of language fused through the lens of second language acquisition levels is a necessary step in developing the second iteration of ELP standards. By integrating language development and language acquisition levels and aligning them to the academic content area on grade level, ELLs will have the linguistic bridge needed to access, comprehend and demonstrate mastery of the academic knowledge needed to succeed, and for the first time the mandate defined by *Lau* can be realized.

This is not an easy task. Most states do not have the resources or the capacity to tackle this additional challenge without the support of consortia, collaborative groups and experts with proven knowledge, not only in the theoretical construct, but the practical application at all levels of the linguistic and academic spectrum.

Second Step - New English Language Proficiency Assessments:

The development of new ELP Standards will necessitate a second generation of ELP assessments that are aligned to the new ELP standards. The new assessments will provide states the opportunity to accurately measure progress ELLs are making toward attaining full English language proficiency based on standards and can provide transparent information and data that will support state decisions related to attainment and exit criteria from language education instruction programs for English language learners.

It is important to note that new ongoing research in this area indicates that ELP standards and assessments developed and implemented in this manner can act as predictors to levels of achievement on state academic content assessments. This is not to say that the ELP assessment can replace the academic content assessment or that one assessment can be used to fulfill accountability reporting requirements for both language acquisition and content.

On the contrary, each assessment is designed for very different purposes and must be adhered to accordingly. A valid and reliable English Language Proficiency Assessment alone cannot effectively measure academic progress of ELLs and conversely, a valid and reliable content assessment alone cannot effectively measure language acquisition progress. Each assessment is designed to provide educators insight into very different processes.

The definition of progress, attainment and academic achievement must be considered separately and should be based on the two parallel demonstrations of mastery defined by each assessment. Therefore, it is not only about developing new assessments, but how new assessments can be used to revolutionize instruction and accountability for ELLs.

Third Step – A New Vision for ELL Accountability:

The development of a new generation of English language proficiency standards that embody essential components language develop integrated with the levels of language acquisition; aligned to grade level academic content standards offers unique opportunity to change Title I and Title III accountability systems for English language learners.

The ability to accurately measure and report on "what a student knows and can demonstrate" in academic content while clearly differentiating an ELL "making progress" and "attainment" in acquiring a second language can have a profound positive impact on the Title III accountability system for ELLs.

NCLB requires states and districts to annually assess and report on all students' ability to attain proficiency in academic grade level content <u>and</u> annually assess and report on an ELLs' progress toward mastering English language proficiency and attainment of English language proficiency. The two processes are separate and distinct but have a profound impact on each other. An ELL cannot

demonstrate full mastery in academic content in English without attaining full English language proficiency. These are not two separate learning processes for an ELL. However, for assessment, reporting and accountability purposes the processes are distinctly different.

In the past, experts have discussed the importance of using linguistic accommodations that mitigate linguistic barriers to access the academic content or reduce the "language load" and not invalidate the assessment. Part of the reason for the need for linguistic accommodations is because ELLs were not taught all the necessary elements of the English language that would provide direct access or a linguistic bridge to the grade level academic content. Hence, it was assumed that ELLs, considered a special population (like SWD) needed some type of accommodation to access the content. Following that theory educators deferred to the special education community and adopted over 70 testing modifications and accommodations for English language learners. Current evidence-based research indicates that the majority of these invalidated a content assessment for ELLs.

The Global Institute for Language and Literacy Development (GILD) offers a different accountability model for ELLs that effectively measures both processes and addresses the complex dynamic of the interplay between English language acquisition and content mastery in English with or without accommodations by combining both Title I and Title III accountability in a different measurement system than what has been used in the past.

Proposed Weighted Index Model: A New Approach to NCLB Accountability for ELLs

Since ELLs must learn English language acquisition and grade level academic content simultaneously and educators can concurrently measure and report the dual nature of ELL language and content mastery; it is important for the measurement experts to consider an accountability system for ELLs in the manner that it is taught and assessed.

One methodology is to establish the Annual Measurable Achievement Objectives (AMAO's) for Title III and the Adequate Yearly Progress (AYP's) for Title I by treating both systems as one. The first step is to take the two AMAO's from Title III – measuring progress and attainment toward English language proficiency and give them more statistical weight for ELLs who are at the beginning levels of English language proficiency. The academic content score from the state annual standardized content assessment would still be included in the AMAO accountability system, but more weight would be assigned to the two Title III language acquisition scores and a lesser weight assigned to the Title I academic content score while the student is still acquiring the English language. The "weight" would shift toward academic content scoring as the ELLs continue to master the English language.

This methodology would require considering a cut score within "making progress" and "attainment" of English language proficiency at a level that could indicate that the ELL is less dependent on learning the English language to demonstrate content mastery and has achieved a level of understanding and mastery of the content in the language it is taught and assessed. This method is fluid; it changes over time and accurately reflects the dual nature of learning language and content simultaneously by accounting for the fluid nature of the learning dynamic.

A few well known researchers have statistically demonstrated and proven the possibilities of shifting the Title I and III accountability system to a weighted index model; GILD and these researchers can demonstrate for the review panel exactly what this entails, how and why establishing an accountability system to reflect the shifting and changing dynamic of ELL language and content learning and reporting would greatly benefit ELLs, their teachers, schools, districts and states.

We thank you for your time and the opportunity to address this vitally important issue.

Signed (electronically),

Kathleen Leos, President
Lisa C. Saavedra, Vice-President
The Global Institute for Language & Literacy Development, LLC. [www.gilld.com]

Written Comments Submitted January 18, 2010
Attachment: PPT Presentation Used During Public Meeting

Race to the Top Assessment Program Public & Expert Input Meeting

Expert Input & Suggestions
January 13, 2010

Kathleen Leos
President
The Global Institute for Language & Literacy Development

Legal Framework & A Look to the Future

1974 Supreme Court Case Lau v Nichols:
"Language may not be a barrier to academic content knowledge."

In order to fulfill the requirements set forth in Lau, ELLs must be provided with . . . .

Instructional Opportunities to Acquire the English Language, and Develop English Language in listening, speaking, reading & writing.

Access to Academic Content, and Opportunities to Demonstrate Mastery of Academic Content on Grade Level, based on their level of language proficiency.

The Global Institute for Language & Literacy Development, LLC

2

The Evolution from Lau v. Nichols to NCLB and Beyond

What was and is today:

- Content standards developed independently of language development and acquisition.
- ELP standards often "linked" not "aligned" to content standards.
- Content assessments developed based on content standards, often not valid or reliable measures of ELLs demonstration of mastery in English.
- ELP assessments often developed without consideration of the English language needed to effectively access and demonstrate mastery.

RESULT:

Over reliance on assessment and accountability; less focus on standards, and the realization of validity & reliability issues & concerns. Thus, incongruent use of results; lack of direct impact to positively increase achievement of ELLs.

A New Direction is Needed:

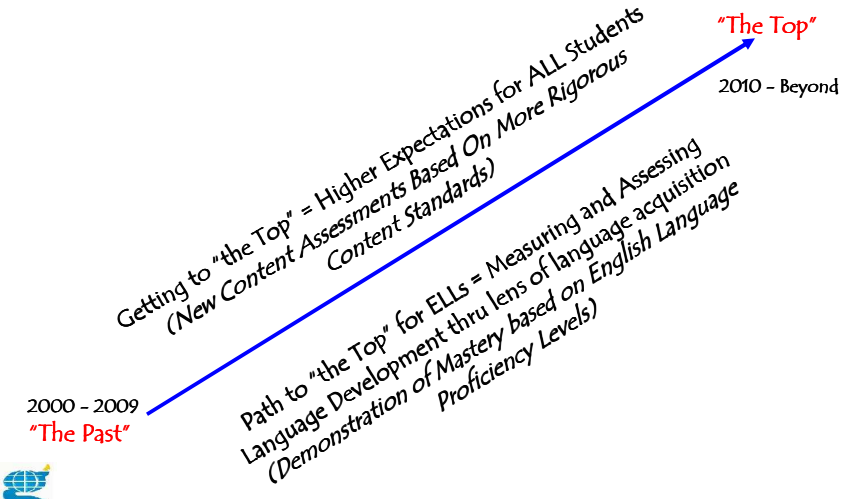
- Focus on standards first, then assessment and accountability.
- If Common Core Content Standards, then Common ELP Standards.
- Alignment of both sets of standards.
- Development of two different assessments based on standards.
- Two assessments linked by revised accountability system from deficit model to transparent growth model.
- Integration of language development and language acquisition.

RESULT:

New generation of standards, assessment and accountability to effectively impact instruction and to increase achievement of ELLs.

The Global Institute for Language & Literacy Development, LLC

"Race to the Top".....Implies U.S.A. Students Effectively Competing in a Global Economy



The Global Institute for Language & Literacy Development, LLC

Consider Importance of "Language" in Assessments

2005 NAEP, Grade 12 Released Item (Reading) SCR:

What was the grandmother seeking in going to London, and did she find it? Support your answer using information from the story?

2007 NAEP, Grade 4 Released Item (Mathematics) MC:

Marty has 6 red pencils, 4 green pencils, and 5 blue pencils. If he picks out one pencil without looking, what is the probability that the pencil he picks will be green?

Essential components of English language development are basis for item development (whether conscious or not).

*In other words, the **structure or form** of English language, the **content** of language and how English language can be **used** can determine the level of difficulty of the item, regardless of the "academic content" item is designed to measure.*

THE CONFLUENCE OF LANGUAGE AND CONTENT ARE NEEDED IN ORDER FOR ELLs TO DEMONSTRATE MASTERY

The Global Institute for Language & Literacy Development, LLC

Valid & Reliable Assessment to Measure ELL Progress

- Valid and reliable content assessments alone cannot effectively measure progress of ELLs.
- Valid and reliable English language proficiency assessments alone cannot effectively measure progress of ELLs.
- Definition of progress should be based on two parallel demonstrations of mastery (assessments):
  - Attainment of proficiency in listening, speaking, reading and writing in English
  - Attainment of grade level content knowledge

The Global Institute for Language & Literacy Development, LLC

It's Not Only About Assessments . . . How Should Two Assessment be Used

- Integrate scores from content and language to measure progress and success.
- Use index model for growth
- Used for accountability purpose
- Allow for the impact of language learning on assessing content knowledge
- Foster the implementation of a transparent model of assessment
- Ensure results may be used for enhancing instruction and thus increasing ELL achievement

The Global Institute for Language & Literacy Development, LLC

GILD's Responses to Some RTTT Assessment Questions

- Provide recommendations for the development and administration of assessments for each content area that are valid and reliable for English language learners
  - *Assessments based on aligned standards (Content & ELP); development of two different assessments; alignment and linkage between two sets of standards and assessment might prove to be a predictor of ELLs success on content assessments.*
- How do you recommend that assessments take into account the variations in English language proficiency of students?
  - *Use of confluence of language and content in test development*
  - *Measuring and assessing language development thru lens of language acquisition levels*

The Global Institute for Language & Literacy Development, LLC

GILD's Responses to Some RTTT Assessment Questions (cont'd)

- How can technology be utilized to make assessments more inclusive to ELL students?
  - *Presentation of assessment items in multiple mediums whereby ELLs may demonstrate mastery of content knowledge based on their level of language proficiency.*
  - *Allowing for use listening/speaking skills v. reading/writing – for items not specifically measuring language domain skill.*
- What are the relative merits of developing and administering content assessments in native language?
  - *Consider the mandates of NCLB and the minimal number of schools in the nation that provide content instruction in the native language in the content areas. (Debra P. v. Turlington: caution about assessing students on content and subject matter they have not been taught.*

The Global Institute for Language & Literacy Development, LLC

THANK YOU!

The Global Institute for Language & Literacy Development, LLC

Visit our website at: www.gilld.com

KATHLEEN LEOS, President
KathleenLeos@gilld.com

LISA C. SAAVEDRA, Vice-President
LisaSaavedra@gilld.com

The Global Institute for Language & Literacy Development, LLC

From: Denis Udall [DUdall@hewlett.org]
Sent: Tuesday, January 19, 2010 8:34 PM
To: Race To The Top Assessment Input
Subject: RTTT Assessment Comments
Attachments: Assessment RFP recommedations.doc; Secretary Duncan letter 1-20-10.docx

Dear Secretary Duncan,

Thank you for the opportunity to submit comments on the competitive Race to the Top Assessment Program to support one or more consortia of States that are working toward jointly developing and implementing common, high-quality assessments.

Attached please find "What a Federal RFP for New Assessments Should Require and Encourage." This document summarizes the recommendations of the Hewlett Foundation, Nellie Mae Education Foundation, Ford Foundation, and the Sandler Foundation regarding the U.S. Department of Education's goal of developing and implementing high quality summative assessments that are aligned with internationally benchmarked standards and that support effective teaching and learning.

We believe that the Department should request that state consortia develop assessment systems that reflect the best practices of high-achieving nations around the world, as well as lessons learned in the United States. The goal should be not only to develop internationally benchmarked Common Core standards, but also internationally benchmarked systems of curriculum, assessment, and instruction by which to implement the standards. These recommendations are offered with that goal in mind.

Respectfully yours,

Denis Udall
Program Officer
William and Flora Hewlett Foundation

What a Federal RFP for New Assessments Should Require and Encourage

Submitted by the Ford Foundation, Hewlett Foundation
Nellie Mae Foundation, and Sandler Foundation

The US Department of Education has initiated an Assessment Program intended, in the Department's words, to support "consortia of States working toward jointly developing and implementing a next generation of common summative assessments that are aligned with a common set of K-12 internationally benchmarked, college and career ready standards that model and support effective teaching and student learning."

The goal of the new assessments is to allow individual students to demonstrate mastery of the knowledge and skills, at each grade level tested, that would enable them to be on track to be college and career ready by the time of high school graduation – as well as to demonstrate growth over time. The RFP is intended to support "innovative and effective" approaches to assessment and to help states create "powerful and useful" systems of assessment.

Finally, the Department intends that information gathered from the assessments should provide information useful for teaching, learning, and program improvement; as well as for understanding student achievement and readiness for college and careers.

To accomplish these goals a number of decisions need to be made about what priorities to establish for the features of the assessments themselves and the system in which they rest, as well as how consortia are formed and what specific tasks consortia should undertake.

If the goal of new assessment systems is to promote strong and equitable learning of the higher order thinking and performance skills needed in contemporary college and career settings, and to improve the capacity of teachers, leaders, and schools to provide powerful educational experiences, it will be critically important that these systems be designed to engender as well as measure deeper and more transferable learning that students can apply to new contexts.

From that perspective, we believe that the Department should request that state consortia develop assessment systems that reflect the best practices of high-achieving nations around the world, as well as lessons learned in the United States. The goal should be not only to develop internationally benchmarked Common Core standards, but also internationally benchmarked systems of curriculum, assessment, and instruction by which to implement the standards. These recommendations are offered with that goal in mind.

Features of Assessment Systems

Developing new, more powerful systems of assessment that support more productive and more equitable learning will require more than simply developing a new test. As high-achieving nations understand, assessment *as, of,* and *for* learning requires a frame of reference that considers how assessment operates to model and provoke useful learning experiences and quality instruction, as well as how it can inform teaching and school improvement. Consequently, the RFP should ask bidders to describe how they will:

- Design assessment systems to support a tightly integrated system of standards, curriculum, formative and summative assessment, instruction, and teacher development. Consortia should be able to articulate how the assessments they define will be integrated with these other, well-aligned elements of state systems, and assessments should be developed with these other aspects of the system in mind. They should also articulate how they will stage and sequence their work to enable assessments to be designed in light of curriculum goals and to appropriately support the necessary changes in instruction and teaching to support teachers' and students' opportunities to learn.

- Consider the role of curriculum focused on higher order thinking and performance skills. Assessments always rest on an implicit conception of what students should learn. Making this conception explicit promotes critically important guidance for assessment design, greater clarity for teaching, greater curriculum equity in the classroom, and greater possibilities for linking formative and summative elements of an assessment system. Consortia states should work together to develop means for standards to be translated into curriculum guidance that is lean, clear, and descriptive of what students should know and be able to *do* as a result of their learning experiences, including applications of knowledge to real-world tasks.

  While specific curricula and materials may vary across states and districts, consortia can support states in developing curriculum frameworks around which they can organize deeper curriculum development at the local level, state and local assessment development, instructional supports, and professional development. Thoughtful curriculum guidance should outline the scaffolding and formative assessment needed to prepare students to succeed on the summative assessments. To the greatest extent possible, the curricula, frameworks, formative assessments and other education resources that are generated by the Assessment Program should be openly licensed, adaptable, and available for reuse and redistribution so that all states and districts can benefit from them.

- Organize assessments (and the curriculum in which they are embedded) around a well-defined set of learning progressions within subject areas that reflect how students build knowledge and skills and can guide teaching decisions as well as evaluation. Systems organized around such progressions can better evaluate student growth in both formative and summative assessments, and can enable teachers to use assessment data to support more effective teaching and learning.

- Ensure that results from the assessment system can be aggregated at the classroom, school, district, and state levels, and support reporting that is timely, accurate, and displayed in ways that are useful for teachers, students, families, schools, districts, states, institutions of higher education, and employers.

- Develop processes for clear, transparent communication with stakeholders and the public about 1) the nature of the new assessment system, 2) how it has changed, 3) what the results mean.

- Take advantage of new technologies to enhance the way assessments are developed, delivered, scored, and used, providing adaptive tools to more accurately measure student achievement and growth, enabling access to information resources for students to

demonstrate their learning, providing appropriate feedback by supporting both teacher scoring and computer-based scoring, and reducing costs for high-quality items / tasks.

- Involve teachers in scoring of constructed responses and performance tasks in order to measure effectively students' mastery of higher-order content and skills and to build teacher expertise and understanding of standards and performance expectations.

- Clarify standards by exemplifying assessment expectations through scoring criteria or rubrics and by samples of student work, which should be made available to educators and others who are to be informed by the assessment system.

- Develop moderation, auditing, or calibration systems that can be used strategically to ensure comparability in scoring and to support teacher and system learning about the implementation of performance standards.

Features of Assessments

In order to encourage valid, appropriate, and instructionally useful evaluation of student learning, the RFP should ask bidders to explain how the assessments they design will:

- Reflect and support good instructional practice by eliciting complex performances as well as applications of knowledge to familiar and new situations, including performance components and constructed response tasks, as well as selected-response items. Bidders should include a rationale for the mix of proposed item and task types across grade levels and subject areas, in terms of learning goals, cost factors, and psychometric considerations.

- Reflect the full range of standards, including those that require student planning, the ability to frame and solve problems, the ability to access and analyze information, and the ability to communicate and defend ideas clearly and accurately.

- Design classroom-based tasks that are part of summative assessment judgments for comparable administration and consistent scoring. Bidders should describe why and how they will strategically use training, moderation, auditing, and/or calibration systems (in-person and via computer) to enhance scoring consistency, teacher learning, and cost-efficiencies.

- Design assessments to more accurately evaluate student growth over time by using well-grounded learning progressions to guide both the construction of "on-demand" test items and more extended classroom assessments, and by exploring the use of computer-based adaptive testing that creates vertically scaled assessments based on the full continuum of learning. Where adaptive testing is proposed, bidders should describe how they propose also to obtain information about the mastery of grade level standards.

- Ensure that assessments will be valid, reliable, and fair, with evidence that they measure well the knowledge and abilities intended, are comparable across contexts and evaluators, and are accessible and unbiased in the language, referents, and approaches they use.

- Incorporate principles of universal design that seek to remove construct-irrelevant aspects of tasks that could increase barriers for non-native English speakers and students with other specific learning needs. In addition, designers who are skilled at developing linguistically supportive assessments and tests for students with learning disabilities should be engaged from the beginning in considering how to develop assessments for maximum access, as well as how to design appropriate accommodations and modifications to enable as many students as possible to be validly assessed within the system.

Features of Consortia

Consortia should have a clear governance structure that defines decision making processes and membership -- including what members have agreed to, what conditions enable or terminate membership, what may define different levels of participation, if applicable.

Consortia of states should organize their work to develop a coherent set of assessments in reading and mathematics from grades 3-8 plus high school. While it has been suggested by some that consortia take on subtopics within these domains (e.g. a few grade levels within a single subject), the likelihood that these attempts would ultimately be integrated into a coherent set of well-sequenced assessments, or that states would be able to implement an integrated system of curriculum and assessments in a timely manner would likely be undermined by this approach.

At the high school level, consortia might choose different approaches – or might support subgroups of states with a given consortium that work together to develop assessments. These could include:

- **Course- or syllabus-based systems** like those in England, Australia, Singapore, Hong Kong, Alberta (Canada), as well as the International Baccalaureate. Generally conceptualized as end-of-course-exams in this country, this approach is a more comprehensive course assessment approach in these other countries. Such an approach would include within-course performance assessments that count toward the examination score, as well as high-quality assessment end-of-course components that feature constructed response as well as selected response items. Within-course performance assessments should tap central modes of inquiry in the disciplines, enabling students to engage in scientific investigations, literary analyses and other genres of writing, speaking and listening; mathematical modeling and applications; social scientific research. Such an approach might require an ELA and math assessment at a key juncture (e.g. grade 10) that evaluates an appropriate benchmark level for high school standards, and then, as in high-achieving nations, allow for pursuit of other courses/ assessments that are selected by students according to their interests and expertise. These could serve as additional information on the diploma for colleges and employers.

- **Standards-driven systems** that might include a more comprehensive benchmark assessment in ELA and mathematics at around grade 10, complemented by collections of evidence that demonstrate students' abilities to meet certain standards within and across the disciplines. This set of assessments would allow more curriculum flexibility in how to meet the standards. Sometimes these sets of evidence are organized into structured portfolios that require specific tasks in each content area, scored with common rubrics and moderation.

- **A mixed model** could combine elements of both course- and standards-driven models, allowing some demonstrations of proficiency to occur in any one of a range of courses (rather than a single, predetermined course) or even outside the bounds of a course, like the efforts by some states to allow students to pass courses via demonstrations of competence rather than seat time. Such a system could also include specific components intended to develop and display research and inquiry skills that might also be interdisciplinary, such as the Project Work requirements in England, Singapore, and the International Baccalaurate, and the Senior Project requirements in Pennsylvania and Ohio.

Any of these approaches (either course based, as part of a certification system, or standards-based, as part of a proficiency system) could be the basis for a set of assessments developed for **Career and Technical Education.** This might be the topic of a separate consortium or of a subgroup within an overarching consortium.

Bidders should be prepared to describe how **states working within the Consortium will:**

- Adopt and augment the Common Core standards as appropriate to their context.
- Create and deploy curriculum frameworks that address the standards—drawing on exemplars and tested curriculum models.
- Build and manage an assessment system that includes both on-demand and classroom-based assessments that evaluate the full range of standards and allow evaluation of student progress. Consortia may develop joint summative assessments (commonly implemented by states) as well as other assessment tasks and items linked to the standards (and grounded in curriculum units) that can be incorporated into states' individual assessment plans for formative or summative purposes.
- Develop rubrics that embody the standards, and clear examples of good work, benchmarked to performance standards.
- Create oversight / moderation / audit systems for ensuring the comparability of locally managed and scored assessment components.
- Ensure that teacher and leader education and development infuse knowledge of learning, curriculum, and assessment.
- Implement high-quality professional learning that includes examination of student work, curriculum and assessment development, and moderated scoring.
- Develop communication plans and materials to enable parents, stakeholders, and the general public to understand the new assessments and their meaning.

Bidders should also describe how **LEAs and schools will:**

- Examine the standards and evaluate current curriculum, assessment, and instructional practice in light of the standards.
- Evaluate state curriculum guidance, and further develop and adapt curriculum to support local student learning, select and augment curriculum materials, and continually evaluate and revise curriculum in light of student learning outcomes.
- Incorporate formative assessments into the curriculum, organized around the standards, curriculum, and learning sequences to inform teaching and student learning.

- Participate in administering and scoring relevant portions of the on-demand and curriculum-embedded components of the assessment system, and examining student work and outcomes.
- Help design and engage in professional development around learning, teaching, curriculum, & assessment.
- Engage in review and moderation processes to examine assessments and student work, within and beyond the school.
- Create processes for explaining the new assessment system and for increasing public understanding of the assessments and their results.

In addition, Consortia should be prepared to develop technology platforms to support the assessment system. Bidders should be prepared to describe how they will develop technology to deliver assessments; to support technology-based tasks; to score results and/or support scoring through the training and calibration of scorers and moderation of scores, as well as to aggregate results in ways that support reporting and research about the responses. Bidders should also be prepared to provide information about how they will use technology and other strategies (economies of scale from multiple states, advanced scoring systems) to create cost savings that render the final system feasible and affordable.

Larry Berger
CEO and Co-Founder
Wireless Generation, Inc.
U.S. Department of Education Public Meeting
Race to the Top Assessment -- Procurement
January 14, 2010
Prepared Remarks

Good morning. My name is Larry Berger and I am the CEO and co-founder of Wireless Generation. Wireless helps more than 200,000 teachers assess 3 million students in all 50 states. Our most recent product uses formative assessment data to provide students with optimized curricula based on individual needs. Wireless Generation also helped to create the country's largest and most comprehensive instructional improvement system for the NYC public schools. Currently, we are working with several researchers to invent a more curriculum-driven approach to summative assessment that we hope will contribute to the Race To The Top Assessment program.

I am likely to disagree with much of what my admittedly more knowledgeable colleagues say today because I believe strongly that there is a vast disconnect between the sorts of procurement schemes likely to be discussed today and the aspirations of the Assessment program.

Robert Mislevy characterized "the test theory that dominates educational measurement today as the application of 20th century statistics to 19th century psychology." The mechanically-scored, multiple-choice bubble sheet dates to 1937, the #2 pencil to 1870. The Race To The Top Assessment Program is our moment to race into the 21st century with an updated understanding of how we learn, with data that matter, and with technologies that are poised to deliver fundamental changes in teaching, learning, and assessing.

This program invites meaningful innovation in numerous dimensions of assessment:

- New and better *items*, *item types*, and *approaches to test construction*;
- Assessment at *different times of year, more often, and more embedded;*
- New *relationships between common standards, curriculum, and assessment;*
- New affordances for *assessing special populations*; and
- New technological platforms.

Innovating on so many fronts at once creates exponential, rather than simply additional, levels of difficulty. For this to be the practical, large scale breakthrough we seek, all of these complex elements will need to be woven together seamlessly into something that teachers are eager to use.

The Race To The Top and the states who will lead it are seeking to buy something that does not yet exist, something that has yet to be *designed*. We may therefore need a new kind of procurement that creates a space in which this kind of design can happen.

While I am not an expert on procurement law, I do spend my days working with educators and education agencies trying to expand the boundaries of educational design. From this experience, I have become convinced of what I call "the 50 hour rule," which asserts that you need a team (in this case a team of teachers, principals, scholars, administrators, designers, engineers, statisticians and user interface experts) to put in at least 50 hours shoulder-to-shoulder at the whiteboard defining the needs, shaping a vision, and prototyping solutions before you really understand the top toward which you could be racing and the path you want to take there. The 50 hour rule is informally confirmed by my observations of the leading innovation efforts in education including the SERP program at the National Academies, the

Design, Educational Engineering, and Development program that Tony Bryk is leading at the Carnegie Foundation for the Advancement of Teaching, and the PEER effort at the Institute for Learning.

The timing of this need for collaborative design runs up against the procurement problem I raised during an earlier public meeting and that hangs over this whole process: that the procurement rules in some states seem to work against doing collaborative design with state agencies. Depending on widely-varying and uncertain state law, the work one might do to generate ideas with a state might preclude one from doing further work should the state obtain federal funding and solicit bids. This is a policy problem that threatens both the success of the federal government's program and state education goals. In some cases it is not just a matter of *when* the 50 hours can happen. It is a matter of *whether* they can happen at all.

There are, as my co-panelists today have discussed, ways to solicit information in a procurement process – and those are certainly helpful. But none of the methods easily fits the intensity of the collaboration between the private sector, the public sector, the research sector, and the technology sector that this moment calls for.

I have only one slide today. It addresses the question at the heart of my testimony: when could the 50 hours of collaborative design take place in the Race To The Top Assessment Program?

The timeline in this slide outlines what I take to be the likely steps in the Race To The Program. I think we have a few options about when this moment of collaborative design could happen. My humble proposal is that it might be enabled by creating a modified kind of

procurement that I'll call an"RFC" – Request for Collaboration. You might think of it as a "design bake-off."

Here's how it could work. A state or states would issue the RFC, and potential partners would respond by submitting sketchbooks or other summaries of their initial design ideas, accompanied by a description of their organizational capacities. On the basis of those submissions, the state or each consortium of states would short-list a small number of likely candidates, and go through the 50-hour collaborative design process with each. The results of those sessions would then help inform the selection of partners to do the implementation work, assuming the resulting proposal wins RTT funding. Participating in multiple fifty hour processes is a lot of time for a busy state, but it is a tiny fraction of the overall time a state invests in its assessment programs. And the time spent will be immensely valuable as a way to expand the thinking that all of us in the room do.

The slide I have prepared illustrates when the RFC might take place. The RFC could happen here, at the beginning of the process. This would make sure that states really understand their options and clarify their needs before joining a consortium. Or it could happen here, after states have defined their needs and affiliated, but in time to shape a breakthrough vision and pick external partners in time for their proposal to the feds. Or it could happen here, after states have defined their vision in some other way. It is this third one that I am dubious about. It worked for earlier consortia when there wasn't as much demand for breakthrough design innovations, but I think this would be too late in the process to achieve the aspirations of the Race To The Top Assessment program.

I think we want this to happen as early in the process as possible.

Doing so offers a way to arrive at clarity, transparency and deep collaboration. We need much greater *clarity* for innovators about how states expect their procurement rules to apply. We need *transparency* so that the formation of consortia is not a result of back room deals but instead takes place in an open marketplace of ideas. And we need a model of procurement that allows time and opportunity for collaboration so that innovation in design can remain at the heart of this race.

I would like to touch briefly upon the important question of how to continue fostering innovation after the initial design phase of the Assessment competition closes. Continued innovation is particularly challenging in light of the fact that neither the federal government nor a State is likely to have the resources to fund it fully. Nor should government be required to carry this burden alone. Rather, the Assessment program should create sufficient incentives for the business community to partner with government in making healthy, ongoing investments in innovation.

Although the Department need not prescribe a particular intellectual property scheme or business model to foster sustained innovation, it should reward creative solutions to this challenge. And it should do so with the knowledge that different models of intellectual property ownership will likely result in different consequences.

For example, a consortium might take a fully "open-source" approach by putting all assessment items and developed technology under an open source license. Although this approach would allow parties other than the original developer to use and build upon the consortia's work, it could limit the amount of investment any one developer is willing to put into the project.

Another consortium might “open source” only content – perhaps assessment items or lessons – while permitting the developer to maintain rights to the technology platform. This approach could incentivize companies to invest and innovate now and later, as they look forward to returns at a later date.

Yet another consortium might build a fully “closed” platform of assessment items and technology, while providing an open “innovation zone” for Application Programming Interfaces (APIs) or similar opportunities for third parties to extend and enhance the platform.

There are other models, and each of them has benefits and disadvantages. The bottom line should be that the government’s program both inspires the most creative and innovative solutions and encourages non-government entities to invest in creating ever more innovative approaches to the education challenges our country faces.

Thank you.

* * *

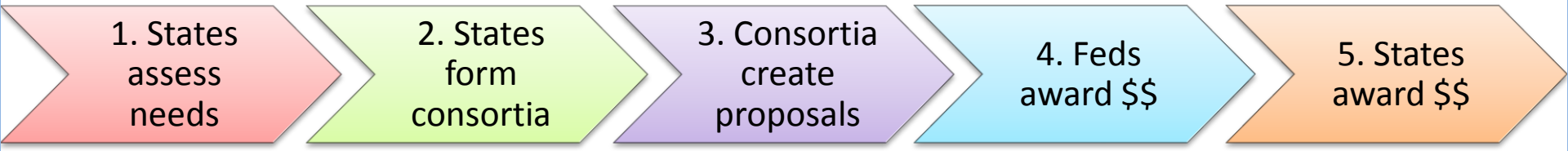
wireless generation®

Larry Berger
CEO, Wireless Generation

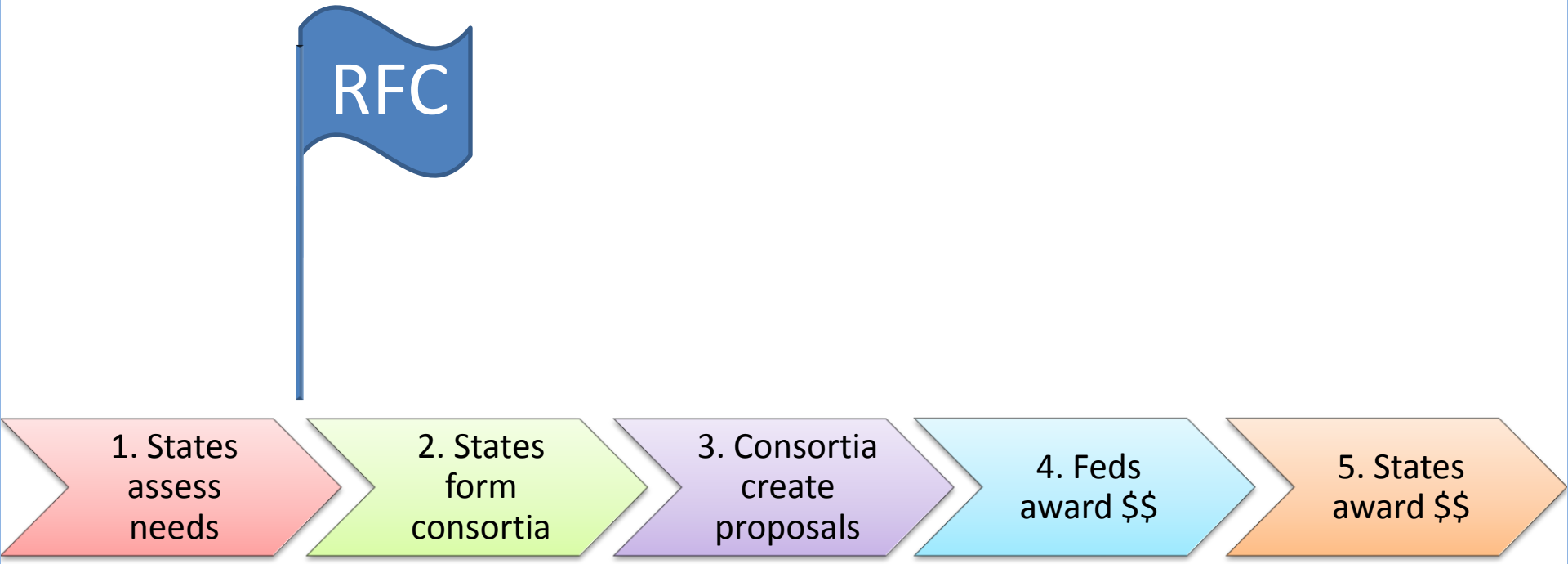
U.S. Department of Education Public Meeting
January 14, 2010

© Wireless Generation, Inc. 2010

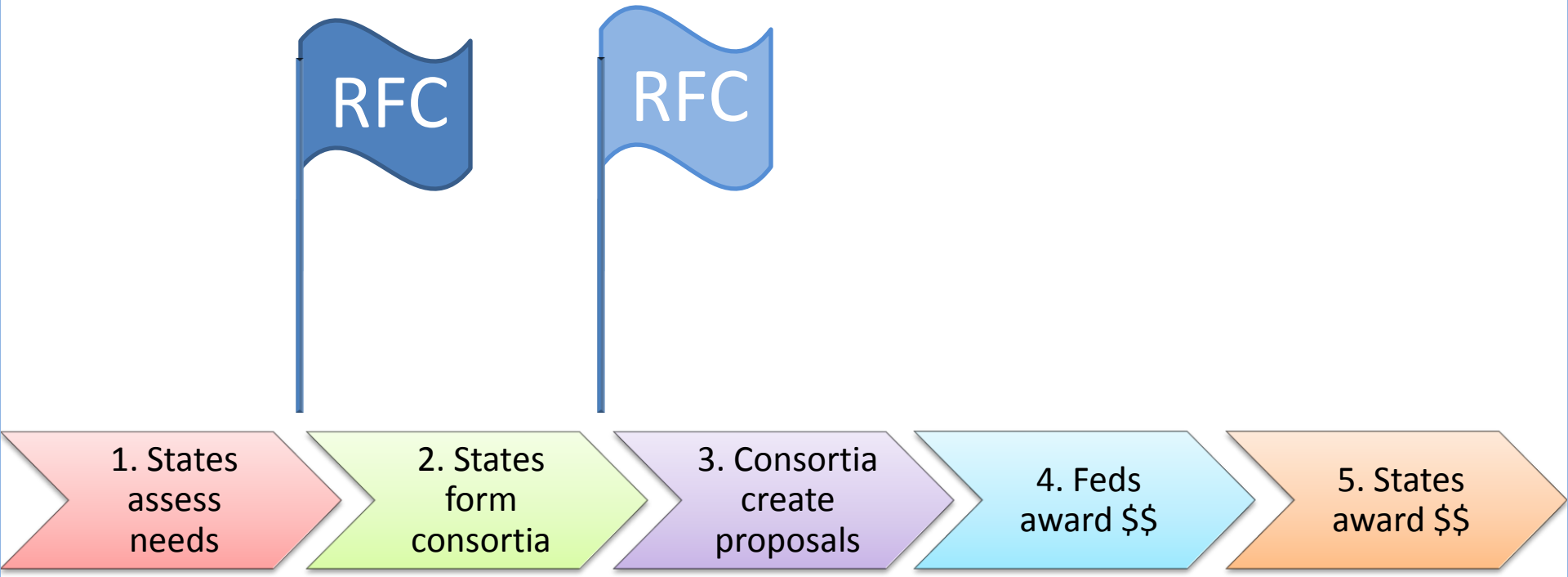
When should the 50 hours happen?



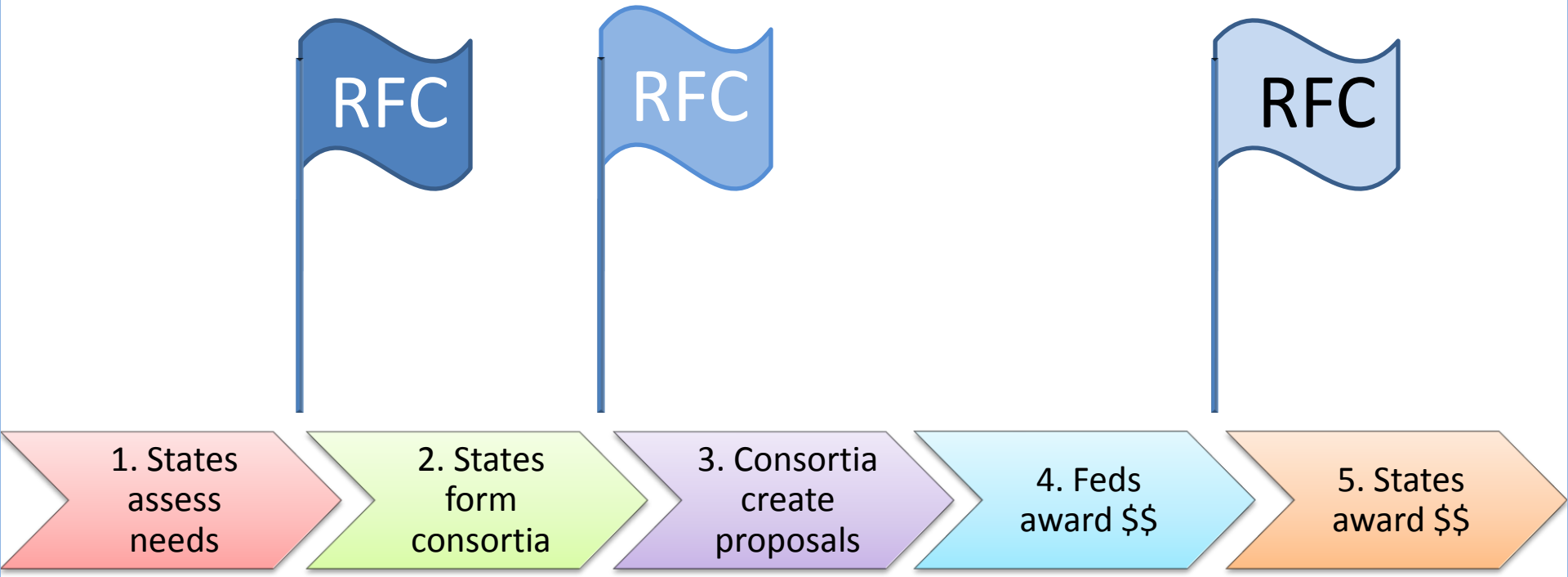
When should the 50 hours happen?



When should the 50 hours happen?



When should the 50 hours happen?



More Potential for Innovation

Less Potential for Innovation

Race to the Top
Assessment Input Meeting
January 20, 2010

Linda Darling-Hammond

Good afternoon. My name is Linda Darling-Hammond. I am a professor at Stanford University and co-director of the Stanford Center for Opportunity Policy in Education and the School Redesign Network. I have conducted extensive research on assessment systems in the United States and abroad, and have been involved in the development of student and teacher assessment programs in a number of states, including CA, NY, and Ohio. I am currently leading a project, sponsored by the Ford Foundation and the Nellie Mae Educational Foundation, which synthesizes what we have learned in the U.S. and other nations about how to design, score, and manage performance assessment systems that are valid, reliable, feasible, and affordable.[1] This project has sponsored a set of 7 papers that will be released at the end of March, 2010.

Along with my comments today, I will be submitting three papers[2] from that set of studies that describe 1) the operations of performance assessment systems in high-achieving countries, 2) technical advances in performance assessments in the U.S. over the last 20 years that allow us to create more valid, reliable, and affordable assessment tools; and 3) cost estimates that illustrate how to make rich, performance-based assessments affordable in an integrated system that supports teaching and learning. In this brief time, I want to make three key points.

[1] That study is guided by a technical advisory panel of assessment experts, education leaders, and policy analysts, including: Stanford Professors Dr. Richard Shavelson (chair) and Dr. Edward Haertel; NAS Education Director, Dr. Michael Feuer; University of Colorado Professor and former NAE president Lorrie Shepard; UCLA Professor and head of CRESST, Dr. Eva Baker; former Commissioner of Rhode Island, Peter McWalters; former Commissioner of New Hampshire and current Nellie Mae president, Nick Donohue; policy analysts Jack Jennings and Christopher Cross; Council of Chief State School Officers executive director, Gene Wilhoit; and National Association of State Boards of Education executive director, Brenda Wellburn.

[2] Linda Darling-Hammond, Benchmarking Learning Systems:Student Performance Assessment in International Context (in press); Suzanne Lane, Performance Assessment: The State of the Art (in press); The Assessment Solutions Group, The Cost of New Higher Quality Assessments: A Comprehensive Analysis of the Potential Costs for Future State Assessments (in press).

First: Creating a system that routinely incorporates performance-based components in both the formative and summative aspects of our assessment systems is essential to improving the quality of teaching and learning in our schools. Students will not learn to become strong writers and communicators, researchers and investigators, problem framers and solvers, thinkers and producers of ideas and products if they do not have opportunities and expectations to engage in these activities on a regular basis. These kinds of activities are a substantial and growing component of the assessment systems in high-achieving nations like Finland, the Netherlands, Hong Kong, Singapore, Australia, New Zealand, and the U.K, among others.

I coined the term "through-course assessments" at the Boston assessment panel meeting in describing what these nations do, largely to make the point that their assessments are typically not only "end of course" tests, but a set of assessments that include – in addition to a sit down test at the end of the year -- curriculum-embedded performance tasks which count as part of the summative judgment. These include science labs and investigations, research papers, a variety of written tasks across genres (e.g. persuasive essays, responses to literature, creative writing), applied mathematics tasks, and so on. These elements that I called "through-course" assessments are <u>not</u> mini-tests that are used to predict success on the final test (as are many of the narrow interim or benchmark tests that are currently in use in the U.S.) They are tasks that represent the central modes of inquiry in a discipline; that are valid and authentic measures of how knowledge and skills will be used outside of school; and that are designed to ensure that students have rich learning opportunities, teachers have good models and a rich evidence base to inform their planning and teaching, and the full range of standards are measured.

If the phrase College and Career-Ready skills means anything, it must include these abilities that are the core expectations of universities and employers in the knowledge industries

that are growing most rapidly. Teaching and learning focused on multiple-choice and short answer tests that emphasize recall and recognition rather than analysis, synthesis, and production will not help students develop these skills, and we have substantial evidence that these are the very skills our students are furthest behind in developing relative to students in high-achieving nations. Furthermore, we have evidence that the use of performance-based assessments, grounded in a curriculum that requires writing, analysis, investigation, and problem solving, change teaching and learning in ways that help students master these abilities. This research is summarized in a paper I authored that will be submitted with my comments.

Second: Creating such a system is much more feasible that it once was. Since the early 1990s, there have been substantial advances in our understanding of how to design tasks and scoring systems so that they measure what is intended, and are generalizable, and scorable in comparable ways. We have seen large advances in technology that allow both for more efficient delivery of performance tasks to students and teachers and for more consistent, effective, and affordable scoring by both teachers and by computers. These advances are described in a paper by Suzanne Lane that I am submitting along with my comments today.

Third: Such a system – which will net great benefits for instruction that cannot be secured in other ways – is both cost-effective and can be made affordable through the economies of scale offered by state consortia, new uses of technologies, and thoughtful approaches to teacher scoring. The cost implications of various approaches to building a system that includes significant performance components across a consortium of states are outlined in a paper that will be further discussed by John Olson this afternoon. This cost modeling study uses carefully researched, real-world cost data to illustrate how it is possible to construct a formative and

summative assessment system that features performance-based items and tasks for less than what most states are spending on systems with much less instructional benefit today.

Among the several consortia that are currently forming, there are three that are working together to conceptualize and implement a balanced assessment system with performance-based components for both formative and summative assessment, and with shared instructional supports, that can take advantage of the affordances of technology. There is great energy in the field to move learning forward in significant ways. The time is ripe for major breakthroughs in assessment in this country if the US DOE reaches for a truly higher standard.

Benchmarking Learning Systems:

Student Performance Assessment in International Context

Linda Darling-Hammond
With the Assistance of Laura Wentworth

Stanford University

Abstract: High-performing nations integrate curriculum, instruction, and assessment to improve both teaching and learning. As a large and increasing part of their examination systems, they use open-ended performance tasks and school-based assessments to give students opportunities to develop 21st century skills: the abilities to find and organize information to solve problems, frame and conduct investigations, analyze and synthesize data, and apply learning to new situations. This paper illustrates how several nations integrate these assessments into the curriculum to create stronger learning for both students and teachers, resulting in higher and more equitable achievement.

Since the release of *A Nation at Risk* in 1983, the United States has launched a set of wide-ranging reforms with the intention of better preparing all children for the higher educational demands of life and work in the 21st century. All 50 states have developed standards for learning and tests to evaluate student progress. No Child Left Behind reinforced the use of test-based accountability to raise achievement, and scores have climbed on state tests used for accountability purposes, yet the U.S. has fallen further behind on international assessments of student learning since the law was passed in 2001.

On the Program in International Student Assessment (PISA) tests in 2006, the U.S. ranked 25th of 30 OECD countries in mathematics and 21st of 30 in science, a decline in both raw scores and rankings from 3 years earlier. (Reading scores were not reported, because of editing problems with the U.S. test.) Furthermore, in each disciplinary area tested, U.S. students scored lowest on the problem-solving items. The U.S. also had a much wider achievement gap than the most highly-ranked jurisdictions,

such as Finland, Canada, Australia, New Zealand, Hong Kong, Korea, and Japan. Singapore, which has not yet participated in PISA, places at the top of the international rankings on the International Education Assessments (TIMSS and PIRLS) where these other nations also excel, and shows a very narrow distribution of achievement, despite its multiethnic, multilingual, socioeconomically diverse population.

It is worth noting that PISA assessments focus explicitly on 21st century skills, going beyond the question posed by most U.S. standardized tests, "Did students learn what we taught them?" to ask, "What can students do with what they have learned?" (Stage, 2005). PISA defines literacy in mathematics, science, and reading as students' abilities to *apply* what they know to new problems and situations. This kind of higher-order learning is increasingly emphasized in other nations' assessment systems, but often discouraged by the multiple-choice tests most U.S. states have adopted.

Policy discussions in Washington often refer to these rankings when emphasizing the need to create more "internationally competitive" standards by benchmarking expectations in the U.S. to those in high-performing nations. Typically, this means looking at topics that are taught at various grade levels in various countries. These analyses reveal that higher-achieving countries teach fewer topics more deeply each year; focus applications of knowledge, rather than recall of facts; and have a more thoughtful sequence of expectations based on developmental learning progressions within and across domains (Schmidt, Wang, & McKnight, 2005; Valverde & Schmidt, 2000).

It is important to examine *how* these topics are taught and assessed — so that we understand how other countries' education systems shape what students actually learn and can do. European and Asian nations that have steeply improved student learning have

focused explicitly on creating curriculum guidance and assessments that focus on the so-called 21st-century skills: the abilities to find and organize information to solve problems, frame and conduct investigations, analyze and synthesize data, apply learning to new situations, self-monitor and improve one's own learning and performance, communicate well in multiple forms, work in teams, and learn independently.

Curriculum differences are reinforced by sharp divergence between the forms of testing used in the U.S. and those used in higher-achieving countries. Whereas U.S. tests rely primarily on multiple-choice items that evaluate recall and recognition of discrete facts, most high-achieving countries rely largely on open-ended items that require students to analyze, apply knowledge, and write extensively. Furthermore, these nations' growing emphasis on project-based, inquiry-oriented learning has led to an increasing prominence for school-based tasks, which include research projects, science investigations, development of products, and reports or presentations about these efforts. These assessments, which are incorporated into the overall examination scoring system, influence the day-to-day work of teaching and learning, focusing it on the development of higher-order skills and use of knowledge to solve problems.

Smaller countries often have a system of national standards that are sometimes — though not always — accompanied by national tests in the upper grades. Top-ranking Finland uses local assessments almost exclusively to evaluate its national standards and manages a voluntary national assessment for college admissions at only one grade level. Larger nations — like Canada, Australia, and China — have state- or provincial-level standards, and their assessment systems are typically a blend of state and local assessments. Managing assessment at the state rather than national level, where it remains

relatively close to the schools, turns out to be an important way of enabling strong teacher participation and ensuring high-quality local assessments that can be moderated to ensure consistency in scoring.

In many cases, local assessments complement centralized “on-demand” tests, constituting up to 60% of the final examination score. Tasks are mapped to the standards or syllabus for the subject and are selected because they represent critical skills, topics, and concepts. They are often outlined in the curriculum guide, but they are generally designed, administered, and scored locally, based on common specifications and evaluation criteria. Whether locally or centrally developed, decisions about when to undertake these tasks are made at the classroom level, so they are used when appropriate for students’ learning process and teachers can get information and provide feedback as needed, something that traditional standardized tests cannot do. In addition, as teachers use and evaluate these tasks, they become more knowledgeable about both the standards and how to teach to them and about their students’ learning needs. Thus, the process improves the quality of teaching and learning.

Like the behind-the-wheel test given for all new drivers, these performance assessments evaluate what students can actually *do*, not just what they know. The road test not only reveals some important things about drivers’ skills; preparation for the test also helps *improve* those skills as novice drivers practice to get better. In the same way, performance assessments set a standard toward which everyone must work. The task and the standards are not secret, so teachers and students know what skills they need to develop and how they will need to be demonstrated.

Finally, these countries typically do not use their examination systems to punish schools or to deny diplomas to students. Following the problems that resulted from the Thatcher government's use of test-based school rankings, which caused a narrowing of the curriculum and widespread exclusions of students from school (Rustique-Forrester, 2005), several countries enacted legislation precluding the use of test results for school rankings. High school examinations provide information for higher education, vocational training, and employment, and students often choose areas in which they will be examined, as a means of demonstrating their qualifications. Because the systems are focused on using information for users of the system and for curriculum improvement, rather than sanctions, governments can set higher standards and work with schools to achieve them, rather than devising tests and setting cut scores at a minimal level to avoid dysfunctional side-effects.

In this paper, we examine the assessment systems of several high-achieving education systems: two Scandinavian nations — Finland and Sweden — plus a group of English-speaking jurisdictions that have some shared approaches to assessment, as well as some interesting variations — Australia, Hong Kong, Singapore, and the United Kingdom. In addition, we describe the International Baccalaureate system, which is now used in many U.S. schools as well as schools around the world. (A summary of assessment system features is included in Table 1 below.) These examples provide interesting lessons about how assessments can be linked to curriculum and integrated into the instructional process to shape and improve learning for students and teachers alike.

Table 1: Examples of International Assessment Systems

Country/ State	Description of Core System	What kinds of assessments are used?	Who designs and grades the assessments?
FINLAND	Student performance is evaluated on a sample basis by the Finnish education authorities at the end of 2nd and 9th grades to inform curriculum and school investments. All other assessments are designed and managed locally, based on the national curriculum.	*National* – Problems and written tasks that ask students to apply their thinking *School-based* – Research tasks, Presentations, & demonstrations	*National* – Designed by teachers through the Finnish Ministry of Education. Graded by teachers. *School-based* -- Teachers design and grade tasks based on recommended assessment criteria and benchmarks for each subject and grade within the national core curriculum.
	A voluntary matriculation examination is taken by most students to provide information to colleges. Students choose which subjects they will sit for (usually at least four), with the test in the students' mother tongue being compulsory.	The tests use mostly open-ended questions to evaluate skills including problem solving, analysis, and writing	The exam is administered, organized and evaluated by The Matriculation Exam Board appointed by the Finnish Ministry of Education. Teachers grade the matriculation exams locally by using the official guidelines and samples of the grades are re-examined by professional raters hired by the Exam Board.
SWEDEN	Students take faculty-designed, nationally approved examinations in year 9 and in the last two years of upper secondary school in Swedish, Swedish as a Second Language, English, and Mathematics. Teachers use these assessments as one factor in determining students' grades at year 9, along with course grades and local assessments. In some cases, local regulations require schools to give an examination in year 5 in these same subjects. All other assessments are designed and managed locally. Teachers weight information from classroom work and assessments they design to determine whether students met the objectives of the national syllabus.	*National* -- Open-ended tasks requiring analysis of materials or problems, and written responses; materials may be given in advance of the test. *School-based* -- Coursework, research projects, diagnostic tasks, essays, problem sets	*National* -- University faculty, with secondary teachers, design the tasks for the national exams at year 9 and the Upper Secondary level. Teachers grade the assessments during time set aside by regional authorities to calibrate grading practices to minimize variation across the region. They incorporate these scores into their grades for their courses. *School-based* -- Teachers design, administer, and grade tasks based on the national curriculum and syllabi, which outline "objectives to achieve" in each subject. Diagnostic materials for assessing syllabus goals are made available on an optional basis.
ENGLAND	National curriculum assessments are enacted primarily as guidance for school-based formative and progress assessments conducted by teachers. A mandatory set of assessments at ages 7 and 11 includes externally developed tasks and observation scales implemented by teachers. Teachers choose which tasks and tests to use and when to use them, within certain parameters. Assessments for primary school are designed and managed locally, based on the national curriculum and guidance provided through the Assessing Pupil Progress (APP) program.	*National* -- Observation scales completed by teachers regarding pupils' work and performance on specific kinds of tasks; written, oral, and performance tasks & tests *School-based* -- Coursework, tests, projects, essays	*National* -- The Qualifications and Curriculum Authority (QCA) manages develops the national assessments, which are implemented scored by teachers, and a range of guidance and supports for in-school assessment. *School-based* -- Teachers evaluate student performance and work samples based on the national curriculum and syllabi. Extensive guidance for documenting pupil performance and progress, with indicators showing relationships to national standards are provided through the Assessing Pupils' Progress project. Regional authorities support teacher training for assessment and in-school moderation.

	Most students voluntarily take a set of exams at year 11 (age 16) to achieve their General Certificate of Secondary Education (GCSE). If they take advanced courses, they may later take A-level exams, which provide information to universities. Students choose the exams they will take based on their interests and areas of expertise. About 75% of the exam grade is based on externally developed tests & 25% is school-based.	*National* -- Essays and open-ended problem solutions, oral language assessments *School-based* - Coursework, tests, projects	*National* -- External exams are designed and graded by examining groups serving different schools (e.g. Oxford Cambridge, Ed Excel, the Assessments and Qualifications Alliance). *School-based* -- Teachers develop and score school-based components based on the syllabus.
SINGAPORE	External examinations are given at the end of primary school (grade 6) in mathematics, science, English, and mother tongue (Malay, Chinese, or Tamil). Results are used to guide course placements in secondary school. All other assessments are school-based.	*National* --Short and long open-ended responses *School-based* – Coursework, research projects, investigations	*National* -- The Singapore Education Assessment Board designs the assessments and manages the assessment system. *School-based* – Designed and graded by the classroom teacher in response to the syllabus.
	After four years of secondary school, students take the GCE N- or O-level examinations. Students choose the elective subject areas in which they want to be examined. Exams have school-based components that comprise up to 20% of the final score. Results are used as information for postsecondary education. GCE A-level examinations may be taken after two years of tertiary education.	*National* -- Short and long open-ended responses and multiple-choice items *School-based* -- Research projects, laboratory investigations and experiments	*National* -- The Singapore Education Assessment Board manages the assessment system. The GCE examinations are developed by the Cambridge International Examinations Group. *School-based* -- Teachers develop, implement, and score projects and other products that complement the external examinations.
HONG KONG	The Hong Kong Territory-wide System Assessments occur at the equivalent of Grades 3, 6, and 9 in Chinese, English, and mathematics. The test is matrix-sampled, and results are reported to schools, but not publicly. Results are not reported for individual students. The goal is to inform curriculum planning within schools and to enable the government to assist schools that are struggling. An on-line bank of tasks is also available for teachers to use for diagnostic assessment of individual students.	*Territory-Wide* -- TSA items are written and oral open-ended items and tasks. *School-based* -- Essays, research projects, investigations	*Territory-Wide* -- Assessments are developed by the Hong Kong Education Examinations Authority and are scored *School-based* -- Assessments are developed and scored by teachers. The Education Bureau encourages schools to develop multiple forms of assessment including projects, portfolios, and observations as well as tests, and looks for the variety of assessments in the performance indicators used for school evaluation.
	The Hong Kong Certificate of Education Examinations are taken at the end of secondary school to provide information to universities and employers. Students choose the areas they will sit for, beyond Chinese, English, mathematics, and liberal studies. These exams and the Diploma of Secondary Education that will replace them in 2012 include school-based assessments, comprising from 10% to 50% of the examination score.	*Territory-Wide* -- Written, speaking, and listening tasks *School-based* -- Projects, portfolios, lab experiments, reading activities, oral tasks	*Territory-Wide* -- Assessments are developed by teachers and higher education faculty, and scored by teachers who are trained as assessors. Tests are allocated randomly to scorers, and essay responses are typically rated by two independent scorers. *School-Based* -- School-based assessments are designed, administered, and scored by teachers in response to syllabus guidelines. Results are statistically moderated to ensure comparability within the province. The assessments are internationally benchmarked, through the evaluation of sample student papers, to peg the results to those in other countries.

AUSTRALIA	At the national level, a literacy and numeracy assessment is given at grades 3, 5, 7, and 9. States and localities manage their own assessment systems.	*National* -- Multiple-choice, short-answer, and extended written responses	*National* -- Designed, administered, and scored by the Curriculum Corporation with questions and prompts contributed by state education agencies.
QUEENSLAND, AUSTRALIA	All additional assessments are school-based, developed by teachers based on the national curriculum guidelines and state syllabi On an optional basis, schools may draw on a bank of "Rich Tasks" from the New Basics project that can be administered across grade levels and scored at the local level, with moderation.	*School-based* -- Open-ended papers, projects, and inquiries -- Rich tasks are complex, interdisciplinary tasks requiring research, writing, and the development of multi-faceted products	*School-based* -- Assessments are developed, administered and scored by teachers. Scoring is moderated by regional panels of teachers and professors that examine scored portfolios of student work representing each score point from each grade level from each school. A state panel also looks at specimens across schools as well. Based on these moderation processes, schools are given instructions to adjust grades for comparability. -- Rich-tasks are developed by teachers with assessment developers; they are accompanied by scoring rubrics and moderation processes by which the quality of student work and scoring can be evaluated.
VICTORIA, AUSTRALIA	All additional assessments are school-based until 11th and 12th grades, when students choose to take exams in different subject areas as part of the Victorian Certificate of Education (VCE), used to provide information to universities and employers. The VCE exams have both external and school-based components. At least 50% of the total examination score is comprised of required classroom-based assignments and assessments given throughout the school year.	*State VCE* -- Multiple-choice (25%) and open-ended (75%) written, oral, and performance elements *School-based* -- Lab experiments, essay, research papers and presentations	The Victoria Curriculum and Assessment Authority (VCAA) establishes courses in a wide range of studies, oversees the development of the external examinations by teachers and university faculty, and ensures the quality of the school-assessed component of the VCE. Teachers score the open-ended items on the external exam and design and score the classroom-based assessments in response to syllabus guidelines. The quality of the tasks assigned by teachers, the work done by students, and the appropriateness of the grades and feedback given to students are audited through an inspection system, and schools are given feedback on all of these elements. In addition, the VCAA uses statistical moderation based on the external exam scores to ensure that the same assessment standards are applied to students across schools, adjusting the level and spread of each school's assessments to match that on the common exam.
INTERNATIONAL BACCALAUREATE	The International Baccalaureate (IB) Diploma Program, a program for students in grades 11 and 12 that is used worldwide, assesses students using school-based assessments throughout the two-year program and externally-developed exams at the end of the two-year program. School-based assessments comprise 20 to 50% of the examination score for each subject.	*External (IB-developed)* -- Essays, open-ended problem solutions, short answer and multiple-choice items *School-based* – Speeches, projects, portfolio, presentations, investigations, labs, artistic performances	*External* -- Designed, administered, and graded by trained IB examiners (usually current or former teachers). *School-based* – Designed and graded by the classroom teacher based on a common syllabus and scoring criteria.

Finland

Finland has been a poster child for school improvement since it rapidly climbed to the top of the international rankings after it emerged from the Soviet Union's shadow. It now ranks first among all the OECD nations on the PISA assessments in mathematics, science, and reading. Leaders in Finland attribute these gains to their intensive investments in teacher education and major overhaul of the curriculum and assessment system. Prospective teachers are competitively selected from the pool of college graduates and receive a three-year graduate-level teacher preparation program, entirely free of charge and with a living stipend. Their master's degree program includes both extensive coursework on how to teach – with a dual focus on inquiry-oriented teaching and teaching that meets the needs of diverse learners – plus at least a full year of clinical experience in a school associated with the university. Preparation includes a strong focus on how to use formative performance assessments in the service of student learning (Laukkanen, 2008; Buchberger & Buchberger, date).

Policy makers decided that if they invested in very skillful teachers, they could allow local schools more autonomy to decide what and how to teach — a reaction against the highly centralized system they sought to overhaul. Finland's national core curriculum is a much leaner document, reduced from hundreds of pages of highly specific prescriptions to descriptions of a small number of skills and core concepts each year. (For example, about 10 pages describe the full set of math standards for all grades.) This guides teachers in collectively developing local curricula and assessments that encourage students to be active learners who can find, analyze, and use information to solve problems in novel situations.

There are no external standardized tests used to rank students or schools. Although it may seem counterintuitive to Americans accustomed to external testing as a means of accountability, Finland's leaders point to its use of school-based, student-centered, open-ended tasks embedded in the curriculum as an important reason for the nation's extraordinary success on international exams (Lavonen, 2008; Finnish National Board of Education, 2007). School-level samples of student performance are evaluated periodically by the Finnish education authorities, generally at the end of the 2nd and 9th grades, to inform curriculum and school investments. All other assessments are designed and managed locally. The national core curriculum provides teachers with recommended assessment criteria for specific grades in each subject and in the overall final assessment of student progress each year (Finnish National Board of Education, June 2008). Local schools and teacher then use those guidelines to craft a more detailed curriculum and set of learning outcomes at each school as well as approaches to assessing benchmarks in the curriculum (Finnish National Board of Education, June 2008). Teachers are treated as "pedagogical experts" who have extensive decision-making authority in the areas of curriculum and assessment as in other areas of school policy and management (Finnish National Board of Education, April 2008).

According to the Finnish National Board of Education (June 2008), the main purpose of assessing students is to guide and encourage students' own reflection and self-assessment. Consequently, on-going feedback from the teacher is very important. Teachers give students formative and summative reports both through verbal feedback and on a numerical scale based on students' level of performance in relation to the objectives of the curriculum. All Finnish schools use a grading scale of 4–10, where 5 is

"adequate" and 10 is "excellent." The recommended assessment criteria are shaped around the grade of 8 or "good." Teachers' reports must be based on multiple forms of assessment, not only exams. Schools are responsible for giving basic education certificates for completing the different milestones of comprehensive school up to 9th grade and additional classes prior to university (European Commission, 2007/2008).

Finland uses assessments to cultivate students' active learning skills by asking open-ended questions and helping students address these problems. In a Finnish classroom, teachers rarely stand at the front of a classroom lecturing students for 50 minutes. Instead, students are generally engaged in independent or group projects, often choosing tasks to work on and setting their own targets with teachers, who serve as coaches (Korpela, 2004). The cultivation of independence and active learning encourages students to develop problem solving and metacognitive skills.

Most Finnish students take a set of voluntary matriculation examinations that provide information for university admissions based on students' abilities to apply problem solving, analytic and writing skills. University and high school faculty members construct the examinations – which are composed of open-ended essays and problem solutions – under the guidance of the Matriculation Exam Board, which is appointed by the Finnish Ministry of Education to organize, manage, and administer the exam (The Finnish Matriculation Examination, 2008). The Board members (about forty in number) are faculty and curriculum experts in the subject areas tested, nominated by universities and the National Board of Education. More than 300 associate members – also typically high school and college faculty – help develop and review the tests. High school teachers

grade the matriculation exams locally using official guidelines, and samples of the grades are re-examined by professional raters hired by the Board (Kaftandjieva & Takala, 2002).

Students take at least four exams, with the test in the students' mother tongue (Finnish, Swedish, or Saami) being compulsory. These tests have a textual skills section, that evaluates students' analytic skills and linguistic expression, and an essay that focuses on the development of thinking, linguistic expression and coherency. They then choose three other tests from among the following: the test in the second national language, a foreign language test, the mathematics test, and one or more tests from the general studies battery of tests in the sciences and humanities (e.g. religion, ethics, philosophy, psychology, history, social studies, physics, chemistry, biology, geography, and health education). The tests also incorporate questions which cross disciplinary boundaries.

It is interesting to note that this system assumes that all students aiming for college (who comprise a majority of will be at least bi-lingual and many will be tri-lingual. The language tests evaluate listening and reading comprehension as well as writing in the language in question.

In addition to choices of which tests to take, students make choices of which items to answer within the exams. In the general battery, they are generally given a set of questions or prompts from which they must respond to six or eight of their choice. On the mathematics test, there are 15 or so problems from which they must choose 10 to answer. Problems require critical thinking and modeling, as well as straightforward problem-solving. For example, the Basic mathematics exam poses this problem:

> A solution of salt and water contains 25 per cent salt. Diluted solutions are obtained by adding water. How much water must be added to one kilogram of the original solution in order to obtain a 10 per cent solution? Work out a graphic representation which gives the amount of water to be added in order to get a

> solution with 2–25 % of salt. The amount of water (in kilograms) to be added to one kilogram of the original solution must be on the horizontal axis; the salt content of the new solution as a percentage must be on the vertical axis.

And the Advanced mathematics exam poses this one:

> In a society the growth of the standard of living is inversely proportional to the standard of living already gained, i.e. the higher the standard of living is, the less willingness there is to raise it further. Form a differential-equation-based model describing the standard of living and solve it. Does the standard of living rise forever? Is the rate of change increasing or decreasing? Does the standard of living approach some constant level?

Assessment is used in Finland to cultivate students' active learning skills by posing complex problems and helping students address these problems. For example, in a Finnish classroom, it is rare to see a teacher standing at the front of a classroom lecturing students for fifty minutes. Instead, in Finnish schools, students are likely to determine their own weekly targets with their teachers in specific subject areas and choose the tasks they will work on at their own pace. A description of a Finnish school (Korpela, 2004) illustrates how students are likely to be walking around, rotating through workshops or gathering information, asking questions of their teacher, and working with other students in small groups. They may be focusing on completing independent or group projects or writing articles for their own magazine. The cultivation of independence and active learning allows students to focus on broad knowledge with emphasis on skills like analytical thinking, problem solving, and metacognitive skills that develop students' thinking. These types of skills, cultivated in Finnish classrooms, are increasingly emphasized on tests like the PISA, which seek to evaluate students' capacities to think independently and creatively in applying knowledge (Lavonen, 2008).

Sweden

Over the past 40 years, Sweden's national assessment system has, like Finland's, shifted from a centralized system based on one test to a more localized system based on multiple forms of assessments. With this change, Sweden hoped to increase Upper Secondary school enrollment and provide more open-access to higher education. Around 1970, Sweden abolished its *studentexamen*, a nationally administered exit exam that ranked Upper Secondary students and placed them in higher education programs (European Commission, 2006-2007). The new policy intended to produce more valid and reliable evidence of whether students would succeed at the university level. Additionally, the country wanted to correct social and educational inequities caused by a one-size-fits-all assessment system (Eckstein and Noah, 1993, p. 230).

Sweden's national curriculum is adjusted for the local context. In "Compulsory School," ages 7 to 16, the common curriculum, which includes nationally approved syllabi for individual subjects, is adapted in each district to address local conditions (Swedish National Agency for Education, 2005). In "Upper Secondary School," ages 17 to 20, Sweden has 17 subject matter curricula consisting of 3-year programs providing a general education and eligibility to study at the post-secondary or university level. The programs include eight core subjects (English, the arts, physical education and health, mathematics, general science, social studies, Swedish or Swedish as a second language, and religion) and a set of subject specific area like "the construction program" or "the business program" that combine general courses with specialized classes. The National Agency for Education determines the required courses for a national specialization and most of the programs require at least 15 weeks of workplace training outside of school.

Sweden pairs its nationally outlined and locally implemented curriculum with multiple layers of assessment controlled by schools and teachers. Assessments in compulsory school consist of several components. First, during each school term, the teacher, student, and parent meet to discuss the student's learning and social development (Swedish National Agency for Education, 2005). Second, students receive grades from their teachers in each term of year 8 (age 15) and the end of the fall term of year 9 (age 16). Teachers base their Year 9 grades on the goals in the syllabi, awarding grades of "pass", "pass with distinction," or "pass with special distinction" based on nationally approved assessment criteria (Swedish National Agency for Education, 2005).

Third, schools can use a number of optional diagnostic materials to assess students in Swedish, Swedish as a Second Language, Mathematics, and English. The diagnostic materials help teachers assess students and support their learning. The diagnostic materials in years 6 through 9 assess where students stand in relation to the goals set by the syllabi (Swedish National Agency for Education, 2005).

Fourth, students take nationally approved examinations in year 9. The exams assess the subjects of Swedish, Swedish as a Second Language, English, and Mathematics. Teachers use these assessments as one factor in determining students' grades. The exam at year 9 is compulsory for schools, but not for students. Sweden uses the scores from the test to ensure the grades given by teachers compare to the national standards (Qualifications and Curriculum Authority, 2008). In some jurisdictions, schools give an examination at the end of year 5 in these same subjects.

Towards the end of their Upper Secondary schooling, Swedish students receive a grade in each course and a final grade or "learning certificate" that acts as a compilation

of all grades awarded for courses and projects completed by the student. Teachers keep extensive records of student progress, using three assessments to aid in their grading at the Upper Secondary school level: 1) coursework, 2) assessments designed by teachers based on the course syllabi, and 3) nationally approved examinations when grading the core subjects of Swedish, English and Mathematics, and selected other areas (Swedish National Agency for Education, 2005). Regional education officials and schools provide time for teachers to calibrate their grading practices to minimize variation across the schools and across the region (Eckstein and Noah, 1993, p. 230).

While university personnel produce the national exams at year 9 and at the Upper Secondary level, teachers help design the tasks and questions and grade the assessments (Eckstein and Noah, 1993; O'Donnell, 2004). Teachers also design school-based assessments based on the nationally determined syllabi, which dictate the content of the coursework students complete in their classes (Qualifications and Curriculum Authority, 2008a; O'Donnell, 2004, p. 23; Eckstein and Noah, 1993, p. 83-84).

The National School Board examinations administered during students' compulsory and Upper Secondary schooling use an open-ended, authentic approach to assessing students. The exam questions are grounded in real world contexts, asking students to use analytic skills and draw on content knowledge learned during their classes. For example, Sweden's native language test at the upper elementary school level asks students about a broad theme. One year, the exam used the theme of "travel" and provided students with a contemporary poem, prose and poetry extracts from a variety of authors, a practical description of how to plan a trip, and data about travel presented in a set of texts, charts, and statistical tables. Schools gave students materials a week in

advance, so students had time to review the materials. Students had five hours to write an essay on the topic of their choice that was evaluated on specific criteria emphasized in the syllabus from their course. The skills assessed included using appropriate language in certain circumstances, comprehending the different purposes of language, persuasive mechanisms, presenting information, as well as creative self expression, word choice and grammar (Eckstein and Noah, 1993, p. 119).

These examples from math assessments illustrate how questions are embedded in real-world contexts. Here is an on-demand sample question from the grade 5 exam that asks students (aged 11-12) to grapple with a problem that they might have in their own lives, both weighing and balancing decisions as well as applying math knowledge:

> Carl bikes home from school at four o'clock. It takes about a quarter of an hour. In the evening he's going back to school because the class is having a party. The party starts at 6 o'clock. Before the class party starts, Carl has to eat dinner. When he comes home, his grandmother calls, who is also his neighbor. She wants him to bring in her post before he bikes over to the class party. She also wants him to take her dog for a walk, then to come in and have a chat. What does Carl have time to do before the party begins? Write and describe below how you have reasoned (Pettersson, 2008).

The mathematics exam from the Upper Secondary level also frames the questions in real world, tangible topics and formats. Students have almost 4 hours to answer 15 questions. The first 10 questions require short answers and the last 5 questions require longer answers for which students show their work.

Swedish Mathematics Exam at the Third Year of Upper Secondary
Short Format Questions
– A coffee blender mixed *x* kg of coffee costing *a* knoner/kg with *y* kg of coffee costing *b* kroner/kg. Give a formula for the price per kg of the blend.
– In 1976 Lena had a monthly salary of 6,000 kr. By 1984 her salary had risen to 9,000 kr. In current prices, her salary had risen by 50%. How large was the percent change in fixed prices? In 1976 the Consumer Price Index (CPI) was 382; in 1984 it was 818.
Long Format Questions
– A business paid into a pension fund at the beginning of every year a sum of 15,000 kr.

> The fund has a yearly growth rate of 10%. The first payment was made in 1987 and the last will be in 2010. The pension fund will continue to grow until 2015. How much more will the business have in the fund at the beginning of 2015, if it pays in the same amount as above, but the rate of growth is 15%?
> (Eckstein & Noah, 1993, pp. 270-272).

Australia

Whereas smaller countries like Finland and Sweden have national curriculum guidance, in much-larger Australia, each state has its own curriculum and assessment program. In most states, local school-based performance assessment is a well-developed part of the system. In some cases, states have also developed centralized assessment with performance components. Two of the highest-achieving states, Queensland and A.C.T., have highly developed systems of local performance assessment. Victoria, which uses a blended model of centralized and school-based assessment, also generally performs well on national and international tests. An effort to develop common national curriculum standards is underway, but states will continue to be responsible for translating these standards into an assessment program at the state and local levels.

Queensland. In Queensland, there has been no assessment system external to schools for 40 years. Until the early 1970s, a traditional "post-colonial" examination system controlled the curriculum. When it was eliminated, all assessments became school-based. School-based assessments are developed, administered and scored by teachers in relation to the national curriculum guidelines and state syllabi (also developed by teachers), and are moderated by panels that include teachers from other schools as well as at least one professor from the tertiary education system.

To create the standards used throughout the province, the central authority gathers groups of teachers and subject experts to write standards that specify different levels of

achievement and describe the characteristics of student work at each level. In the excerpt from Queensland's science standards shown below, the left side describes the objectives or "Essential Learnings" that must be taught and assessed by teachers. The objectives convey the knowledge or skill expected at each standard. The standard descriptors to the right detail the expected characteristics and quality of the work. The teachers and experts also develop samples of work used as exemplars to show the different levels. These standards guide the assessments teachers develop and their scoring.

Excerpt from Queensland Science Standards

What must be taught and assessed

Expected quality of the work

Investigative processes	The Standards descriptors	
Students need to ... investigate scientifically ... judge the worth of quantitative and qualitative data ... interpret ... apply the outcomes. Students manipulate and review data and scientific techniques ... improve their scientific knowledge. They synthesise research that they have generated to discuss the outcomes in relation to their initial purpose. By the end of the course, students should be able to:	**Standard A** The student work has the following characteristics:	**Standard C** The student work has the following characteristics:
• **conduct and appraise Physics research tasks** — this includes the abilities to formulate questions, hypothesise, plan, manage, evaluate, refine and justify decisions made during investigations, as well as the critical reflection required to fulfil research goals	• formulation of **justified significant** questions/ hypotheses which inform **effective** and **efficient** design, refinement and management of investigations	• formulation of questions and hypotheses to select and manage investigations
• **operate scientific equipment and technology safely** — this includes the abilities to safely select, adapt and apply technologies, laboratory and fieldwork equipment, and consider its limitations; it also incorporates the ability to do this individually and in groups	• assessment of risk, **safe** selection and adaptation of equipment, and appropriate application of technology to gather, record and **process valid** data	• assessment of risk, safe selection of equipment, and appropriate application of technology to gather and record data
• **use primary and secondary data** — this includes the abilities to analyse and extrapolate from data, and to identify relationships, patterns and anomalies in primary and secondary data.	• **systematic** analysis of primary and secondary data to **identify relationships** between patterns, trends, errors and anomalies.	• analysis of primary and secondary data to identify obvious patterns, trends, errors and anomalies.

The syllabi seek to strike a balance between "informed prescription" and "informed professionalism." They spell out a small number of key concepts and/or skills to be learned in each course, and what kinds of projects or activities (including minimum assessment requirements) students should be engaged in. Each school designs its

program to fit the needs and experiences of its own students, choosing specific texts and topics with this in mind. However, all schools must use shared, specific criteria for evaluation of student work, based on the course objectives and specific standards for an A, B, C, D, and E.

As the criteria from the physics syllabus (below) indicate in substantial detail, in the category of *Knowledge and conceptual understanding*, work that meets an "A" standard demonstrates interpretation, comparison, and explanation of complex concepts, theories and principles, whereas work at an "E" standard is characterized by reproduction of isolated facts and application of simple, given algorithms. In this particular course, objectives also include *Investigative Processes*, and *Evaluating and Concluding*,, with indicators spelled out for all of these objectives. The expectations of work quality are challenging, as shown in the example below.

Extended Experimental Investigations

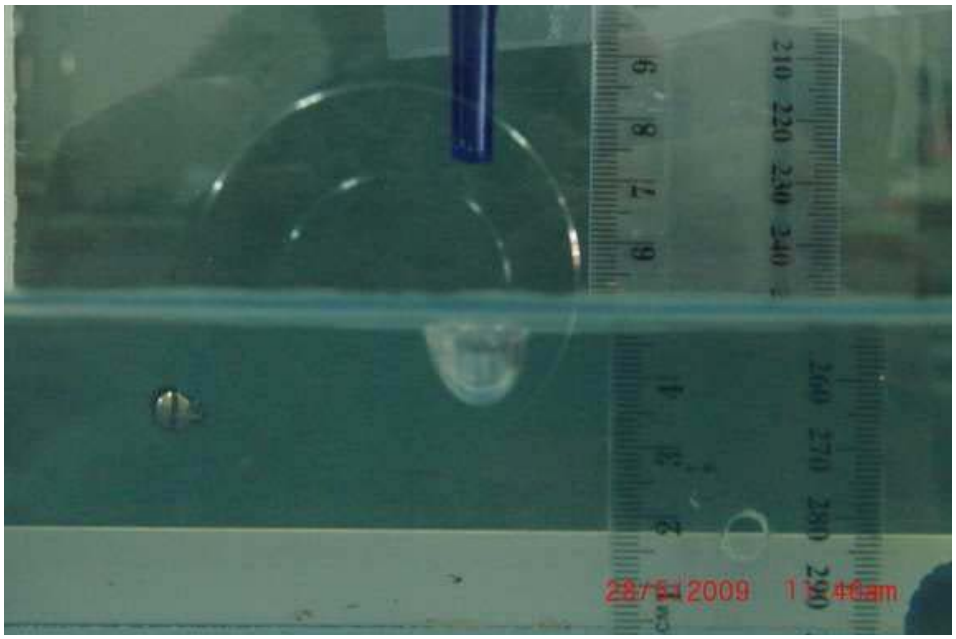
In Queensland science courses, students must complete an extended experimental investigation. The instructions for the task read:

Within this category, instruments are developed to investigate a hypothesis or to answer a practical research question. The focus is on planning the extended experimental investigation, problem solving and analysis of primary data generated through experimentation by the student. Experiments may be laboratory or field based. An extended experimental investigation may last from four weeks to the entirety of the unit of work. The outcome of an extended experimental investigation is a written scientific report. *Aspects of each of the three criteria should be evident in the investigation.* For monitoring, the discussion/conclusions/evaluation/recommendations of the report should be between 1500 and 2000 words.

To complete such an investigation the student must:
- develop a planned course of action
- clearly articulate the hypothesis or research question, providing a statement of purpose for the investigation
- provide descriptions of the experiment
- show evidence of modification or student design

· provide evidence of primary and secondary data collection and selection
· execute the experiment(s)
· analyze data
· discuss the outcomes of the experiment
· evaluate and justify conclusion(s)
· present relevant information in a scientific report.

An example from a year 12 paper shows how a student investigated a problem entitled, “The Air Pocket.” The assessment starts with a picture of a vertical air jet from a straw producing a cavity on a water surface.



The student investigated the parameters that would affect the volume of the cavity, preparing a 32 page paper meeting the criteria described earlier, including evaluating the problem theoretically and empirically, presenting data through tables and charts, analyzing findings both by summarizing individual results and developing a regression to evaluate the combined effects of several variables on the volume of the cavity, and evaluating the results, along with the potential errors and additional research needed.

Overall, the paper more closely resembles a research report from a scientific laboratory than a traditional high school physics test. The student concluded:

> It was determined through initial theoretical research that the predominant influences on the cavity's volume were air speed, diameter of nozzle/straw and distance between straw/nozzle and water. Upon testing the effects of changing an individual parameter with respect to volume, every possible variation was tried, such that eventually a complete set of values was obtained. To combine the different parameters into a single equation, a multiple regression was used; to determine both the constant factor and the powers to which each of the variables should be raised. The resultant r2 value was 0.96 indicating an excellent fit for the data while the average percentage error was 1.59% and the median percentage error, 6.71%. … [In future experiments], it would be suggested to do the experiments on a larger scale as this would virtually eliminate the effects of surface tension while cutting down unfounded accuracy in the model (the volume could be measured in cubic centimetres or cubic metres, resulting in a more realistic fit, with data that is not required to be impossibly precise. Finally, it would be suggested to trial the effects of the different orientation of the straw/nozzle, as tilting it would give a completely differently shaped cavity (due to the dispersion characteristics of air).

Thus, students go beyond their own empirical data and conclusions to reflect on the accuracy of their findings and means for improving their investigation. These kinds of extended responses are demanded in all of the subject areas, shaped by the core concepts and modes of inquiry of the disciplines. Student reflection is also a common element of the assessments. Consistent scoring of such intellectually ambitious work is made possible in part by internal and external moderation processes (described below), and in part by the clear guidance of the syllabi and rubrics used to set standards for the work.

Physics

Syllabuses are developed to show a clear alignment between the General Objectives (shaded) that state what students should be able to achieve as a result of completing the course and the standards associated with Exit Criteria. Schools must directly teach and assess the General Objectives.

General objectives in the syllabus	**Standards**				
	Standard A	**Standard B**	**Standard C**	**Standard D**	**Standard E**
Knowledge and conceptual understanding Students should acquire knowledge and construct understanding of facts, theories, concepts and principles of physics. To work scientifically, students need to have an understanding of underlying scientific knowledges, including the associated mathematical skills. They need to engage with the processes and phenomena observed in Physics through characteristics of data analysed. Students need to make informed judgments based on sound reasoning in order to direct them in their scientific endeavours and to engage with problem solving. By the end of the course, students should be able to: ▪ **recall and interpret concepts, theories and principles of Physics** — this includes the abilities to remember, reproduce and interpret subject matter such as facts, definitions, formulas, terminology, concepts, theories, principles, laws, procedures, sequences, events, diagrams, symbols, figures, systems and patterns ▪ **describe and explain processes and phenomena of Physics** — this includes the abilities to compare and classify the concepts, theories and principles being explored, based on primary and secondary data ▪ **link and apply algorithms, concepts, theories and schema of Physics** — this includes the abilities to adapt, translate and reconstruct understandings in order to find solutions.	The student work has the following characteristics: • reproduction and interpretation of complex and challenging concepts, theories and principles • comparison and explanation of complex concepts, processes and phenomena • linking and application of algorithms, concepts, principles, theories and schema to find solutions in complex and challenging situations.	The student work has the following characteristics: • reproduction and interpretation of complex or challenging concepts, theories and principles • comparison and explanation of concepts, processes and phenomena • linking and application of algorithms, concepts, principles, theories and schema to find solutions in complex or challenging situations.	The student work has the following characteristics: • reproduction of concepts, theories and principles • explanation of simple processes and phenomena • application of algorithms, principles, theories and schema to find solutions in simple situations.	The student work has the following characteristics: • reproduction of simple ideas and concepts • description of simple processes and phenomena • application of algorithms, principles, theories and schema.	The student work has the following characteristics: • reproduction of isolated facts • recognition of isolated simple phenomena • application of simple given algorithms.

General objectives in the syllabus	Standards				
	Standard A	Standard B	Standard C	Standard D	Standard E
Investigative processes Students need to recognise the methodologies available to them to investigate scientifically. They need to be able to judge the worth of quantitative and qualitative data and interpret and apply the outcomes of such data. Students require the skills to manipulate and review data and scientific techniques so that they may improve their scientific knowledge. They need to synthesise the research that they have generated and be able to discuss the outcomes in relation to their initial purpose. By the end of the course, students should be able to: ▪ **conduct and appraise Physics research tasks** — this includes the abilities to formulate questions, hypothesise, plan, manage, evaluate, refine and justify decisions made during investigations, as well as the critical reflection required to fulfil research goals ▪ **operate scientific equipment and technology safely** — this includes the abilities to safely select, adapt and apply technological, laboratory and fieldwork equipment, and consider its limitations; it also incorporates the ability to do this individually and in groups ▪ **use primary and secondary data** — this includes the abilities to analyse and extrapolate from data, and to identify relationships, patterns and anomalies in primary and secondary data.	The student work has the following characteristics: • formulation of justified significant questions/hypotheses which inform effective and efficient design, refinement and management of investigations • assessment of risk, safe selection and adaptation of equipment, and appropriate application of technology to gather, record and process valid data • systematic analysis of primary and secondary data to identify relationships between patterns, trends, errors and anomalies.	The student work has the following characteristics: • formulation of justified questions/hypotheses which inform design and management of investigations • assessment of risk, safe selection of equipment, and appropriate application of technology to gather, record and process data • analysis of primary and secondary data to identify patterns, trends, errors and anomalies.	The student work has the following characteristics: • formulation of questions and hypotheses to select and manage investigations • assessment of risk, safe selection of equipment, and appropriate application of technology to gather and record data • analysis of primary and secondary data to identify obvious patterns, trends, errors and anomalies.	The student work has the following characteristics: • implementation of given investigations • safe use of equipment and technology to gather and record data • identification of obvious patterns and errors.	The student work has the following characteristics: • guided use of given procedures • safe directed use of equipment to gather data • recording of data.

General objectives in the syllabus	Standards				
	Standard A	**Standard B**	**Standard C**	**Standard D**	**Standard E**
Evaluating and concluding Students who are working scientifically need to be able to make decisions about the knowledge they have gained and generated. They need to distinguish between a plausible conclusion and one based on pure supposition. Students need to be able to synthesise their thoughts and the thinking of others into a coherent whole, from which they can make judgments and propose future possibilities. They need to reach conclusions and explain the world in which they live, using science. They need to be able to adhere to communication and scientific conventions in communicating their decisions to selected audiences. By the end of the course, students should be able to: ▪ **determine, analyse and evaluate the interrelationships involved in applications of Physics** — this includes the abilities to identify the physics involved, to determine the simple and complex relationships that exist between concepts, principles, theories and schema and then to critically examine the associated implications ▪ **predict outcomes and justify conclusions and recommendations** — this includes the abilities to explore scenarios and consider possible outcomes, and then to provide justifications of conclusions and recommendations ▪ **communicate information in a variety of ways** — this includes the abilities to select, use and present data and ideas to convey meaning, an argument or a case to selected audiences in a range of formats.	The student work has the following characteristics: • analysis and evaluation of complex scientific interrelationships • exploration of scenarios and possible outcomes with justification of conclusions/ recommendations • discriminating selection, use and presentation of scientific data and ideas to make meaning accessible to intended audiences through innovative use of range of formats.	The student work has the following characteristics: • analysis of complex scientific interrelationships • explanation of scenarios and possible outcomes with discussion of conclusions/ recommendations • selection, use and presentation of scientific data and ideas to make meaning accessible to intended audiences in range of formats.	The student work has the following characteristics: • description of scientific interrelationships • description of scenarios and possible outcomes with statements of conclusion/ recommendation • selection, use and presentation of scientific data and ideas to make meaning accessible in range of formats.	The student work has the following characteristics: • identification of simple scientific interrelationships • identification of scenarios or possible outcomes • presentation of scientific data or ideas in range of formats.	The student work has the following characteristics: • identification of obvious scientific interrelationships • statements about outcomes • presentation of scientific data or ideas.

At the end of the year, teachers collect a portfolio of each student's work, which includes the specific assessment tasks, and grade it on a 5-point grading scale. To calibrate these grades, teachers put together a selection of portfolios from each grade level – one from each of the 5 score levels plus borderline cases -- and send these to a regional panel for moderation. The panel of five teachers re-scores the portfolios and confers about whether the grade is warranted, making a judgment on the spread. State review panels also look at a sample of student work from each district to insure that schools implement the standards across all districts. Based on this analysis, and at year 12, a standardized state-wide test called the Queensland Core Skill (QCS) Test, the Queensland authority confirms the levels of achievement proposed by school programs and may adjust it if it does not calibrate to the standards.

Aiming for even more applied, interdisciplinary work, Queensland developed a "Rich Tasks" approach to standards and assessment, which was introduced as a pilot in 2003. Part of the "New Basics" project, this effort has created extended, multi-disciplinary tasks that are developed centrally and used locally when teachers determine the time is right and they can be integrated with locally-oriented curriculum (Queensland Government, 2001). These are "specific activities that students undertake that have real-world value and use, and through which students are able to display their grasp and use of important ideas and skills." Rich Tasks are defined as:

> A culminating performance or demonstration or product that is purposeful and models a life role. It presents substantive, real problems to solve and engages learners in forms of pragmatic social action that have real value in the world. The problems require identification, analysis and resolution, and require students to analyze, theorize and engage intellectually with the world. As well as having this connectedness to the world beyond the classroom, the tasks are also rich in their application: they represent an educational outcome of demonstrable and substantial intellectual and educational value. And, to be truly rich, a task must

be transdisciplinary. Transdisciplinary learnings draw upon practices and skills across disciplines while retaining the integrity of each individual discipline.

One task description is summarized below.

Science and Ethics Confer

Students must identify, explore and make judgments on a biotechnological process to which there are ethical dimensions. Students identify scientific techniques used as well as significant recent contributions to the field. They will also research frameworks of ethical principles for coming to terms with an identified ethical issue or question. Using this information they prepare pre-conference materials for an international conference that will feature selected speakers who are leading lights in their respective fields.

In order to do this students must choose and explore an area of biotechnology where there are ethical issues under consideration and undertake laboratory activities that help them understand some of the laboratory practices. This enables them to:

A) Provide a written explanation of the fundamental technological differences in some of the techniques used, or of potential use, in this area (included in the pre-conference package for delegates who are not necessarily experts in this area).
B) Consider the range of ethical issues raised in regard to this area's purposes and actions, and scientific techniques and principles and present a deep analysis of an ethical issue about which there is a debate in terms of an ethical framework.
C) Select six real-life people who have made relevant contributions to this area and write a 150-200 word précis about each one indicating his/her contribution, as well as a letter of invitation to one of them.

This assessment measures research and analytic skills; laboratory practices; understanding biological and chemical structures and systems, nomenclature and notations; organizing, arranging, sifting through, and making sense of ideas; communicating using formal correspondence; précis writing with a purpose; understanding ethical issues and principles; time management, and much more.

A bank of these tasks now exists across grade levels, along with scoring rubrics, and moderation processes by which the quality of the tasks, the student work, and the scoring can be evaluated. Extensively researched, this system has had excellent success as a tool for school improvement. Studies found stronger student engagement in learning

in schools using the Rich Tasks. On traditional tests, these "New Basics" students scored about the same as students in the traditional program, but they performed notably better on assessments designed to gauge higher order thinking.

The Singapore government has employed the developers of the Queensland system to focus the new school improvement strategies upon performance assessments. High-scoring Hong Kong has also begun a process of expanding its already-ambitious school-based assessment system in collaboration with Queensland assessment developers.

Victoria. In Victoria, a mixed system of centralized and decentralized assessment combines these kinds of school-based assessment practices with a set of state exams guided by the Victoria Essential Learning Standards. The AIM program, used at years 3, 5, 7, and 9, provides an indication of how well the literacy and numeracy skills of students are developing. The results provide information used to plan new programs and a useful source of feedback and guidance to students, parents and teachers. Assessment tasks include extended open-ended writing responses, as well as some multiple-choice responses.

At the secondary level, the Victorian Certificate of Education (VCE) provides information that guides pathways to further study at the university, Technical and Further Education (TAFE) and to the world of work. Some students undertake a school-based apprenticeship or traineeship within the VCE. The Victoria Curriculum and Assessment Authority establishes courses in a wide range of studies, develops the external examinations and ensures the quality of the school-assessed component of the VCE.

VCAA conceptualizes assessment as “of,” “for,” and “as” learning. Teachers are involved in developing assessments, along with university faculty in the subject area, and all prior year assessments are public, in an attempt to make the standards and means of measuring them as transparent as possible. Before the external examinations are given to students, teachers and academics sit and take the exams themselves, as if they were students. The external subject-specific examinations, given in grades 11 and 12, include about 25% machine-scored items; the remaining items are open-ended, and are scored by the classroom teacher. The exams may include written, oral, and performance elements.

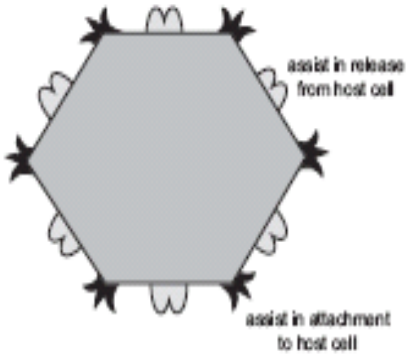
In addition, at least 50% of the total examination score is comprised of classroom-based tasks that are given throughout the school year. These required assignments and assessments – lab experiments and investigations on central topics as well as research papers and presentations – are designed by teachers in response to syllabus expectations. These required classroom tasks ensure that students are getting the kind of learning opportunities which prepare them for the assessments they will later take, that they are getting feedback they need to improve, and that they will be prepared to succeed not only on these very challenging tests but in college and in life, where they will have to apply knowledge in these ways.

An example from the Victoria biology test, for example, describes a particular virus to students, asks them to design a drug to kill the virus and, in several pages, explain how the drug operates (complete with diagrams), and then to design an experiment to test the drug.

Victoria, Australia High School Biology Exam

When scientists design drugs against infectious agents, the term “designed drug” is often used.
A. Explain what is meant by this term. _____
_____
_____

Scientists aim to develop a drug against a particular virus that infects humans. The virus has a protein coat and different parts of the coat play different roles in the infective cycle. Some sites assist in the attachment of the virus to a host cell; others are important in the release from a host cell. The structure is represented in the following diagram:



The virus reproduces by attaching itself to the surface of a host cell and injecting its DNA into the host cell. The viral DNA then uses the components of host cell to reproduce its parts and hundreds of new viruses bud off from the host cell. Ultimately the host cell dies.

B. Design a drug that will be effective against this virus. In your answer outline the important aspects you would need to consider. Outline how your drug would prevent continuation of the cycle of reproduction of the virus particle. Use diagrams in your answer. Space for diagrams is provided on the next page.

C. Before a drug is used on humans, it is usually tested on animals. In this case, the virus under investigation also infects mice. Design an experiment, using mice, to test the effectiveness of the drug you have designed.

In preparation for this on-demand test, students taking Biology will have been assessed on six pieces of work during the school year covering specific outcomes in the syllabus. For example, they will have conducted “practical tasks” like using a

microscope to study plant and animal cells by preparing slides of cells, staining them, and comparing them in a variety of ways, resulting in a written product with visual elements. They also will have conducted practical tasks on enzymes and membranes, and on the maintenance of stable internal environments for animals and plants. Finally, they will have completed and presented a research report on characteristics of pathogenic organisms and mechanisms by which organisms can defend against disease. These tasks link directly to the expectations that students will encounter on the external examination but go well beyond what that examination can measure in terms of how students can apply their knowledge. The syllabus instructs teachers as follows:

The tasks are graded according to criteria set out in the syllabus. The quality of the tasks assigned by teachers, the work done by students, and the appropriateness of the grades and feedback given to students are audited through an inspection system, and schools are given feedback on all of these elements. In addition, the VCAA uses statistical moderation to ensure that the same assessment standards are applied to students across schools. The external exams are used as the basis for this moderation, which adjusts the level and spread of each school's assessments of its students to match the level and spread of the same students' collective scores on the common external test score. The result is a rich curriculum for students with extensive teacher participation and a comparable means for examining student learning.

Hong Kong

In collaboration with educators from Australia, the UK, and other nations, Hong Kong's assessment system is evolving from a highly centralized examination system to one that increasingly emphasizes school-based, formative assessments that expect

students to analyze issues and solve problems. The government has decided gradually to replace the Hong Kong Certificate of Education Examinations, which most students sit for at the end of their 5-year secondary education, with a new Hong Kong Diploma of Secondary Education that will place even more emphasis on school-based assessments. In addition, the Hong Kong Territory-wide System Assessment (TSA), which provides assesses lower-grade student performance, in Chinese, English, and mathematics, is developing an on-line bank of assessment tasks to enable schools to assess their students and receive feedback on their performance on their own timeframes. The formal TSA assessments, which include both written and oral components, occur at Primary Grades 3 and 6 and Secondary Grade 3 (the equivalent of grade 9 in the U.S.).

As outlined in Hong Kong's "Learning to learn" reform plan, the goal of the reforms is to shape curriculum and instruction around critical thinking, problem-solving, self-management skills, and collaboration. A particular concern is to develop meta-cognitive thinking skills, so students may identify their strengths and areas needing additional work (Education Bureau, September 2001, Chan, et al., 2008).

School-based assessments have been part of Hong Kong's examination system since 1978, but will increase in importance. The Hong Kong Education Examinations Authority explains the rationale for growing use of school-based assessments (SBA):

> The primary rationale for SBA is to enhance the validity of the assessment, by including the assessment of outcomes that cannot be readily assessed within the context of a one-off public examination. SBA can also reduce dependence on the result of public examinations, which may not always provide the most reliable indication of the actual abilities of candidates. Obtaining assessments based on student performance over an extended period of time and developed by those who know the students best - their subject teachers - provides a more reliable assessment of each student. Another reason for including SBA is to promote a positive impact on teaching and learning. It can serve to motive students by engaging them in meaningful activities; and for teachers, it can reinforce

> curriculum aims and good teaching practices, and provide structure and significance to an activity that they are in any case involved in on a daily basis, namely assessing their own students….
>
> Teachers know that SBA, which typically involves students in activities such as making oral presentations, developing a portfolio of work, undertaking fieldwork, carrying out an investigation, doing practical laboratory work or completing a design project, help students to acquire important skills , knowledge and work habits that cannot readily be assessed or promoted through paper-and-pencil testing. Not only are they outcomes that are essential to learning within the disciplines, they are also outcomes that are valued by tertiary institutions and by employers. Moreover, they are activities that students find meaningful and enjoyable. (HKEAA, 2009).

By 2007, Curriculum and Assessment Guides were published for four core subjects and 20 elective subjects, and assessments in the first two subjects – Chinese Language and English Language – were revised. These became criterion-referenced, performance-based assessments featuring not only the kinds of essays previously used on the exams, but also new speaking and listening components, the composition of written papers testing integrated skills, and a school-based component generally worth 20 to 30% of the examination score. Although the existing assessments already use open-ended responses (see the example of a Physics examination question in Appendix A), the proportion of such responses will increase in the revised assessments.

These new assessments feature "standards-referenced reporting," in which examinations are scored on a scale from 1 to 5 pegged to specific criteria, rather than the norm-referenced approach that produced earlier grades. The Hong Kong Examinations and Assessment Authority explains that, "By providing both students and teachers with a specific set of level descriptors, SRR not only clarifies the standards expected at various levels of attainment but also helps set targets for improvements in learning and teaching" (HKEAA, 2007).

Like the existing assessments, the new assessments are developed by teachers with the participation of higher education faculty, and they are scored by teachers who are trained as assessors. Tests are allocated randomly to scorers, and essay responses are typically rated by two independent scorers (Dowling, n.d.). Results of the new school-based assessments are statistically moderated to ensure comparability within the province. The assessments are internationally benchmarked, through the evaluation of sample student papers, to peg the results to those in other countries. Many of the new assessments are also to be scored on-line, which the Examinations Authority notes is now the common practice in 20 of China's mainland provinces, as well as in the UK.

The Education Bureau encourages schools to develop assessments focused on learning, and is shifting its education policies to underscore this focus. For example, the Bureau promotes the use of multiple forms of assessment in schools including projects, portfolios, observations, and examinations, and looks for the variety of assessments in the performance indicators used for school evaluation (Chan, et al., 2008; Quality Assurance Division of the Education Bureau, 2008).

To guide the process of assessment reform, the Education Bureau implemented a School Development and Accountability Framework in the 2003-2004 school year. The framework emphasizes School Self-Evaluation, as well as an external peer evaluation, using a set of performance indicators. For example, with respect to curriculum and assessment, the performance indicators ask: "Is the school able to adopt varied modes of assessment and effectively assess students' performance in respect of knowledge, skills, and attitude?" and "How does the school make use of curriculum evaluation data to inform curriculum planning?" (Quality Assurance Division of the Education Bureau,

2008). The Education Bureau also conducts a Quality Assurance Inspection, which involves personnel from other schools and from the Education Bureau in on-site examinations of how each school meet the performance indicators.

Singapore

Greater emphasis on school-based assessment integrated into large-scale testing systems has also occurred in Singapore. Singapore's education system has been a source of intense interest for policy analysts since its students took first place in the TIMSS (Trends in International Mathematics and Science Study) assessments in mathematics and science in 1995, 1999 and 2003. These rankings are based on strong achievement for all of the country's students, including the Malay and Tamil minorities, who have been rapidly closing what was once a yawning achievement gap (Dixon, 2005). About 90% of Singapore's students scored above the international median on the TIMSS tests. This accomplishment is even more remarkable given that fewer than half of Singapore's students routinely speak English, the language of the test, at home. Most speak one of the other four official national languages of the country – Mandarin, Malay, or Tamil – and some speak one of several dozen other languages or dialects.

Intensive investment and reform over the period of thirty years have transformed the Singaporean education system, broadening access and increasing equality while orchestrating a system that includes a complex system of private, "autonomous," and public schools, some of them inherited from the colonial era, all of which receive government subsidies. These schools are intentionally diverse in many ways, as local schools are urged to innovate, but purposefully common in instructional expectations and supports, using a common national curriculum for core subjects.

Since the prime minister introduced the "thinking schools, learning nation" initiative in 1997, Singapore's explicit focus in its reforms of curriculum, assessment, and teaching has been to develop a creative and critical thinking culture within schools, by explicitly teaching and assessing these skills for students – and by creating an inquiry culture among teachers as well, who are supported to conduct action research on their teaching and to continually revise their teaching strategies in response to what they learn. This initiative was married to commitments to integrating technology into all aspects of education – a mission nearly fully accomplished a decade later – and to dramatically opening up college and university admissions.

Higher education is now available to virtually all Singaporeans. Based on their interests, labor force needs, and the results of their grades, O-level exams, and other accomplishments, students pursue one of three pathways after 10th grade when secondary school ends: about 25% attend Junior College for two years, followed by university, which leads to professional paths such as teaching, science, engineering, medicine, law, and the civil service; about 60% attend a polytechnic college for 3 years, after which about half go on to the university while the others go into jobs in technical and engineering fields; and the remainder – about 15% -- attend an Institute of Technical Education for two years, and some then continue onto college or university. Virtually everyone finishes one of these pathways.

Historically, the schools have operated a modified British-style system. Students sit for national exams administered by the Singapore Examinations and Assessment Board (SEAB). At the end of Year 6 (age 12), students take the Primary School Leaving Examinations (PSLE). These are open-ended written and oral examinations in four core

subject areas: mathematics, science, English and a “mother tongue” language, which are administered and scored by teachers in moderated scoring sessions. The exams in the English and native languages include four components – two written essays of at least 150 words, listening comprehension, language comprehension, and an oral exam that requires students to engage in a conversation on a set topic for 15 minutes. Two examiners observe the candidates and grade the oral proficiency of the student. In math, students have to demonstrate the steps in solving a problem.

Students then take the General Certificate of Examinations Normal or Ordinary Level (GCE N/O-Level) at the end of year 10 (age 16). The GCE O-level examinations are based on common course syllabi that outline what is to be taught; they require short and long open-ended responses and essays across a wide range of content areas from which students choose the ones in which they want to be examined. Although the results are used to guide postsecondary admissions, and are not used to determine graduation from high school, they exert substantial influence on the high school curriculum. Recent reforms are changing the curriculum and assessment system to make it more explicitly focused on creativity and independent problem solving.

Students attending Junior College (grades 11 and 12) en route to university take the GCE Advanced Level (A-Level) exams at the end of year 12 (age 18). A new ‘A’ level curriculum and examination system was introduced in 2002. The new exams are meant to encourage multi-disciplinary learning by requiring that students “select and draw together knowledge and skills they have learned from across different subject areas, and apply them to tackle new and unfamiliar areas or problems” (Singapore Examinations and Assessment Board, 2006, p. 2).

The A-level curricular framework includes Core Content Areas in which students take courses and associated exams: humanities, mathematics and sciences, and languages. It also includes Life Skills – emphasizing leadership, enrichment, and service to others – and Knowledge Skills, evaluated through a general paper, project work, and a course in knowledge and inquiry. A typical A-level student is evaluated in three compulsory subjects – a general paper, project work, and a native language assessment -- along with four content subjects.

The newer areas of Life Skills and Knowledge Skills are intended to develop the more advanced thinking skills thought to be underrepresented in the traditional content-based curriculum and examinations system. They represent the goals of reforms launched in 1997 as part of the "thinking schools, learning nation" initiative, which created a number of changes:

> Syllabi, examinations and university admission criteria were changed to encourage thinking out of the box and risk-taking. Students are now more engaged in project work and higher order thinking questions to encourage creativity, independent, and inter-dependent learning (Ng, 2008, p. 6).

The content courses are also evolving to include more critical thinking, inquiry and investigation, along with mastery of content. A number of the high school content tests are accompanied by school-based tasks, such as research projects and experiments designed and conducted by students. Each of the science courses now includes a component called the "School-based Science Practical Assessment" (SPA). These school-based components, which are managed and scored by teachers according to specifications provided by the Examinations Board, count for up to 20% of the

examination grade. Scoring is both internally and externally moderated. The goal is for students to be able to:

1. Follow a detailed set or sequence of instructions and use techniques, apparatus and materials safely and effectively;

2. Make and record observations, measurements, methods and techniques with precision and accuracy;

3. Interpret and evaluate observations and experimental data; and

4. Identify a problem, design and plan investigations, evaluate methods and techniques, and suggest possible improvements in the design.

The projects can be submitted to the university as part of the application, and universities are encouraged to examine evidence about student accomplishments beyond examination scores. Below we describe some of these innovations in the examination system.

Project Work

Project Work (PW) is **an interdisciplinary subject** that is compulsory for all pre-university students. There is dedicated curriculum time for students to carry out their project tasks over an extended period. As an interdisciplinary subject, it breaks away from the compartmentalization of knowledge and skills to focus on interdisciplinary outcomes by requiring students to draw knowledge and apply skills from across different subject domains. The goals for this experience are embedded in the requirements for the task and its assessment, which are centrally set by the Singapore Examinations and Assessment Board. The tasks are designed to be sufficiently broad to allow students to carry out a project that they are interested in while meeting the task requirements:

- **It must foster collaborative learning through group work.** Together as a group which is randomly formed by the teacher, students brainstorm and evaluate each others' ideas, agree on the project that the group will undertake and decide on how the work should be allocated amongst themselves.
- **Every student must make an oral presentation:** Individually and together as a group, each student makes an oral presentation of his / her group project in the presence of an audience
- **Both product and process are assessed:** There are 3 components for assessment:

-- the **Written Report** which shows evidence of the group's ability to generate, analyze and evaluate ideas for the project;

-- the **Oral Presentation** in which each individual group member is assessed on his/her fluency and clarity of speech, awareness of audience as well as response to

questions. The group as a whole is also assessed in terms of the effectiveness of the overall presentation;

-- the **Group Project File** in which each individual group member submits three documents related to 'snaphsots' of the processes involved in carrying out the project. These documents show the individual student's ability to generate, analyze and evaluate (i) preliminary ideas for a project (ii) a piece of research material gathered for the chosen project and (iii) insights and reflections on the project.

In carrying out the PW assessment task, students are intended to acquire self-directed inquiry skills as they propose their own topic, plan their timelines, allocate individual areas of work, interact with teammates of different abilities and personalities, gather and evaluate primary and secondary research material. These PW processes reflect life skills and competencies such as knowledge application, collaboration, communication and independent learning, which prepare students for the future workplace.

About 12,000 students complete this task annually. Assessment is school-based and criterion-referenced. While task setting, conditions, assessment criteria, achievement standards and marking processes are externally specified by SEAB, the assessment of all three components of PW is carried out by classroom teachers, using a set of assessment criteria provided by the board. All schools are given exemplar material that illustrates the expected marking standards. The Board provides training for assessors and internal moderators. Like all other assessments, the grading is both internally and externally moderated.

Knowledge and Inquiry is a Humanities subject that seeks to develop in students

- **An understanding of the nature and construction of knowledge**: Students are expected to show that they have read widely and have understood and can apply the concepts involved. They are expected to demonstrate skill in selecting relevant material with which to tackle the assessment tasks.
- **Critical Thinking:** Students are expected to demonstrate skills of critical thinking. They are expected to analyze different kinds of arguments and information, identify and evaluate assumptions and points of view, verify claims and provide reasoned and supported arguments of their own.
- **Communication Skills:** Students are expected to communicate their ideas and arguments clearly and coherently in good English. They are expected to structure their arguments, and select an appropriate style of presentation, to communicate responses which are fully relevant to the questions asked and demonstrate clear ability to engage with different aspects of these questions.

There are three assessment components:

Essay: This paper gives candidates the opportunity to demonstrate their ability to apply the concepts they have learned in their study of the nature and construction of knowledge. It covers the theoretical aspects of areas of exploration identified in the

syllabus and the questions set will require candidates to draw on knowledge they have gained during their study of the following key questions:

Why ask questions?
What is knowledge?
How is knowledge constructed?
What makes knowledge valid?
How is knowledge affected by society?
How should knowledge be used?

Critical Thinking: This paper requires students to critically analyze different kinds of arguments and information presented in the Material, identify and evaluate assumptions and points of view, and verify claims, and to provide reasoned and supported arguments. Students must use language appropriately and effectively to communicate a clear and well structured argument.

Independent Study: The Independent study component allows students to demonstrate their understanding of the nature and construction of knowledge as it relates to their chosen area of study, apply this understanding in addressing the specific context, select appropriate material and show that they have engaged in relevant reading during the course of their research by presenting a literature review and applying what they have read to support the arguments they present. Students must use language appropriately and effectively to communicate a clear and well structured argument. At the end of the 6 months of independent research study, they submit an extended essay of 2,500 to 3,000 words.

The kinds of more intellectually challenging school-based assessment in the high school examinations are also encouraged in the earlier grades as well. In the curriculum and assessment guidelines that accompany the national standards, teachers are encouraged to engage in continual assessment in the classroom, using a variety of assessment modes, such as classroom observations, oral communication, written assignments and tests, and practical and investigative tasks. The Ministry has developed a number of curriculum and assessment supports for teachers. For example, SAIL (Strategies for Active and Independent Learning) aims to support more learner-centered project work in classrooms, and provides assessment rubrics to clarify learning expectations. All schools have received training for using these tools.

The Ministry's 2004 Assessment Guides for both primary and lower secondary mathematics contain resources, tools and ideas to help teachers incorporate strategies such as mathematical investigations, journal writing, classroom observation, self-assessment and portfolio assessment into the classroom. Emphasis is placed on the assessment of problem solving and on meta-cognition, the self-regulation of learning that will enable students to internalize standards and become independent learners (Kaur, 2005). The Institute of Education has held a variety of workshops to support learning about the new assessments and integrated the new strategies into teacher development programs.

United Kingdom

The move toward more school-based assessment has also occurred in various ways in the United Kingdom, which has, for more than a century, had some influence on examination systems in English-speaking countries around the world. Assessments have typically been open-ended essay and constructed-response examinations, but the nature of the tasks and of the administration have been changing over the last two decades to include more school-based tasks and projects.

England

England's assessment system is managed at the national level by an organization called the Qualifications and Curriculum Authority (QCA). Schools teach and assess students using a national curriculum, which includes syllabi for specific courses. Teachers assess pupils' progress continuously and assemble evidence for external reporting in the national data system at ages 7, 11, and 14 (Key Stages 1, 2, and 3). This evidence is based on classroom-based assignments, observations, and tasks, the results of

which are evaluated in terms of indicators of performance outlined in learning progressions for each of several dimensions of learning within each subject area.

At key stage 1, ages six to seven, student progress is evaluated based on classroom evidence and results from centrally-developed open-ended tests and tasks in English and mathematics. These tests and tasks are marked by teachers and moderated within the school and by external moderators. At key stage 2, ages eight through eleven, student progress is evaluated based on teachers' summary judgments and results from open-ended tests in English, mathematics, and science. These tests are externally marked and the results reported on a national level. At key stage 3, England recently abolished external tests and now relies on teacher assessments to report achievement levels in all subjects. Teacher judgments are moderated and results are reported on a national level.

The Assessing Pupils' Progress program that guides this work is described by the QCA in this way:

> APP is the new structured approach to teacher assessment, developed by QCA in partnership with the National Strategies, which equips teachers to make judgments on pupils' progress. It helps teachers to fine-tune their understanding of learners' needs and to tailor their planning and teaching accordingly, by enabling them to: use diagnostic information about pupils' strengths and weaknesses to improve teaching, learning and pupils' progress; make reliable judgments related to national standards drawing on a wide range of evidence; and track pupils' progress.
>
> The APP subject materials for teachers include assessment guidelines for assessing pupils' work in relation to national curriculum levels. These provide a simple recording format providing assessment criteria for each of the assessment focuses in the subject, and standards files, which are annotated collections of pupils' day-to-day work that exemplify national standards at different levels. These help teachers reach consistent and reliable judgments about national curriculum levels (Qualifications and Curriculum Authority, 2009, p. 1.)

Some nationally-developed tasks are designed and distributed to schools to support teacher assessment. At Key Stage 2 (age 11), a set of these tasks and tests must be used to evaluate students, in combination with the other evidence teachers assemble from the classroom. In other years, the use of the tasks is optional. As described by the QCA: "The tasks are designed to support teacher assessment. They can be used to indicate what pupils are able to do and inform future learning and teaching strategies. Individual tasks can be used to provide a basis for discussion by teachers and pupils on what has been achieved and to identify the next steps. They can support day-to-day assessment and generate outcomes which can contribute to the breadth of evidence which is used as the basis for periodic and transitional assessment."

At key stage 4, ages fifteen to sixteen, the national qualification framework includes multiple pathways for students and consequently multiple measures of student achievement. There are four pathways based on students' aspirations after graduation: apprenticeship, diploma, the General Certificate of Secondary Education (GCSE), and the A-Level examinations. Some students go on to a Further Education college to take vocationally related courses. They usually take the National Vocational Qualification using the apprenticeship model.

Most students take the GCSE, a two year course of study evaluated by assessments both within and at the end of courses or unit. Students may take as many single-subject or combined-subject assessments as they like, and they choose which ones they will take based on their interests and areas of expertise. The exams involve constructed response items and structured, extended classroom-based tasks which comprise from 25 to 60% of the final examination score. England is currently piloting

new tasks for the GCSE with an increased emphasis on functional skills like problem solving, team building, and communication as well as personal learning and thinking skills across subjects. These new tasks, called "controlled assessments" are either designed by the awarding body and marked by teachers or designed by teachers and marked by the awarding body. Either way teachers determine the timing of controlled assessments.

These classroom-based assessments comprise 25% of the total examination score in subjects like business studies, classical civilization, English literature, geography, history, humanities, or statistics, and 60% of the total examination score in subject areas such as applied business, music and dance, design and technology, drama, engineering, English, English Language, expressive arts, health and social care, home economics, ICT, manufacturing, media studies, and modern foreign languages. Below are examples of classroom-based tasks in English and ICT.

Example of Tasks: GCSE English	
Unit and Assessment	**Tasks**
Reading literacy texts Controlled assessment (coursework) 40 marks	Responses to three texts from choice of tasks and texts. Candidates must show an understanding of texts in their social, cultural and historical context
Imaginative Writing Controlled assessment (coursework) 40 marks	Two linked continuous writing responses from a choice of Text Development or Media
Speaking and Listening Controlled assessment (coursework) 40 marks	Three activities: a drama-focussed activity; a group activity; an individual extended contribution. One activity must be a real-life context in and beyond the classroom
Information and Ideas Written exam 80 marks (40 per section)	Non-Fiction and Media: Responses to unseen authentic passages Writing information and Ideas: One continuous writing response – choice from 2 options

GCSE Controlled Assessment Task in ICT
(Interactive Computer Technology)

Litchfield Promotions works with over 40 bands and artists to promote their music and put on performances in England. The number of bands they have on their books is gradually expanding. Litchfield Promotions needs to be sure that each performance will make enough money to cover all the staffing costs and overheads as well as make a profit. Many people need to be paid: the bands; sound engineers; and, lighting technicians. There is also the cost of hiring the venue. Litchfield Promotions needs to create an ICT solution to ensure that they have all necessary information and that it is kept up to date. Their solution will show income, outgoings and profit.

Candidates will need to: 1) Work with others to plan and carry out research to investigate how similar companies have produced a solution. The company does not necessarily have to work with bands and artists or be a promotions company. 2) Clearly record and display your findings. 3) Recommend a solution that will address the requirements of the task. 4) Produce a design brief, incorporating timescales, purpose and target audience.

Produce a solution, ensuring that the following are addressed: 1) It can be modified to be used in a variety of situations. 2) It has a friendly user interface. 3) It is suitable for the target audience. 4) It has been fully tested. You will need to: 1) incorporate a range of: software features, macros, modeling, and validation checks - used appropriately. 2) Obtain user feedback. 3) Identify areas that require improvement, recommending improvement, with justification. 4) Present information as an integrated document. 5) Evaluate your own and others' work.

During key stage 4, most students take five or more GCSE exams. Their performance determines the level of the diploma they receive, and whether they will go on to Advanced Studies that are later evaluated by A-level exams that qualify students for university admissions. England has 45 areas for A-level exams. The exam questions require extended answers aimed at assessing deeper levels of understanding and applications of knowledge to real-world problems, as illustrated in the example below.

Sample A-Level Question from a Probability and Statistics 1 Exam

A City council attempted to reduce traffic congestion by introducing a congestion charge. The charge was set for 4 pounds for the first year and was then increased by 2 pounds each year. For each of the first eight years, the council recorded the average number of vehicles entering the city center per day. The results are shown in the table:

Charge, (Pounds) x	4	6	8	10	12	14	16	18
Average number of vehicles per day, y million	2.4	2.5	2.2	2.3	2.0	1.8	1.7	1.5

1. Calculate the product moment correlation coefficient for these date.

2. Explain why x is the independent variable.
3. Calculate the equation of the regression line of *y* on *x*.
4a. Use your equation to estimate the average number of vehicles, which will enter the city center per day when the congestion charge is raised to 20 pounds.
4b. Comments on the reliability of your estimate.
5. The council wishes to estimate the congestion charge required to reduce the average number of vehicles entering the city per day to 1.0 million. Assuming that a reliable estimate can be made by extrapolation, state whether they should use the regression line of *y* on *x* or the regression line of *x* on *y*. Give a reason for your answer.

.

Most of the exams take the form of essay questions. The mathematics exams include questions that ask students to show their reasoning behind their answers. Foreign language exams require oral presentations. The 'A' Level exam in English literature asks students to show their skills and knowledge in four sections: poetry, drama, prose, and general, analyzing works of literature they have read as part of their curriculum in terms of their meaning and interpretation as well as literary devices and writing strategies. Coursework accounts for 25 to 30% of the 'A' Level score, depending on the course. Students must now also complete an independently-designed extended research project as part of the A-level assessments. Assessments are marked by teachers in a moderated process managed by the five examination agencies that organize sets of examinations.

While England has moved to include some school-based assessments in its increasingly performance-oriented assessment system, Scotland, Wales, and Northern Ireland have gone even further in revising their approaches to assessment.

Scotland

Scotland has a separate governing body for its educational system from the United Kingdom. Scotland uses a set of assessments called the Scottish Survey of Achievement administered in the third, fifth, and seventh years of primary school and standardized courses and benchmark exams in secondary school. The assessment tasks for the primary

courses and general secondary courses are designed and marked by teachers and lecturers. Schools use external assessments for the intermediate and advanced secondary courses. The Scottish Qualifications Authority designs and scores those assessments which may take the form of an examinations, project work, or portfolios (Scottish Qualifications Authority, March 2004; The Scottish Government, 2008).

Wales

Wales only recently separated from the system used in the United Kingdom and now has a separate governing body for its educational system (Archer, 2006). The more centralized system introduced in England under the Thatcher administration (later modified during the Blair administration as described above) inspired policies like No Child Left Behind in the United States, on the one hand, but also caused consternation among countries in the United Kingdom that favored a different approach. Wales broke from the British system and opted to abolish national exams for children through age 14.

Much like Finland, during the primary years, Welsh schools have a national school curriculum supported by teacher-created, administered, and scored assessments. During the secondary years, teachers create and manage all assessment of 14-year-old students, while students 16 years and older are encouraged to participate in the relevant GCSE exams and A Level courses and exams administered by the U.K.'s Qualifications and Curriculum Authority (Welsh Assembly Government, 2008a and 2008b). With these changes to its assessment system, Wales hopes to increase student engagement and reduce teaching to the test (Archer, 2006).

Northern Ireland

Northern Ireland is in the process of implementing an approach at all levels called "Assessment for Learning." This approach emphasizes locally developed, administered and scored assessments and focuses on five key actions:

1) *Sharing learning intentions* where students and teacher agree upon learning intentions to give them ownership over their learning.
2) *Sharing and negotiating success criteria* where students and teacher create the criteria for successful completion of a task together to help with self-assessment.
3) *Feedback* where teacher provide on-going feedback during formative assessment sessions.
4) *Effective questioning* where teachers introduce strategies like using open-ended questions and giving more thinking time so students will feel more confident thinking aloud and explaining their reasoning.
5) *How pupils reflect on their learning* where teachers provide students with strategies to think about what they have learned.

Northern Ireland does not require schools to externally assess students up through age 14, but it provides teachers with the option to give students end of Stage 3 assessments that are externally graded through the Northern Ireland Council for the Curriculum Examinations and Assessments (CCEA). These are largely open-ended assessments that evaluate how students reason, think, and problem solve. CCEA provides multiple assessments for Stage 4, according to which pathway a student chooses to follow including taking the GCSE exam and A level courses and exams from the U.K. system (i.e. whether towards university or a vocational degree) (Council for the Curriculum Examinations and Assessment, 2008a and b).

International Baccalaureate Diploma Program

The International Baccalaureate Diploma Programs is offered in 1600 Schools serving 100,000 students in 125 countries. Like systems in many high-achieving nations, the IB high school curriculum is standards-based and syllabus-based, integrating assessment within the curriculum in a tightly constructed teaching and learning system that blends classroom-based and external examinations. The IB program generally serves students in grades 11 and 12, assessing students using school-based assessments throughout the two-year program and externally-based exams at the end of the two-year course of study. Both types of assessments measure students' individual performance on the objectives outlined by syllabi, or "subject outlines" written by the International Baccalaureate Organization (IBO), the non-profit educational foundation that provide services and support to IB schools worldwide.

In almost all of the subjects, teachers conduct school-based assessments by grading individual pieces of course work based on the objective set out by the IB subject outlines. School-based assessments contribute between 20 and 30% of the total grade in most subjects and as much as 50% in arts courses like music, theater arts, and visual arts. Coursework graded by teachers includes such assessments as oral exercises in language subjects, projects, student portfolios, class presentations, practical laboratory work, mathematical investigations and artistic performances (International Baccalaureate Organization, 2008).

The externally-based exams usually consist of essays, structured problems, short-response questions, data-response questions, text-response questions, case-study questions, and a limited use of multiple choice questions. There are a limited number of

externally assessed pieces of work (i.e., a theory of knowledge essay and extended essay and world literature assignment) that students complete over an extended period of time under teacher supervision, but marked by external evaluators, or "IB Examiners," personnel trained and organized by the IBO.

Teachers design, administer, and grade school-based assessment tasks. Externally-based assessment tasks are designed, administered, and graded by IB examiners. IB uses criterion-based exams that rate students' performance in comparison to a set of standards set by the IBO (International Baccalaureate Organization, 2008).

IB externally-based exams ask students to apply the analytical and problem solving skills and content knowledge they gain in their IB coursework to specific problems or analyses. For example, a sample "English A1 – Higher Level" essay asks students to answer one essay question and base their answer on a least two of three works studied in class. Students pick from 5 categories: drama, poetry, prose: the novel and short story, prose: other than the novel and short story, general questions on literature. On the English exam, students may choose from essay questions like the following:

1) Using two or three of the works you have studied, discuss how and to what effect writers have used exaggeration as a literary device.

2) Acquiring material wealth or rejecting its attractions has often been the base upon which writers have developed interesting plots. Compare the ways the writers of two or three works you have studied have developed such motivations.

3) Discuss and compare the role of the speaker or persona in poems you have studied. You must refer closely to the work of two or three poets in your study and base your answer on a total of three or four poems (International Baccalaureate Organization, 2005).

A typical "Mathematics, Standard Level, Paper 2" essay asks students to show their work and support their answers with work and explanations. It also asks students to

draw any graphs they create on their graphing calculator, and tell students that they could receive credit for using the correct method even if another error creeps in to produce an incorrect final answer. The exam includes five multi-step questions. The following question addresses multiple math skills including probability, proportions, and algebra:

> In a large school, the heights of all fourteen-year-old students are measured. The heights of the girls are normally distributed with mean 155cm and standard deviation 10cm. The heights of the boys are normally distributed with mean 160cm and standard deviation 12cm.
> (a) Find the probability that a girl is taller than 170 cm.
> (b) Given that 10% of the girls are shorter than x cm, find x.
> (c) Given that 90% of the boys have heights between q cm and r cm where q and r are symmetrical about 160 cm, and $q < r$, find the value of q and of r.
>
> In the group of fourteen-year-old students, 60% are girls and 40% are boys. The probability that a girl is taller than 170 cm was found in part (a). The probability that a boy is taller than 170 cm is 0.202.
> A fourteen-year-old student is selected at random.
> (d) Calculate the probability that the student is taller than 170 cm.
> (e) Given that the student is taller than 170 cm, what is the probability the student is a girl? (International Baccalaureate Organization, 2006).

In addition to these kinds of assessments of specific course-based learning, the IB course of studies, like the British A-levels and the Singaporean exams, requires students to complete an extended essay. This paper is an independent, self-directed piece of research, culminating in a 4,000-word paper. According to the IB organization, it is intended to provide practical preparation for the kinds of undergraduate research required in college, and an opportunity for students to engage in an in-depth study of a topic of interest within a chosen subject. Emphasis is placed on the research process: formulating an appropriate research question, engaging in a personal exploration of the topic, communicating ideas, and developing an argument. Participation in this process is intended to develop students' capacity to analyse, synthesize, and evaluate knowledge.

Conclusion

An examination of assessment practices in a number of high-achieving nations and rigorous programs like the International Baccalaureate illustrates the importance of assessment *of*, *for* and *as* learning, rather than as a separate disjointed element the education enterprise. High-quality assessments provide feedback to students, teachers, and schools about what has been learned, and they "feedforward" information that can shape future learning, as well as guide college and career-related decision making.

These systems closely align curriculum expectations, subject and performance criteria, and desired learning outcomes. They engage teachers in assessment development and scoring as a way to improve their professional practice and their capacity to support student learning and achievement. They engage students in authentic assessments to improve their motivation and learning. They seek to advance student learning in higher-order thinking skills and problem-solving by using a wider range of instructional and assessment strategies. And they privilege quality over quantity of standardized testing – moving systems from 'accounting' to more useful 'accountability' for learning.

References

Archer, J. (December 19th, 2006). "Wales Eliminates National Exams for Many Students." *Education Week.* Retrieved on September 11th, 2008 from http://www.edweek.org/ew/articles/2006/12/20/16wales.h26.html?qs=Wales.

Buchberger, F. & Buchberger, I. (date). Problem Solving Capacity of a Teacher Education System as a Condition of Success? An Analysis of the "Finnish Case." In F. Buchberger and S. Berghammer (eds.), *Education Policy Analysis in a Comparative Perspective*, pp. 222-37. Linz: Trauner.

Chan, J.K., K.J. Kennedy, F.W. Yu, P. Fok. (2008). "Assessment Policy in Hong Kong: Implementation Issues for New Forms of Assessment." *The Hong Kong Institute of Education*. Retrieved on September 12th, 2008 from http://www.iaea.info/papers.aspx?id=68

Council for the Curriculum Examinations and Assessment. (2008a). "Curriculum, Key Stage 3, Post-Primary Assessment." Retrieved on September 12th, 2008 from http://www.ccea.org.uk/

Council for the Curriculum Examinations and Assessment. (2008b). "Qualifications." Retrieved on September 12th, 2008 from http://www.ccea.org.uk/

Dixon, Q. L. (2005). Bilingual Education Policy in Singapore: An Analysis of its Sociohistorical Roots and Current Academic Outcomes. International Journal of Bilingual Education and Bilingualism, 8 (1): 25-47.

Dowling, M. (undated). Examining the Exams. Retrieved on Sept. 14, 2008 from http://www.hkeaa.edu.hk/files/pdf/markdowling_e.pdf.

Eckstein, M.A. and H.J. Noah. (1993). *Secondary School Examinations: International Perspectives on Policies and Practice*. New Haven: Yale University Press.

European Commission. (2007/2008). "Eurybase, The Information Database on Education Systems in Europe, The Education System in Finland."

European Commission. (2006/2007). "Eurybase, The Information Database on Education Systems in Europe, The Education System in Sweden."

The Finnish Matriculation Examination. (2008). Retrieved on September 8th, 2008 from http://www.ylioppilastutkinto.fi/en/index.html

Finnish National Board of Education. (November 12th, 2007). "Background for Finnish PISA Success." Retrieved on September 8th, 2008 from http://www.oph.fi/english/SubPage.asp?path=447,65535,77331

Finnish National Board of Education. (April 30th, 2008). "Teachers." Retrieved on September 11th, 2008 from http://www.oph.fi/english/page.asp?path=447,4699,84383

Finnish National Board of Education. (June 10th, 2008). "Basic Education." Retrieved on September 11th, 2008 from http://www.oph.fi/english/page.asp?path=447,4699,4847

Hong Kong Examinations Assessment Authority (HKEAA) (2007a). 2007 Annual Report. Introduction, p. 6. Retrieved on September 14, 2008 from http://eant01.hkeaa.edu.hk/hkea/redirector.asp?p_direction=body&p_clickurl=http%3A%2F%2Fwww%2Ehkeaa%2Eedu%2Ehk%2Fen%2Fannual%5Freport%2Ehtml

HKEAA (2007b, January 28). School-based assessment: Changing the assessment culture. http://www.hkeaa.edu.hk/en/hkdse/School_based_Assessment/SBA/

International Baccalaureate Organization. (2008). "Diploma program assessment: Methods" Retrieved on May 31st, 2008 from http://www.ibo.org/diploma/assessment/methods/

International Baccalaureate Organization. (May 2006). "IB Diploma Programme: Mathematics, Standard Level, Paper 2." Retrieved on May 31st, 2008 from http://www.ibo.org/diploma/curriculum/examples/samplepapers/documents/gp5_mathssl2.pdf

International Baccalaureate Organization. (November 2005). "IB Diploma Programme: English A1 – Higher Level – Paper 2." Retrieved on May 31st, 2008 from http://www.ibo.org/diploma/curriculum/examples/samplepapers/documents/gp1_englisha1hl2.pdf

Kaur, B. (2005). Assessment of Mathematics in Singapore Schools – The Present and Future. Singapore: National Institute of Education.

Kaftandjieva, F. & Takala, S. (2002). "Relating the Finnish Matriculation Examination English Test Results to the CEF Scales."

Korpela, Salla. (December 2004). "The Finnish school – a source of skills and well-being: A day at Stromberg Lower Comprehensive School." Retrieved on September 11th, 2008 from http://virtual.finland.fi/netcomm/news/showarticle.asp?intNWSAID=30625

Laukkanen, R. (2008). Finnish Strategy for High-Level Education for All, in N. C. Soguel and P. Jaccard (eds.), *Governance and Performance of Education Systems*. (Springer).

Lavonen, J. (2008). "Reasons behind Finnish Students' Success in the PISA Scientific Literacy Assessment." University of Helsinki, Finland. Retrieved on September 8th, 2008 from

http://www.oph.fi/info/finlandinpisastudies/conference2008/science_results_and_reasons.pdf

Murad, L.C. (2008). "Hong Kong's Education System: Challenge for the Future." Retrieved on September 12th, 2008 from http://www.lehigh.edu/~incntr/publications/perspectives/v20/Murad.pdf

Ng, P.T. (2008). Educational reform in Singapore: from quantity to quality. Education Research on Policy and Practice,7:5–15

O'Donnell. (December, 2004). International Review of Curriculum and Assessment Frameworks. Comparative tables and factual summaries – 2004. Qualifications and Curriculum Authority and National Foundation for Educational Research. Retrieved on September 11th, 2008 from http://www.inca.org.uk/pdf/comparative.pdf

Petterson, A. (2008). The National Tests and National Assessment in Sweden. Stockholm, Sweden. Stockholm Institute for Education. PRIM gruppen. Retrieved on May 31st, 2008 from http://www.prim.su.se/artiklar/pdf/Sw_test_ICME.pdf

Qualifications and Curriculum Authority (2008a). "Sweden: Assessment arrangements." Retrieved on September 11th, 2008 from http://www.inca.org.uk/690.html

Qualifications and Curriculum Authority. (2008b). "England: Assessment arrangements." Retrieved on May 27th, 2008 from http://www.inca.org.uk/1315.
http://education.qld.gov.au/corporate/newbasics/html/richtasks/richtasks.html html

Qualifications and Curriculum Authority (2009). Assessing pupils' progress: Assessment at the heart of learning. Retrieved on 5/23/09 from
http://www.qca.org.uk/libraryAssets/media/12707_Assessing_Pupils_Progress_leaflet_-_web.pdf

Quality Assurance Division of the Education Bureau. (2008). "Performance Indicators for Hong Kong Schools, 2008 with Evidence of Performance." Retrieved on September 12th, 2008 from
http://www.edb.gov.hk/FileManager/EN/Content_6456/pi2008%20eng%205_5.pdf

Queensland Government (2001). "New Basics: The Why, What, How and When of Rich Tasks." Retrieved on September 12th, 2008 from
http://education.qld.gov.au/corporate/newbasics/pdfs/richtasksbklet.pdf

Rustique-Forrester, E. (2005). Accountability and the Pressures to Exclude: A Cautionary Tale from England. *Education Policy Analysis Archives*. Retrieved on 10/1/09 from http://epaa.edu/epaa/v13n26.

Schmidt, W. H., Wang, H.C. & McKnight, C. (2005). Curriculum Coherence: An Examination of U.S. Mathematics and Science Content Standards from an International

Perspective. *Journal of Curriculum Studes, 37* (5); 525-59.

The Scottish Government. (2008). "Schools: Attainment." Retrieved on September 11th, 2008 from http://www.scotland.gov.uk/Topics/Education/Schools/curriculum/Attainment

Scottish Qualifications Authority. (March 2004). "Scotland's National Qualifications: Quick Guide." Retrieved on September 11th, 2008 from http://www.sqa.org.uk/files_ccc/NQQuickGuide.pdf

Singapore Examinations and Assessment Board (2006). 2006 A-Level Examination. Singapore: Author.

Singapore Ministry of Education (2007). From the Ministry of Education website. http://www.moe.gov.sg/corpora/mission_statement.htm (Accessed 8/25/08)

State of Colorado House Bill 03-1108. (2003). Retrieved on May 31st, 2008 from http://www.state.co.us/gov_dir/leg_dir/olls/sl2003a/sl_153.htm

Swedish Institute. (March 1984). "Primary and Secondary Education in Sweden. Fact Sheets on Sweden." Stockholm, Sweden.

Swedish National Agency for Education. (2005). "The Swedish School System: Compulsory School." Retrieved on May 31st, 2008 from http://www.skolverket.se/sb/d/354/a/959

Texas Education Code, Section 51.968. (2005). Retrieved on May 31st, 2008 from http://www.legis.state.tx.us/tlodocs/79R/billtext/html/HB00130I.htm.

Valverde, G.A. & Schmidt, W.H. (2000). Greater Expectations: Learning from Other Nations in the Quest for 'World-Class Standards' in U.S. School Mathematics and Science. *Journal of Curriculum Studies, 32* (5): 651-87.

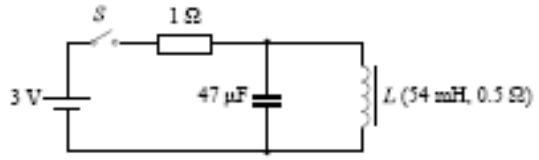
Welsh Assembly Government. (2008a). "Primary (3-11)." Retrieved on September 12th, 2008 from http://old.accac.org.uk/eng/content.php?cID=5

Welsh Assembly Government. (2008b). "Secondary (11-16)." Retrieved on September 12th, 2008 from http://old.accac.org.uk/eng/content.php?cID=6

Hong Kong High School Physics Test

8. In Figure 8.1, a 47 μF capacitor, an inductor L and a 1 Ω resistor are connected with a cell of e.m.f. 3 V and negligible internal resistance. The inductor L is of inductance 54 mH and resistance 0.5 Ω. Initially the capacitor is uncharged.

Figure 8.1



(a) Find the current flowing in the 1 Ω resistor

(i) when the switch S is just closed ;

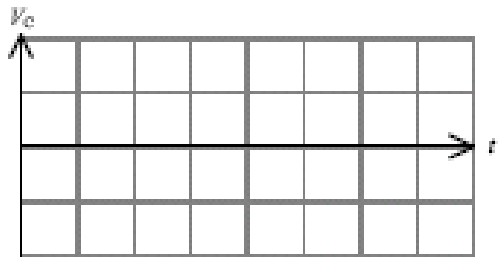
(ii) a few minutes after the switch S is closed.

Explain briefly. (4 marks)

(b) (i) Calculate the maximum p.d. across the capacitor. (2 marks)

(ii) Find the energy stored in the inductor at the steady state. (2 marks)

(iii) If switch S is now opened, sketch the time variation of the p.d. V_C across the capacitor. (2 marks)



(c) State how you would modify the circuit so as to demonstrate that a large induced e.m.f. is produced across the inductor when switch S is suddenly opened. Explain briefly. (3 marks)

Performance Assessment: The State of the Art

Suzanne Lane
University of Pittsburgh

This paper was supported by the project, Investigating Teaching and Learning Systems Around the World, funded by the Ford Foundation. The opinions expressed in this paper do not necessarily reflect the position of the Ford Foundation. This paper extends the work presented in two documents: 1) Lane & Stone, C.A. (2006). Performance Assessments. In B. Brennan (Ed.), *Educational Measurement*. American Council on Education & Praeger: Westport, CT. and 2) Lane, S. (in press). Issues in the Design and Scoring of Performance Assessments that Assess Complex Thinking Skills. In G. Schraw (Ed.). *Assessment of Higher Order Thinking Skills.*

Educational reform in the 1980's was based on the premise that too many students knew how to repeat facts and concepts, but were unable to apply those facts and concepts to solve realistic problems that require complex thinking and reasoning skills. Assessments needed to better reflect students' competencies in applying their knowledge and cognitive skills to solve substantive, meaningful tasks. Promising advances in the study of both cognition and learning in content domains and of psychometrics also prompted individuals to think differently about how students process and reason with information and how assessments can be designed to capture meaningful aspects of student learning. Performance assessments that assess complex cognitive skills were also considered to be valuable tools for educational reform by policy makers and advocates for curriculum reform (Linn, 1993; Resnick & Resnick, 1992). They were thought of as vehicles that could help shape sound instructional practice by modeling to teachers what is important to teach and to students what is important to learn. Carefully crafted performance assessments that measure complex thinking and reasoning skills can serve as exemplars of assessments that stimulate and enrich learning rather than just serve as indicators of learning (Bennett & Gitomer, in press; Black &William, 1998). Performance assessments are needed to assess the types of thinking and reasoning skills that are valued by educators, and cannot be assessed by other item formats such as multiple-choice items.

The use of performance tasks in large-scale assessments has declined with the requirements of the NCLB Act of 2001 (U.S. Department of Education, 2002). Under the NCLB Act, states are required to test all students from grades 3 through 8 annually in reading and mathematics, and students in high school at one grade level. Students also need to be tested in science at one grade level in elementary, middle and high school. An example of a successful performance assessment program prior to NCLB was the Maryland School Performance Assessment Program (MSPAP) that was designed for grades 3, 5 and 8 to measure school-level performance and provides information for school accountability and improvement (Maryland State Board of Education, 1995). These assessments were designed to promote performance-based instruction and classroom assessment in reading, writing, mathematics, science and social studies. The performance assessment tasks were interdisciplinary, required students to produce both short and extended written responses, and some required hands-on activities and collaboration with peers. Maryland's performance-based assessment was no longer tenable given the constraints imposed by the NCLB Act.

This chapter addresses design, scoring and psychometric advances in performance assessment that allow for the assessment of 21st century skills. Although performance assessments for classroom purposes are discussed, the focus of the chapter is on the use of performance assessments in large-scale

assessment programs. Advances in the integration of cognitive theories of learning and measurement models as they apply to the design of performance assessments are considered. The chapter begins with a discussion on the advances in the design of performance assessments, including a description of the important learning outcomes that can be assessed by performance assessments, and not by other assessment formats. The second section discusses advances in the scoring of performance assessments, including both the technical and substantive advances in automated scoring methods that allow for timely scoring of student performances to innovative item types. The third section addresses issues related to the validity and fairness of the use and interpretation of scores derived from performance assessments. The type of evidence needed to support the validity of score interpretations and use, such as content representation, cognitive complexity, fairness, generalizability and consequential evidence, is discussed. It should be noted, however, that validity and fairness are addressed throughout the chapter. The last section briefly addresses additional psychometric advances in performance assessments, including advances in measurement models used to capture student performance and rater inconsistencies as well as advances in linking performance assessments.

Design of Performance Assessments

In the design of any assessment, the type of score inferences one wants to make should first be delineated. This includes deciding on whether one wants to generalize to the larger construct domain of interest, or to provide evidence of a particular accomplishment or performance. The former requires sampling tasks from the domain to ensure content representativeness which will contribute to the validity of the score generalizations. This approach is typically used in the design of large-scale assessments, and the challenges of ensuring valid generalizations to the content domain using scores derived from performance assessments will be addressed later in the chapter. The latter approach requires the specification of a performance assessment that allows for the demonstration of a broader ability or performance which is similar to a “merit badge” approach. This approach, performance demonstration, is commonly used for classroom purposes such as a high school project or paper.

This section focuses on design issues that need to be considered to ensure that performance assessments are capable of eliciting the cognitive processes and skills that they are intended to measure, and to ensure the coherency among curriculum, instruction and assessment. Advances in the design of computer-based simulation tasks are addressed. The use of computers allows for the modeling of performance tasks that engage students in meaningful problem solving and reasoning skills and for the monitoring and scoring of student performances. Computer simulation tasks are ideal for capturing the multidimensionality of content domains and scores can be generated for the different dimensions being assessed. Advances in the use of learning progressions in assessment design are also addressed. The use

of learning progressions in the design of assessments is invaluable for monitoring an individual student or group's progress, and for informing instruction and learning. The importance of expert review and field testing performance assessments is also discussed. Throughout this section examples of performance assessments are also provided, some of which have been used in large-scale assessment programs.

Description of Performance Assessment

Performance assessments can measure students' cognitive thinking and reasoning skills and their ability to apply knowledge to solve realistic, meaningful problems. They are designed to more closely reflect the performance of interest, allow students to construct or perform an original response, and use predetermined criteria to evaluate student work. The close similarity between the performance that is assessed and the performance of interest is the defining characteristic of a performance assessment as described by Kane, Crooks, and Cohen (1999). As stated by the *Standards for Educational and Psychological Testing*, performance assessments attempt to "emulate the context or conditions in which the intended knowledge or skills are actually applied" (AERA, APA, & NCME, 1999, p.137). As this definition indicates, performance assessments do not have to assess complex reasoning and problem solving skills. As an example, if the targeted domain is the speed and accuracy at which students can keyboard, a measure that captures the accuracy and speed of students' keyboarding would be considered a performance assessment. Clearly, keyboarding is not a high level thinking skill but a learned automated procedural skill. The focus of this chapter will be on performance assessments that are designed to assess complex reasoning and problem solving skills in academic disciplines, and can be used for large-scale assessments. The Maryland School Performance Program (MSPAP) was an excellent example of a performance assessment that consisted of interdisciplinary tasks that assessed problem-solving, reasoning and evaluation skills (Maryland State Board of Education, 1995). As an example for a grade 5 Science MSPAP task, students were asked to investigate how a hydrometer can be used to measure different levels of saltiness (salinity) in various water samples, predict how the hydrometer might float in mixtures of fresh and salt water, and determine how the hydrometer could be used to establish the correct salinity for an aquarium. This hands-on task allowed students to conduct several investigations, make predictions, evaluate their work, and provide explanations for their responses.

Performance assessments require students to perform a task such as conducting a science investigation as described above, or to construct an original product or response such as writing an explanation of one's solution to a mathematics problem or writing a persuasive essay. Proficiency can be explained by the cognitive processes and skills involved in solving the performance task as well as the strategies chosen for solution, having the potential to provide rich information for diagnosing strengths as well as gaps in understanding for individual students as well as groups of students. When working on well

designed performance tasks, students may be engaged in applying their knowledge to real world problems, evaluating different approaches to solving problems, and providing reasoning for their solutions. A prevailing assumption underlying performance assessments is that they serve as motivators in improving student achievement and learning, and that they encourage instructional strategies that foster reasoning, problem-solving and communication (Frederiksen & Collins, 1989; National Council on Education Standards and Testing, 1992; Resnick & Resnick, 1982). Further, Mislevy (1993) pointed out that they allow for better measurements of change, both quantitatively and qualitatively. A hypothetical example of a quantitative measure of change for a mathematics performance assessment would be that a student used a recursive strategy only 2 out of 10 times during the first administration of a math assessment, but 6 out of 10 times during the second administration. An example of a qualitative evaluation of change would be that the student switched from a less effective strategy to the more sophisticated recursive strategy after instruction.

Performance assessments are contextualized, linking school activities to real world experiences (Darling-Hammond, Ancess, & Falk (1995), and can include opportunities for self reflection and collaboration as well as student choice, such as choosing a particular topic for a writing assignment (Baker, O'Neil, & Linn, 1993; Baron, 1991). Collaborative efforts were required on the Maryland State Performance Assessment Program (MSPAP, Maryland State Board of Education, 1995) in that students worked together on conducting science investigations and evaluated each other's essays. Collaboration is required on many performance assessments outside of this country, such as a mathematics assessment in Denmark aimed at 16 year olds (Black & William, 2006). It can be argued that these types of collaborations on performance assessments better reflect skills required in the 21st century. Performance assessments may also allow for a particular task to yield multiple scores in different content domains, which has practical as well as pedagogical appeal. Tasks that are designed to elicit scores in more than one content domain may not only reflect a more integrated approach in instruction, but also motivate a more integrated approach to learning. As an example, MSPAP tasks were integrated and would yield scores in two or more domains (Maryland State Board of Education, 1995). Practical implications for tasks that afford multiple scores may be reduced time and cost for task development, test administration and scoring by raters (Goldberg & Roswell, 2001). It is important, however, to provide evidence that each score represents the construct it is designed to measure and does not include construct-irrelevant variance (Messick, 1989).

Performance assessments, in particular, writing assessments, have been included in some large-scale assessment programs in this country for monitoring students' progress towards meeting national or state content standards, promoting educational reform, and holding schools accountable for student learning.

High-stakes decisions are typically required as well as an evaluation of changes in performance over time, which requires a level of standardization of the content to be assessed, of the administration of the assessment, and of the scoring of student performances over time. Thus, extended time periods, collaborative work, choice of task, and use of ancillary material may challenge the standardization of the assessment and consequently, the accuracy of the score interpretations. Large-scale performance assessment programs, such as MSPAP, however, have included these attractive features of performance assessments while ensuring the quality and validity of the score interpretations at the school level.

Another consideration in the design of assessments of complex skills is whether a portfolio approach will be used. The Advanced Placement (AP) Studio Art portfolios provide an excellent example of a large-scale portfolio assessment that has been sustained over time (Myford & Mislevy, 1995). As an example, in the 3-D Design portfolio students are required to submit a specified series of images of their 3-D artworks and their artworks are evaluated independently according to their quality (demonstration of form, technique and content), breadth (demonstration of visual principles and material techniques), and concentration (demonstration of depth of investigation and process of discovery). Using a well-delineated scoring rubric for each of these three areas, from three to seven artist-educators evaluate the submitted images of the artwork. The portfolios that are submitted are standardized in that specific instructions are provided to students that specify what type of artwork is appropriate and the students are provided with detailed scoring rubrics that delineate what is expected for each of the dimensions being assessed.

Performance assessments that are aligned with curriculum and instruction can provide valuable information to guide the instructional process. Thus, it is imperative to ensure that both classroom and large-scale assessments are aligned to the curriculum and instruction. A rich, contextualized curriculum-embedded performance assessment that is used for classroom purposes does not require the level of standardization as the typical large-scale performance assessment. These curriculum-embedded assessments allow teachers to more fully understand the ways in which students understand the subject matter, and can help guide day-to-day instruction. Large-scale assessments can also inform instruction, but at a broader level for both individual students and groups of students (e.g. classrooms).

Cognitive Theories in the Design of Performance Assessments

The need for models of cognition and learning and quantitative psychometric models to be used together to develop and interpret achievement measures has been widely recognized (Embretson, 1985; Glaser, Lesgold & Lajoie, 1987; National Research Council, 2001). The deeper the understanding of how individuals acquire and structure knowledge and cognitive skills, and how they perform cognitive tasks, the better able we are to assess students' cognitive thinking and reasoning and obtain information that will lead to improved learning. Substantial theories of knowledge acquisition are needed in order to design

assessments that can be used in meaningful ways to guide instruction and monitor student learning. Several early research programs that have had a direct impact on the assessment of achievement studied the difference between experts' and novices' knowledge structures (e.g., Simon & Chase, 1973; Chi, Feltovich, & Glaser, 1981). Chi and her colleagues (1981) demonstrated that an expert's knowledge in physics is organized around central principals of physics, whereas a novice's knowledge is organized around the surface features represented in the problem description. It is important to point out that much of what is known in the development of expertise is based on studies of students' acquisition of knowledge and skills in content domains.

Other early approaches that link cognitive models of learning and psychometrics have drawn upon work in the area of artificial intelligence (e.g., Brown & Burton, 1978). As an example, Brown and Burton (1978) represented the complex procedures underlying addition and subtraction as a set of component procedural skills, and described proficiency in terms of these procedural skills. Using artificial intelligence, they were able to uncover procedural errors or bugs in students' performance that represented misconceptions in their understanding. Cognitive task analysis using experts' talk alouds (Ericcson & Smith, 1991) has also been used to design performance assessments in the medical domain (Mislevy, Steinberg, Breyer, Almond, & Johnson, 1999). Features of the expert's thinking, knowledge, procedures, and problem posing are considered to be indicators of developing expertise in the domain (Glaser, Lesgold & Lajoie, 1987), and can be used systematically in the design of assessment tasks. These features can then in turn be used in the design the scoring rubrics by embedding them in the criteria at each score level. Experts need not be professionals in the field. Instead, in the design of K-12 assessments, experts are typically considered students who have attained competency within the content domain.

While there is the recognition that theories of cognition and learning should serve as the foundation for the design and interpretation of assessments, widespread use of cognitive models of learning in assessment design has not been realized. As summarized by Bennett and Gitomer (in press), there are three primary reasons for this: 1) the disciplines of psychometrics and of cognition and learning have developed separately are just beginning to merge, 2) theories of the nature of proficiency and learning progressions are not fully developed, and 3) there are both economic and practical constraints. There are promising assessment design efforts; however, that are taking advantage of what has been learned about the acquisition of student proficiency. A systematic approach to designing assessments that reflect theories of cognition and learning is embodied in Mislevy and his colleagues' (Mislevy, Steinberg, and Almond, 2003) evidence-centered design (ECD) in which evidence observed in student performances on

complex problem solving tasks (that have clearly articulated cognitive demands) is used to make inferences about student proficiency. Some of these design efforts will be discussed in this chapter.

Delineation of a Conceptual Framework for Design

A well designed performance assessments begins with the delineation of the conceptual framework. The extent to which the conceptual framework considers cognitive theories of student proficiency and is closely aligned to the relevant curriculum will affect the validity of score interpretations. The delineation of the conceptual framework includes a description of the construct to be assessed, the purpose of the assessment, and the intended inferences to be drawn from the assessment results (Lane & Stone, 2006). Construct theory as a guide to the development of an assessment provides a rationale basis for specifying features of assessment tasks and scoring rubrics as well as for expecting certain empirical evidence, such as the extent of homogeneity of item responses and the relationship between scores with other measures (Messick, 1994; Mislevy, 1996; National Research Council, 2001). Two general approaches to designing performance assessments has been proposed, a construct-centered approach and a task-centered approach (Messick, 1994). Under the construct-centered approach, the construct is specified by identifying the complex set of knowledge and skills that need to be assessed and are valued in instruction. The performances or responses that should be elicited by the assessment are then identified. By using this design approach, the construct guides the development of the tasks as well as the specification of the scoring criteria and rubrics.

By focusing on the construct in assessment design, the test designer can pay attention to both construct-irrelevant variance and construct underrepresentation which may have an impact on the validity of score inferences (Messick, 1994). Construct underrepresentation occurs when the assessment does not fully capture the targeted construct, and therefore the score inferences may not be generalizable to the larger domain of interest. Construct-irrelevant variance occurs when one or more irrelevant constructs is being assessed in addition to the intended construct. For example, students' writing ability may have an unwanted impact on performance on a mathematics assessment. The section on validity and fairness of assessment later in the chapter provides a discussion on construct-irrelevant variance and construct underrepresentation. Scores derived from a construct-centered approach may be more generalizable across variations in tasks, settings, and examinee groups than scores derived from a task-centered approach because of the attention to reducing construct-irrelevant variance and to increasing the representation of the construct (Messick, 1994).

For large-scale educational assessments, the conceptual framework is typically defined by content standards delineated at the state or national level. The grain at which the content standards are specified impacts on whether narrow bits of information will be assessed or whether broader, more contextualized

understanding of the content domain will be assessed. This is because the content standards guide the development of the test specifications that include the content, cognitive processes and skills, and psychometric characteristics of the tasks. Thus, the extent to which an assessment is valued will be dependent on the quality of the content standards.

Test specifications need to clearly articulate the cognitive demands of the tasks, problem solving skills and strategies that can be employed, and criteria to judge performance. This includes the specification of knowledge and strategies that are not only linked closely to the content domain, but also those that are content domain independent (Baker, 2007). Carefully crafted and detailed test specifications are even more important for performance assessments than multiple-choice tests because there are fewer performance tasks and typically each is designed to measure something that is relatively unique (Haertel & Linn, 1996). The use of detailed test specifications can also help ensure that the content of the assessment is comparable across years so as to allow for measuring change over time. The performance tasks and scoring rubrics are then developed iteratively based on a well-delineated conceptual framework and test specifications (Lane & Stone, 2006).

The use of conceptual frameworks in designing performance assessments leads to assessments that are linked to educational outcomes and provide meaningful information that can guide curriculum and instructional reform. As an example, a construct-centered approach was taken in the design of a mathematics performance assessment that required students to show their solution processes and explain their reasoning (Lane, 1993; Lane et al., 1995). The conceptual framework that Lane and her colleagues proposed then guided the design of the performance tasks and scoring rubrics. Cognitive theories of student mathematical proficiency provided a foundation for defining the construct domain of mathematics. Four components were specified for the assessment and task design and further delineated: cognitive processes, mathematical content, mode of representation and task context. To reflect the complex construct domain of mathematical problem solving, reasoning and communication, for example, a range of cognitive processes were specified including discerning mathematical relations, using and discovering strategies and heuristics, formulating conjectures, and evaluating the reasonableness of answers. Performance tasks were then developed to assess one or more of these skills.

Underlying performance assessments is a continuum that represents different degrees of structure versus open-endedness in the response (Messick, 1996). The degree of structure for the problem posed and the response expected should be considered in the design of performance assessments. Baxter and Glaser (1998) characterized performance assessments along two continuums with respect to their task demands. One continuum represents the task demand for cognitive processes ranging from open to constrained, and the other continuum represents the task demand for content knowledge from rich to lean.

A task is process open if it promotes opportunities for students to develop their own procedures and strategies, and a task is content rich if it requires substantial content knowledge for successful performance. These two continuums are crossed to form four quadrants so that tasks can be designed to fit one or more of these quadrants. This allows for clearly articulated cognitive and content targets in task design, and for the evaluation of tasks in terms of their alignment with these targets (Baxter & Glazer, 1998). In the design of performance assessments that assess complex cognitive thinking skills, design efforts can be aimed primarily in the quadrant that reflects tasks that are process open and content rich; however, familiarity with these types of tasks in instruction and the age of the student needs to be considered in design efforts. The two continuums (content knowledge and cognitive processes) could allow for more than four quadrants so as to examine students' progression in understanding within a content area.

Templates for Task Design. It is beneficial to develop templates for task design to ensure that the cognitive skills that are of interest are assessed. Templates can be developed for performance tasks that allow for tasks to be designed that assess the same cognitive processes and skills, and a scoring rubric can then be designed for the tasks that can be generated from a particular template. The use of templates for task design allows for an explicit delineation of the cognitive skills to be assessed, and can improve the generalizability of the score inferences. A model-based assessment approach that uses task templates has been proposed by Baker (2007). The major components of the model are the cognitive demands of the task, criteria to judge performance derived by competent performance, and a content map that describes the subject matter, including the interrelationships among concepts and the most salient features of the content. The cognitive demands of the tasks can then be represented in terms of families of tasks (or task templates) such as reasoning, problem solving, and knowledge representation tasks (Baker, 2007). As an example, the explanation task template asks students to read one or more texts that require some prior knowledge of the subject domain, including concepts, principles and declarative knowledge, in order to understand them, and to evaluate and explain important issues introduced in the text (Neimi, Baker, & Sylvester, 2007). A task from the explanation family that was developed for assessing student proficiency in Hawaii is provided below (Neimi, Baker, & Sylvester, 2007, p. 199).

> Imagine you are in a class that has been studying Hawaiian history. One of your friends, who is a new student in the class, has missed all the classes. Recently, your class began studying the Bayonet Constitution. Your friend is very interested in this topic and asks you to write an essay to explain everything that you have learned about it.
>
> Write an essay explaining the most important ideas you want your friend to understand. Include what you have already learned in class about Hawaiian history and what you have learned from the texts you have just read. While you write, think about

what Thurston and Liliuokalani said about the Bayonet Constitution, and what is shown in the other materials.

Your essay should be based on two major sources:

1. The general concepts and specific facts you know about Hawaiian history, and especially what you know about the period of Bayonet Constitution.
2. What you have learned from the readings yesterday.

Prior to receiving this task, students were required to read the primary source documents that were referred to in the prompt. This task requires students to not only make sense of the material from multiple sources, but to integrate material from these multiple sources in their explanations. This provides just one example of a task that can be generated from the explanation task template. Task templates can also be used to design computer-based simulation tasks.

Design of Computer-Based Simulation Tasks

Computer-based simulations have made it possible to assess complex thinking skills that cannot be measured well by more traditional assessment methods. Using extended, integrated tasks, a large problem-solving space with various levels of complexity can be provided in an assessment (Vendlinski, Baker, & Niemi, 2008). Computer-based simulation tasks can assess student competency in formulating and testing and evaluating hypotheses, selecting an appropriate solution strategy, and when necessary, adapting strategies based on the degree of success to solution. An attractive feature of computer-based simulation tasks is that they can include some form of immediate feedback to the student according to the course of actions taken by the student. Other important features of computer-based simulations include the variety of the types of interactions that a student has with tools in the problem solving space and the monitoring and recording of how student use these tools (Vendlinski et al., 2008). Technology used in computer-based simulations allow assessments to provide more meaningful information by capturing students' processes and strategies, as well as their products. Information on how a student arrived at an answer or conclusion can be valuable in guiding instruction and monitoring the progression of student learning (Bennett, Persky, Weiss, & Jenkins, 2007). The use of automated scoring procedures for evaluating student performances to computer-based simulations tasks addresses the cost and time demands of human scoring.

Several issues need to be considered in the design of computer-based simulations such as the examinee's familiarity with the navigation rules and controls imposed by the computer interface and testing network requirements, the potential requirement of examinees' to record their answers in an unusual manner, and the large amount of data that needs to be summarized in a meaningful way (Bennett, et al., 2007; DeVore, 2002). Like all assessments, computer-based tasks have the potential to measure factors which are irrelevant to the construct that is intended to be assessed, and therefore the validity of

the score interpretations can be hindered. It is important to ensure that the computer interface is one in which examinees are familiar with, and students have had the opportunity to practice with the computer interface and navigation system. It is also important to ensure that the range of cognitive skills and knowledge assessed are not narrowed to those that are more easily assessed using computer technology. Further, the automated scoring procedures need to reflect important features of proficiency so as to ensure that the generated scores provide accurate interpretations (Bennett, 2006; Bennett & Gitomer, in press). The use of test specifications that delineate the cognitive skills and knowledge that are intended to be assessed by the computer-based simulations will help ensure representation of the assessed content domain in both the tasks and scoring procedures so as to allow for valid score interpretations. Further, task templates can be used to ensure that the tasks and scoring rubrics embody the important cognitive demands.

The advancements of computer technology have made it possible to use performance-based simulations, which assess problem solving and reasoning skills in large scale, high-stakes assessment programs. The most prominent large-scale assessments that use computer-based simulations are licensure examinations in medicine, architecture and accountancy. As an example, computer-based case simulations have been designed to measure physicians' patient-management skills, providing a dynamic interaction simulation of the patient-care environment (Clyman, Melnick, & Clauser, 1995). In this assessment, the examinee is first presented with a description of the patient and then the examinee must manage the patient case by selecting history and physical examination options or by making entries into the patient's chart to request tests, treatments, and/or consultations. The condition of the patient changes in real time based on the patient's disease and the examinee's course of actions. The computer-based system generates a report that displays each action taken by the examinee and the time that the action was ordered. The examinee performance is then scored by a computer automated scoring system according to the appropriateness of the sequence of the ordered actions. It is apparent that this licensure examination captures some essential and relevant problem-solving, judgment and decision making skills that are required of physicians.

A research project that used the architecture computer-based exam demonstrated how the format of a task can affect the problem solving and reasoning skills that are used by examinees (Martinez &Katz, 1996). Differences in the cognitive skills assessed by computer-based figural responses items as compared to multiple-choice items in the architecture exam were observed. As an example, for one figural response item, a building site which is surrounded by icons that represent a parking lot, playground, and library are presented on the computer screen. The examinee is asked to select a tool, such as one that rotates or moves an icon, and through a series of mouse movements and clicks the

examinee is asked to arrange the icons to meet particular criteria. Other figural response items require students to draw lines or arrows or attach labels to parts of a diagram. The results of their study suggest that on items that required students to use their own strategies, the skills used to solve the tasks differed dependent on whether it was a figural response item or a multiple-choice item. For the figural response items, students devised a strategy, generated a response and evaluated it based on the criteria, whereas on the multiple-choice items students just examined each alternative with respect to the criteria. The cognitive demands of the item formats were clearly different, with the skills engaged by students on the figural response items being better aligned to the skills of interest than those used on the multiple-choice items.

Using evidence-centered design (Mislevy, Steinberg, and Almond, 2003), computer simulations tasks in the physics domain were developed in the context of a NAEP research project (Bennett, Persky, Weiss, & Jenkins, 2007). The goal of this project was to examine the feasibility of including computer-based simulations on the NAEP science assessment. The computer simulation tasks were designed to represent exploration features of real-world problem solving, and incorporated "what-if" tools that students used to uncover underlying scientific relationships. To assess scientific inquiry skills, students were required to design and conduct experiments, interpret results, and formulate conclusions. As part of the simulations, students needed to select values for independent variables and to make predictions as they designed their experiments. To interpret their results students needed to develop tables, graphs and formulate conclusions. In addition to these scientific inquiry tasks, tasks were developed to assess students' search capabilities on a computer. One 8th grade inquiry computer-based simulation task required students to investigate why scientists use helium gas balloons to explore out of space and the atmosphere (Bennett et al., 2007). An example of an item within this task that required students to search a simulated World Wide Web is provided below (Bennett et al., 2007, p. 41).

> Some scientists study space with large helium gas balloons. These balloons are usually launched from the ground into space but can also be launched from a spacecraft near other planets.
>
> Why do scientists use these gas balloons to explore outer space and the atmosphere instead of using satellites, rockets, or other tools? Be sure to explain at least three advantages of using gas balloons.
>
> Base your answer on more than one web page or site. Be sure to write your answer in your own words.

This task assesses students' research skills using a computer which is typical of what is expected in their instructional experiences. An example of a related scientific inquiry task that required students to evaluate

their work, form conclusions and provide rationales after designing and conducting a scientific investigation is provided below (Bennett et al., 2007, p. 46).

> How do different amounts of helium affect the altitude of a helium balloon?
>
> Support your answer with what you saw when you experimented.

These simulation tasks were based on models of student cognition and learning and allowed for the assessment of problem-solving, reasoning and evaluation skills that are valued within the scientific discipline. It should be noted that for the 2009 science NAEP, a sample of the students were administered these types of computer-based simulation tasks, requiring them to engage in the processes of scientific inquiry by working on a simulated experiment, recording data and critiquing a hypothesis (Sawchuk, 2009).

Computer-based simulation tasks in the reading, mathematics and writing domains are being designed and evaluated for their potential inclusion in an integrated accountability and formative assessment system (Bennett & Gitomer, in press; O'Reilly & Sheehan, in press). In the reading domain, a cognitive model of reading competency serves at the basis for both assessing learning and advancing learning. Three assessment design features that are aimed at assessing deeper processing by requiring students to actively construct meaning from text, and are based on a cognitive model of reading are described by O'Reilly and Sheehan (in press). First, in the assessment a scenario is provided that describes the purpose of reading. Because students engage in the reading process in meaningfully different ways dependent on the purpose of reading, the purpose of reading is clearly articulated. Second, students are required to read multiple texts so as to encourage students to integrate and synthesize information across texts. Lastly, to assess students' evaluation skills texts of varying quality are provided.

One of the four important components assessed in their reading competency model is the student's ability to extract discourse structure (the other three are understanding vocabulary, drawing necessary inferences, and identifying important details). As O'Reilly and Sheehan (in press) pointed out, requiring students to construct a lengthy written summary may be more appropriate in the assessment of writing and not reading since the quality of students' response to a reading task can be affected by their writing ability. Instead they use graphical representations for students to map out the structure of the text, including graphic hierarchical organizers and construct maps. The use of graphical representations instead of written summaries helps ensure that a student's writing ability does not unduly affect their performance on the reading tasks. This may help in minimizing construct-irrelevant variance. Further, the use of graphical representations will more easily allow for computer automated scoring procedures to be

used in the scoring of students' competency in organizing and summarizing information that they have read from one or more texts. This research program draws on models of cognition and learning and advances in technology and measurement to design assessments that capture students' complex thinking skills and thus has the capacity to provide meaningful information to guide instruction. As the researchers have indicated there are a number of things that are being addressed in the design of these computer-based simulation tasks so as to ensure the validity of the score interpretations. Response formats are being chosen to minimize the extent to which writing is affecting the scores on reading and mathematics tasks, and that allow for automated scoring. Also, careful attention is being paid to representing the content and cognitive skills across the tasks so as to ensure the validity of the score generalizations.

Design of Assessments that Measure Learning Progressions

Assessments that reflect learning progressions are capable of identifying where students are on the learning progression and the skills and knowledge they need to acquire to become more competent. There have been some recent advances in assessment design efforts that reflect learning progressions or sometimes referred to as construct maps. Learning progressions indicate what it means to acquire understanding within a construct domain, and they identify where a student is on the continuum of the underlying construct. More specifically, they have been defined as "descriptions of successively more sophisticated ways of reasoning within a content domain based on research syntheses and conceptual analyses" (Smith, Wiser, Anderson, & Krajcik, 2006, p. 1), and should be organized around central concepts or big ideas within a content domain. Empirically validated models of cognition and learning can be used to design assessments that monitor students' learning as they develop understanding and competency in the content domain. These models of student cognition and learning across grade levels can be reflected in a coherent set of content standards across grade levels. This will help ensure the continuity of the assessment of students across grades, and will allow for monitoring student understanding and competency and for informing instruction and learning.

An issue in the design of learning progressions is that there may be multiple paths to proficiency; however some paths typically are followed by students more often than others (Bennett & Gitomer, in press; NRC, 2006). The use of these common paths to define learning progressions and the ways in which students gain a deep understanding of the content domain can be used as the foundation for designing assessments that monitor student achievement and learning. Learning progressions that are based on cognitive models of learning and are supplemented by teacher knowledge of student learning within content domains can inform the design of assessments that will elicit evidence to support inferences about student achievement at different points along the learning progression (NRC, 2006).

Further, they have the potential to lead to more meaningful scaling of assessments that span grade levels, and thus more valid score interpretations regarding student growth.

Wilson and his colleagues have designed an assessment system that incorporates information from learning progressions and advances in both technology and measurement referred to as the BEAR Assessment System (Wilson, 2005; Wilson & Sloane, 2000). One application of this assessment system is for measuring student's progression for one of the three "big ideas" in the domain of chemistry, namely *matter* which is concerned with describing molecular and atomic views of matter (Wilson, 2005). The two other "big ideas are *change* and *stability*, the former is concerned with kinetic views of change and the conservation of matter during chemical change, and the latter is concerned with the system of relationships in conservation of energy. Figure 1 illustrates the construct map for the *matter* big idea for two of its substrands, visualizing and measuring.

Insert Figure 1

Level 1 in the table is the lowest level of proficiency and reflects students' lack of understanding of atomic views of matter, reflecting only their ability to describe some characteristics of matter, such as differentiating between a solid and a gas (Wilson, 2005). At level 2, students begin to use a definition or simple representation to interpret chemical phenomena, and at level 3 students begin to combine and relate patterns to account for chemical phenomena. Items are designed to reflect the differing achievement levels of the learning progression, or construct map, and empirical evidence is then collected to validate the construct map. A task designed to assess the lower levels of the construct map depicted in Figure 1 asks students to explain why two solutions with the same molecular formula have two very different smells. The task presents students with the two solutions, butyric acid and ethyl acetate, their common molecular formula, $C_4H_8O_4$, and a pictorial representation depicting that one smells good and the other bad. The students are required to respond in writing to the following prompt (Wilson, 2005, p. 11).

> Both of the solutions have the <u>same molecular formulas</u>, but <u>butyric acid smells bad and putrid while ethyl acetate smells good and sweet.</u> Explain why these two solutions smell differently.

By delineating the learning progressions within each of the "big ideas" of chemistry based on models of cognition and learning, assessments can be designed so as to provide evidence to support inferences about student competency at different achievement levels along the learning progressions. Performance assessments are well-suited for capturing student understanding and thinking along these learning progressions. Smith and her colleagues (Smith et al., 2006) proposed a learning progression around three key questions and six big ideas within the scientific topic of matter and atomic-molecular theory and

provided examples of performance tasks that can assess different points along the continuum of understanding and inquiry within this domain.

Learning progressions are also considered in the BioKIDS project that is based on the Principled Assessment Designs for Inquiry (PADI) system. Within this system, three main design patterns for assessing scientific inquiry were identified, including formulating scientific explanations from evidence, interpreting data, and making hypotheses and predictions (Gotwals & Songer, 2006). Tasks based on a specific design pattern have many features in common. As an example, the design pattern, Formulating Scientific Explanations from Evidence, has two dimensions that are crossed: the level of inquiry skill required for the task and the level of content knowledge required for the task. This allows for the design of assessment tasks in 9 cells, each cell representing a task template. There are three inquiry skill steps, from Step 1 to Step 3: "students match relevant evidence to a given claim, students choose a relevant claim and construct a simple explanation based on given evidence (construction is scaffolded), students construct a claim and explanation that justifies the claim using relevant evidence (construction is unscaffolded)" (Gotwals & Songer, 2006, p. 13). The level of content knowledge required for the task is classified as simple, moderate or complex, requiring minimal content knowledge and no interpretation to applying extra content knowledge and interpretation of evidence. This is similar to Baxter and Glaser's (1998) conceptualization of four quadrants that differ in terms of content-richness and level of inquiry skills, but further divides these two dimensions into 9 quadrants. To better reflect scientific inquiry, Gotwal and Songer (2006) have proposed a matrix for each of the three design patterns (formulating scientific explanations from evidence, interpreting data, and making hypotheses and predictions).

In designing performance tasks, the amount of scaffolding needs to be considered. Scaffolding is a task feature that is manipulated explicitly in the BioKIDS design patterns (Gotwals & Songer, 2006; Mislevy & Haertel, 2006). For example, Figure 2 presents two scientific inquiry assessment tasks that require scientific explanations from the BioKIDS project. The first task requires more scaffolding than the second task. The amount of scaffolding built into a task depends on the age of the students and the extent to which students have had the opportunity in instruction to solve tasks that require explanations and complex reasoning skills. The first task in the figure represents the second step in the level of inquiry skills and the second level of content knowledge (moderate) in that evidence is provided (pictures of invertebrates that must be grouped together based on their characteristics), but the student needs to choose a claim and construct the explanation, and they must interpret evidence and/or apply additional content knowledge (need to know what characteristics are relevant for classifying animals). The second task is at Step 3 of the level of inquiry skill and the third level of content knowledge (complex) in that the student needs to construct a claim and an explanation that requires the interpretation of evidence and application

of additional content knowledge (Gotwals & Songer, 2006). More specifically, in this second task, "Students are provided a scenario, and they must construct (rather than choose) a claim and then, using their knowledge of food web interactions, provide evidence to back up their claim" (Gotwals & Songer, p. 16).

Insert Figure 2

Additional Examples of Performance Tasks. This section provides additional examples of performance tasks that draw on cognitive theories of student thinking and learning. The National Assessment of Educational Progress (NAEP) has included hands-on performance tasks in their science assessment (U.S. Department of Education, 2005) so as to better measure complex problem solving and reasoning skills. These tasks require students to engage in scientific inquiry, and to record their observations and conclusions by answering both multiple-choice and constructed-response items. As an example, a public release 4th grade task, Floating Pencil, provides students with a set of materials, including bottles of freshwater, salt water and "mystery" water. Students are required to perform a series of investigations to determine the properties of salt and freshwater, and to determine whether the bottle of mystery water is salt water or freshwater. After responding to a number of questions throughout their investigation, the students are asked (U.S. Department of Education, 2005, p. 10):

> Is the mystery water fresh water or is it salt water?
> How can you tell what the mystery water is?
>
> When people are swimming, is it easier for them to stay afloat in the ocean or in a freshwater lake?
> Explain your answer.

The use of these hands-on performance tasks and constructed-response items allow NAEP to better assess scientific inquiry skills. As previously discussed, NAEP is currently examining the use of computer-based simulations to assess scientific inquiry.

The most commonly used large-scale performance assessments in this country are writing assessments. Writing assessments may consist of stand-alone writing prompts or text-based writing prompts. Stand-alone writing prompts require students to produce a written response to a given brief topic or prompt. Whereas, text-based writing prompts reflect the reading and writing connection, in that students are asked to read about a topic from one or more sources, analyze it from a particular perspective, and then write a response (Nelson & Calfee, 1998). It has been argued that text-based writing assessments are more aligned to the writing that occurs in most classrooms in grades 6 through

12, higher education and the workplace. An example of a writing assessment that includes both stand-alone and text-based writing prompts is the Delaware Student Testing Program (DSTP; Delaware Department of Education, 2000). A text-based writing task in the Delaware state assessment is linked to a passage in the reading assessment, and student responses to the task are scored twice, once for reading and once for writing. An example 8th grade text-based persuasive writing prompt from the DSTP which requires students to read an article prior to writing is provided below.

> The article you have just read describes some problems and possible solutions for dealing with grease. Do you think grease should be classified or labeled as a pollutant?
>
> Write a letter to the Environmental Protection Agency explaining whether or not grease should be classified as a pollutant. Use information from this article to support your position. (Delaware Department of Education, 2005, p. 5)

This task is aligned to the reading and writing connection that occurs in instruction in Delaware classrooms. Students are first asked to read about a topic and then to use the information that they have read to support their position in their written product.

Another example of an assessment program that reflected the reading and writing connection was the Maryland School Performance Assessment Program (MSPAP; Maryland State Board of Education, 1995). It consisted of integrated performance tasks in writing, reading, social studies and science. For example, a writing task may have required the student to read one or more text in the social studies or science domains, analyze the text from a particular perspective, and then write an essay. A complex integrated performance task from MSPAP that addressed issues related to child labor is provided in the appendix. This is an integrated task that assesses reading, writing, language usage and social studies. For this task students were required to read "A Letter to Hannah" and "Mill Children" (only the first two pages of the two texts are included), and were asked to respond to a series of questions based on these readings, some of the questions required students to integrate their understanding of the two texts. They were also provided with two maps and were asked to respond to questions based on their understanding of the texts, maps, and general social studies content knowledge. A persuasive writing prompt was also given in which students needed to use information from the texts to support their views on child labor. The writing portion of the task also required students to engage in the different stages of the writing process (prewriting, writing, review and editing, and final version) as well as peer review. As indicated previously, MSPAP was the only completely state performance-based assessment program in multiple content areas that sustained success over a number of years.

Review and Field-testing Performance Assessments

Performance assessments need to be appraised with regard to the quality and comprehensiveness of the content and processes being assessed and with regard to potential issues of bias in task content, language and context. The review process is an iterative process in that when tasks are developed they may be reviewed by experts and modified a number of times prior to and after being field-tested. This involves logical analyses of the tasks to help evaluate whether they are assessing the intended content and processes, worded clearly and concisely, and free from anticipated sources of bias. The development process also includes field testing the tasks and scoring rubrics to ensure they elicit the processes and skills intended.

It is important to field test items individually as well as in a large scale administration. For example, protocol analysis in which students are asked to think aloud while solving a task or to describe retrospectively the way in which they solved the task can be conducted to examine whether the intended cognitive processes are elicited by the task (Chi, Glaser, & Farr, 1988; Ericsson & Simon, 1984). These individual pilots afford rich information from a relatively small number of students regarding the degree to which the tasks evoke the content knowledge and complex thinking processes that they were intended to evoke, and allows for additional probing regarding the processes underlying student performance. The individual piloting of tasks also provides an opportunity for the examiner to pose questions to students regarding their understanding of task wording and directions, and to evaluate their appropriateness for different subgroups of students, such as students whose first language is not English.

A large-scale field testing provides additional information regarding the quality of the tasks including the psychometric characteristics of the items. Student work from constructed-response items or essays can also be analyzed to ensure that the tasks evoke the content knowledge and cognitive processes that they are intended to evoke, and the directions and wording are as clear as possible. Multiple variants of tasks can also be field-tested to further examine the best way to phrase and format tasks to ensure that all students have the opportunity to display their reasoning and thinking. Any one of these analyses may point to needed modifications to the tasks.

Large-scale field testing of performance tasks poses a risk to security because they tend to be memorable to students. To help ensure the security of performance assessments, some state assessment programs have field tested new tasks in other states. As an example, the initial field-testing of writing prompts for the Maryland Writing Test (MWT) occurred in states other than Maryland (Ferrara, 1987). The state's concern about the comparability of the out-of-state sample with respect to demographics, motivation and writing instruction however led them to an in-state field test design. While security issues such as students sharing the field-test prompts with other students were considered problematic, the improvement of the field test data outweighed security concerns (Ferrara, 1987). For example, in 1988,

22 new prompts were field tested on a sample of representative 9th grade students in Maryland with each student receiving 2 prompts. The anchor prompts were spiraled with the field test prompts in the classrooms and each prompt was exposed only to approximately 250 students (Maryland State Department of Education, 1990). Field test prompts that were comparable to the anchor prompts (e.g., similar means and standard deviations) were selected for future operational administrations (Ferrara, personal communication, July 30, 2009) and sophisticated equating procedures were not used. Enough prompts produced similar mean scores and standard deviations so as to be considered interchangeable.

To help maintain the security of the Maryland Writing Test prompts a number of procedures were implemented (Ferrara, personal communication, July 30, 2009). First, the number of students who were exposed to any one prompt was small (approximately 250), and the number of teachers involved in the field test was relatively small. Second, the prompts were field tested two to three years before they were administered on an operational test. Third, there was rigorous enforcement of security regulations.

For the field testing of essay topics for the new SAT, a number of steps are implemented to help ensure the security of the prompts (Educational Testing Service, 2004). First, approximately 78 topics are pretested each year to a representative sample of junior and seniors in high-schools across the country. No more than 175 students are involved in the field test in a participating school, and only 3 prompts are administered in any one school with each student receiving only 1 prompt. Each prompt is field tested in approximately 6 schools so that only 300 students are administered a given prompt. Second, prompts are field tested at least 2 years prior to being on an operational form of the SAT. Third, rigorous security procedures are used for shipping and returning the field-test prompts. Lastly, several security procedures are implemented during the pretest readings such as the requirement of signed confidentiality statements by all prescreened readers who have served on College Board writing committee.

Security issues need to be considered for assessment programs for which the intent is to generalize from the score to the broader content domain. If security is breeched and the assessment tasks are known prior to the administration of the assessment, some scores will be artificially inflated which will have an impact on the validity of the score interpretations. Prior exposure to the task is not a security issue for performance demonstrations such as a high school project that requires students to demonstrate competency within a discipline. However, other issues need to be considered for performance demonstrations, such as ensuring the demonstration reflects the examinee's work and not others unless a specified amount of collaboration was permitted.

Scoring Performance Assessments

As in the design of performance tasks, the design of scoring rubrics is an iterative process and involves coordination across grades as well as across content domains to ensure a cohesive approach to

student assessment (Lane & Stone, 2006). Much has been learned about the design of quality scoring rubrics for performance assessments. First, it is critical to design scoring rubrics that include criteria that are aligned to the processes and skills that are intended to be measured by the assessment tasks. Unfortunately, it is not uncommon for performance assessments to be accompanied by scoring rubrics that focus on lower levels of thinking rather than the more complex reasoning and thinking skills that the tasks are intended to measure, and therefore the benefits of the performance tasks are not fully realized. Typically, scoring rubrics should not be developed to be unique to specific tasks nor generic to the entire construct domain, but should be reflective of the "classes of tasks that the construct empirically generalizes or transfers to" (Messick, 1994, p. 17). Thus, a scoring rubric can be designed for a family of tasks or a particular task template. As previously indicated, underlying performance on a task is a continuum that represents different degrees of structure versus open-endedness in the response, and this needs to be considered in the design of the scoring rubric and criteria (Messick, 1996).

The design of scoring rubrics requires the specification of the criteria for judging the quality of performances, the choice of a scoring procedure (e.g., analytic or holistic), ways for developing criteria, and procedures used to apply the criteria (Clauser, 2000). The ways for developing criteria include the process used for specifying the criteria and who should be involved in developing the criteria. For large-scale assessments in K-12 education, typically, the scoring criteria are developed by a group of experts as defined by their knowledge of the content domain and experience as educators. Often these experienced educators have been involved in the design of the performance tasks and have knowledge of how students of differing levels of proficiency would perform on the task. There are alternative approaches to specifying the criteria such as analyses of experts' thinking and reasoning when solving tasks. Cognitive task analysis using experts' talk alouds (Ericsson & Smith, 1991) has been used to design performance tasks and scoring criteria in the medical domain (Mislevy, Steinberg, Breyer, Almond, & Johnson, 1999). Features of expert's thinking, knowledge, procedures, and problem posing are considered to be indicators of developing expertise in a domain (Glaser, Lesgold, & Lajoie, 1987), and can be used systematically in the design of assessment tasks and scoring criteria. As mentioned previously, these experts can be students who have demonstrated competency within the domain. Two ways in which the criteria can be applied rely on the use of trained raters and computer-automated scoring procedures (Clauser, 2000). This section discusses the specification of the criteria, different scoring procedures, research on scoring procedures, and computer-automated scoring systems.

Specification of the Criteria

The criteria specified at each score level should be linked to the construct being assessed, and depend on a number of factors including the cognitive demands of the tasks in the assessment, the degree of

structure or openness expected in the response, the examinee population, the purpose of the assessment, and its intended score interpretations (Lane & Stone, 2006). Further, the number of scores each performance assessment yields needs to be considered based on how many dimensions are being assessed. Performance assessments are well suited for measuring multiple dimensions within a content domain. For example, a grade 5 mathematics assessment may be designed to yield information on students' strategic knowledge, mathematical communication skills, and computational fluency. Separate criteria would be defined for each of these dimensions and a scoring rubric would then be developed for each dimension.

The number of score levels used depends on the extent to which the criteria across the score levels can distinguish among various levels of knowledge and skills. The knowledge and skills reflected at each score level should differ distinctly from those at other score levels. When cognitive theories of learning have been delineated within a domain, the learning progression can be reflected in the criteria. The criteria specified at each score level are then guided by knowledge of how students acquire understanding and competency within a content domain.

A generic rubric may be designed that reflects the skills and knowledge underlying the defined construct. The development of the generic rubric begins in the early stages of the performance assessment design, and then guides the design of specific rubrics for each family of tasks (task template) or a particular task that capture the cognitive skills and content assessed by the family of tasks or the particular task. An advantage of this approach is that it helps ensure consistency across the specific rubrics and is aligned with a construct-centered approach to test design. Typically, student responses that cover a wide range of competency are then evaluated to determine the extent to which the criteria reflect the components displayed in the student work. The criteria for the generic and/or specific rubrics may then be modified, and/or the task may be redesigned to ensure it assesses the intended content knowledge and processes. This may require several iterations to ensure the linkage among the content domain, tasks, and rubrics.

Scoring Procedures

The design of scoring rubrics has been influenced considerably by efforts in the assessment of writing. There are three major types of scoring procedures for direct writing assessments, holistic, analytic, and primary trait scoring (Huot, 1990; Miller & Crocker, 1990; Mullis, 1984). The choice of a scoring procedure depends on the defined construct, purpose of the assessment, and nature of the intended score interpretations. With holistic scoring, the raters make a single, holistic judgment regarding the quality of the writing and assign one score, using a scoring rubric with criteria and benchmark papers anchored at each score level. With analytic scoring, the rater evaluates the writing according to a number

of features, such as content, organization, mechanics, focus, and ideas, and assigns a score indicating level of quality to each one. Some analytic scoring methods weight the domains, allowing for domains that are assumed to be more pertinent to the construct being measured, such as content and organization, to contribute more to the overall score. As summarized by Mullis (1984), "…holistic scoring is designed to describe the overall effect of characteristics working in concert, or the sum of the parts, analytic scoring is designed to describe individual characteristics or parts and total them in a meaningful way to arrive at an overall score" (p. 18). Although the sum of the parts of writing may not be the same as an overall holistic judgment, the analytic method has the potential to provide information regarding potential strengths and weaknesses of the examinee. Evidence, however, is needed to determine the extent to which the domain scores are able to differentiate aspects of students' writing ability.

Primary trait scoring was developed by the National Assessment of Educational Progress (NAEP; Lloyd-Jones, 1977). The primary trait scoring system is based on the premise that most writing is addressed to an audience with a particular purpose, and levels of success in accomplishing that purpose can be defined concretely (Mullis, 1984). As an example, three common purposes of writing are informational, persuasive, and literary. The specific task determines the exact scoring criteria, although criteria are similar across similar kinds of writing (Mullis, 1984). The design of a primary trait scoring system involves the identification of one or more traits relevant for a specific writing task. For example, features selected for persuasive writing may include clarity of position and support, whereas characteristics for a literary piece may include plot, sequence, and character development. Thus, the primary trait scoring system reflects aspects of a generic rubric as well as task-specific rubrics. By first using a construct-centered approach, the construct, and in this case the type of writing, guides the design of the scoring rubrics and criteria. The development of primary trait rubrics then allows for the general criteria to be tailored to the task allowing for more consistency in raters' application of the criteria to the written response. Thus, in the end there may be one scoring rubric for each writing purpose. This would be analogous of having one scoring rubric for each family of tasks or task template.

Examples of Scoring Rubrics. In the design of performance assessments, Baker and her colleagues (Baker, 2007; Neimi, Baker, and Sylvester, 2007) have represented the cognitive demands of the tasks in terms of classes or families of tasks such as reasoning, problem solving, and knowledge representation tasks (Baker, 2007). To ensure a coherent link between the tasks and the score inferences, they have designed a scoring rubric for each of these families of tasks. In the adoption of a construct-driven approach to the design of a mathematics performance assessment, Lane and her colleagues (Lane, 1993; Lane et al., 1995) used this approach to the design of their holistic scoring rubric. They first developed a generic rubric, as shown in Figure 3 that reflects the conceptual framework used in the design of the

assessment, including mathematical knowledge, strategic knowledge and communication (i.e., explanations) as overarching features. These features guided the design of families of tasks: tasks that assessed strategic knowledge, tasks that assessed reasoning, and tasks that assessed both strategic knowledge and reasoning. Mathematical knowledge was assessed across these task families. The generic rubric guided the design of each task specific rubric that reflected one of these three families. The use of task specific rubrics helped ensure the consistency in which raters applied the scoring rubric and the generalizability of the score inferences to the broader construct domain of mathematics.

Insert Figure 3

Another important issue in the design of scoring rubrics is that each of the score levels addressed each of the important scoring criteria. As can be seen in Figure 3, at each of the score levels criteria are specified for each of the overarching features: mathematical knowledge, strategic knowledge and communication.

The scoring rubric for the tasks that assess student learning in the matter strand in the chemistry domain (Wilson, 2005) discussed previously is presented in Figure 4. The scoring rubric is reflective of the construct map, or learning progression, depicted in Figure 1, with students progressing from the lowest level of describe to the highest level of explain. Score levels 1 (describe) and 2 (represent) in the rubric further differentiate students into 3 levels.

Insert Figure 4

A constructed-response that reflects level 2 (represent) is (Wilson, 2005, p. 16):

> They smell differently b/c even though they have the <u>same molecular formula</u>, they have <u>different structural formulas</u> with different arrangements and patterns.

This example response is at the level 2 because it "appropriately cites the principle that molecules with the same formula can have different arrangements of atoms. But the answer stops short of examining structure-property relationships (a relational, level 3 characteristic)" (Wilson, 2005, p. 16). A major goal of the assessment system is to be able to estimate, with a certain level of probability, where a student is on the construct map or learning progression. Students and items are located on the same construct map which allows for student proficiency to have substantive interpretation in terms of what the student knows and can do (Wilson, 2005). The maps can then be used to monitor the progress of an individual student as well as groups of students. Thus, valid interpretations of a student's learning or progression require a carefully designed assessment system that has well-conceived items and scoring rubrics that represent the various levels of the construct continuum as well as the empirical validation of the construct map, or learning progression. As previously indicated, students do not necessarily follow the same progression in

becoming proficient within a subject domain. Consequently, in the design of assessments, considerations should be given to identifying the range of strategies used for solving problems in a content domain, with an emphasis on those strategies that are more typical of the student population (Wilson, 2005). This assessment design effort provides an interesting example of the integration of models of cognition and learning, and of measurement models in the design of an assessment system that can monitor student learning and inform instruction. Further, a measurement model called the saltus (Latin for leap) model developed by Wilson (1989) can incorporate developmental changes (or conceptual shifts in understanding) as well as the incremental increases in skill in evaluating student achievement and monitoring student learning.

In an effort to assess complex science reasoning in middle and high school, a systematic assessment design procedure was adopted by Liu, Lee, Hofstetter, and Linn (2008). First, they identified an important construct within scientific inquiry, *science knowledge integration*. A comprehensive, integrated system of inquiry-based science curriculum modules, assessment tasks and scoring rubric were then developed to assess *science knowledge integration*. A scoring rubric was designed so that the different levels captured qualitatively different kinds of scientific cognition and reasoning that focused on elaborated links rather than individual concepts. Their assessment design is similar to the modeling of construct maps, or stages in learning progressions, described by Wilson and his colleagues (Wilson, 2005; Wilson & Sloane, 2000). The knowledge integration scoring rubric is shown in Figure 5.

Insert Figure 5

The rubric is applied to all the tasks that represent the task template for *science knowledge integration*, allowing for score comparisons across different items (Liu et al., 2008). As they indicate, having one scoring rubric that can be applied to the set of items that measure knowledge integration makes it more accessible for teachers to use and provides coherency in the score interpretations. The authors also provided validity evidence for the learning progression reflected in the scoring rubric.

Research on Analytic and Holistic Scoring Procedures

The validity of score interpretation and use depends on the fidelity between the constructs being measured and the derived scores (Messick, 1989). Validation of the scoring rubrics includes an evaluation of the match between the rubric and the targeted construct or content domain, how well the criteria at each score level captures the defined construct and the extent to which the domains specified in analytic scoring schemes each measure some unique aspect of student cognition. Lane and Stone (2006) provide a brief summary of the relative advantages of both analytic and holistic scoring procedures for writing assessments. As an example, Roid (1994) used Oregon's direct-writing assessment to evaluate its analytic scoring rubric in which students' essays were score on six dimensions. The results suggested

that each dimension may not be unique, in that relative strengths and weaknesses for some students were identified for combinations of dimensions. Thus, some of the dimensions could be combined in the scoring system without much loss of information while simplifying the rubric and the scoring process. Other researchers have suggested that analytic and holistic scoring methods for writing assessments may not necessarily provide the same relative standings for examinees. Vacc (1989) reported correlations between the two scoring methods ranging from .56 to .81 for elementary school students' essays. Research that has examined factors that affect rater judgment of writing quality have shown that holistic scores for writing assessments are influenced most by the organization of the text and important ideas or content rather than domains related to mechanics and sentence structure (Breland & Jones, 1982; Huot, 1990; Welch & Harris; 1994). Breland and colleagues (Breland, Danos, Kahn, Kubota, & Bonner, 1994) reported relatively high correlations between holistic scores and scores for overall organization (approximately .73), supporting ideas (approximately .70), and noteworthy ideas (approximately .68).

In the science domain, Klein et al. (1998) compared analytic and holistic scoring of hands-on science performance tasks for grades 5, 8, and 10. The correlations between the total scores obtained for the two scoring methods were relatively high, .71 for grade 5 and .80 for grade 8. The correlations increased to .90 for grade 5 and .96 for grade 8 when disattenuated for the inconsistency among raters within a scoring method. The authors suggested that the scoring method has little unique influence on the raters' assessment of the relative quality of a student's performance. They further suggested that if school performance is of interest, the use of one scoring method over the other probably has little or no effect on a school's relative standing within a state given the relatively high values of the disattenuated correlations. The time and cost for scoring for both of the methods was also addressed. The analytic method took nearly three times as long as the holistic method to score for a grade 5 response and nearly five times as long to score for a grade 8 response, resulting in higher costs for scoring using the analytic method.

The results of these studies suggest that the impact of the choice of scoring method (e.g., analytic versus holistic) may vary depending on the similarity of the criteria reflected in the scoring methods and for the use of the scores. The more closely the criteria for the analytic method resemble the criteria delineated in the holistic method, the more likely it is that the relative standings for examinees will be similar. The research also suggests that analytic rubrics typically are capable of providing distinct information for only a small number of domains or dimensions (i.e., two or three), and thus providing scores for a small number of domains has the potential for identifying overall strength and weaknesses in student achievement and for informing instruction. As previously suggested, scores derived from

performances on computer-based simulation tasks also allow for addressing different aspects of students' thinking.

Human Scoring

The scoring of student responses to performance assessments may be done by human scorers or automated scoring procedures that have been trained by human scoring. Lane and Stone (2006) provide an overview of the training procedures and methods for human scorers, and discuss rating sessions that may involve raters spending several days together evaluating student work as well as on-line rating of student work. A consideration in human scoring of performance assessments is rater variability or inconsistency, in particular, with writing assessments. As summarized by Eckes (2008), raters may differ in the extent to which they implement the scoring rubric, the way in which they interpret the scoring criteria, and the extent to which they are severe or lenient in scoring examinee performance; as well as in their understanding and use of scoring categories, and their consistency in rating across examinees, scoring criteria, and tasks (Bachman & Palmer, 1996; McNamara, 1996; Lumley, 2005). Thus, construct representation of the assessment can be jeopardized by the raters' interpretation and implementation of the scoring rubric as well as by features specific to the training session. For example, the pace at which raters are expected to score student responses may affect their ability to use their unique capabilities to accurately evaluate student responses or products (Bejar, Williamson, & Mislevy, 2006). Carefully designed scoring rubrics and training procedures however can help alleviate errors in human scoring.

Freedman and Calfee (1983) pointed out the importance of understanding rater cognition and proposed a model of rater cognition for evaluating writing assessments that consisted of three processes, reading the students' text to build a text image, evaluating the text image, and articulating the evaluation. Wolfe (1997) elaborated on Freedman and Calfee's model of rater cognition and proposed a cognitive model of rater cognition for scoring essays which included a framework for scoring and a framework for writing. He proposed that understanding the process of rating would allow for better design of scoring rubrics and training procedures. The framework of scoring is a "mental representation of the processes through which a text image is created, compared to the scoring criteria, and used as the basis for generating a scoring decision" (Wolfe, 1997, p. 89). The framework for writing, which includes the rater's interpretation of the criteria in the scoring rubrics, emphasized that raters have different interpretations of the scoring rubric and therefore are not equally proficient at rating student essays (Wolfe, 1997). Through training however raters begin to share a common understanding of the scoring rubric so as to apply it consistently. Wolfe (1997) also observed that proficient scorers were better able to withhold judgment as they read an essay and focused their efforts more on the evaluation process than less proficient scorers. This shared common framework for writing and high level of scoring proficiency

can lead to a high level of agreement among raters, and has implications for the training of raters (Wolfe, 1997). Thus, raters can be trained to internalize the criteria in a similar manner and to apply it consistently so as to ensure scores that allow for valid interpretations of student achievement.

Automated Scoring Systems

Automated scoring systems have supported the use of computer-based performance assessments such as computer-delivered writing assessments and computer-based simulation tasks, as well as paper and pencil assessments that are scanned. Automated scoring procedures have a number of attractive features. They apply the scoring rubric consistently, but more importantly they allow for the test designer to control precisely the meaning of scores (Powers, Burstein, Chodorow, Fowles, & Kukich, 2002). In order to accomplish this, they "need to elicit the full range of evidence called for by an appropriately broad definition of the construct of interest" (Bejar, Williamson, & Mislevy, 2006, p 52.). Automated scoring procedures allow for an expanded capacity to collect and record many features of student performance from complex assessment tasks that can measure multiple dimensions (Williamson, Bejar, & Mislevy, 2006). A very practical advantage is that they allow for scores to be generated in a timely manner. Automated scoring is defined by Williamson, Bejar and Mislevy (2006) as "any computerized mechanism that evaluates qualities of performances or work products" (p. 2). Automated scoring of complex constructed-response computerized tasks has proven effective been for large-scale assessments as well as for classroom assessment purposes. Project essay grader (1994, 2003) developed by Page in the 1960's paved the way for automated scoring systems for writing assessments, including e-rater (Burstein, 2003), Intelligent Essay Assessor (IEA; Landauer, Foltz, & Laham, 1998; Landauer, Laham, & Folz, 2003), and Intellemetric (Elliot, 2003).

Automated scoring procedures have also been developed to score short constructed-response items. C-rater is an automated scoring method for scoring constructed-response items that elicit verbal responses that range from one sentence to a few paragraphs, have rubrics that explicitly specify the content required in the response, but do not evaluate the mechanics of writing items (Leacock & Chodorow, 2003, 2004). It has been used successfully in Indiana's state end-of-course grade 11 English assessment, the NAEP Math Online project that required students to provide explanations of their mathematical reasoning, and the NAEP simulation study that required students to use search queries (Bennett et al., 2007; Dean, 2006). C-rater is a paraphrase recognizer in that it can determine when a student constructed response matches phrases in the scoring rubric regardless of their similarity in word use or grammatical structure (Dean, 2006). In the NAEP study that used physics computer-based simulations, c-rater models were built using student queries and then cross-validated using a sample of queries that were independently hand-scored. The agreement between human raters and c-rater for the cross-validation study was 96%.

Automated scoring procedures have also been developed and used successfully for licensure examinations in medicine, architecture and accountancy. These exams use innovative computer-based simulation tasks which naturally lend themselves to automated scoring. The previously mentioned assessment that uses computer-based case simulations to measure physicians' patient-management skills (Clyman, Melnick, & Clauser, 1995) and the figural response items in the architecture assessment (Martinez & Katz, 1996) are excellent examples of the feasibility in using automated scoring procedures for innovative item types.

Scoring Algorithms for Writing Assessments. The most widely used automated scoring systems are those that assess students' writing. Typically, the design of the scoring algorithms for automated scoring essay systems requires humans to first rate a set of student essays to a prompt. The student essays and their ratings then serve as calibration data that are used by the software to train it for scoring. The scoring algorithm is designed to analyze specific features of essays and weights are assigned for each of these features. As summarized by Deane and Gurevich (2008), the fields of computational linguistics, artificial intelligence, and natural language processing have produced a number of methods for investigating the similarity of text content, including latent semantic analysis (LSA) and content vector analysis (CVA). These text similarity methods have been applied to automated essay scoring applications. As an example, E-rater, developed by the Educational Testing Service, uses natural language processing techniques and identifies linguistic features of text in the evaluation of the quality of an essay (Burstein, 2003; Attali & Burstein, 2006). The first version of e-rater used over 60 features in the scoring process, whereas the latter versions use only "a small set of meaningful and intuitive features" (Attali & Burstein, 2006, p. 7) that better captures the qualities of good writing, and thus simplifying the scoring algorithm. The scoring system uses a model building module to analyze a sample of student essays to determine the weight of the features for assigning scores.

Evaluation of Automated Scoring Procedures. As with any assessment procedure, validation studies are imperative for automated scoring systems so as to provide evidence for appropriate score interpretations. Yang, Buckendahl, Juszkiewicz, and Bhola (2002) identified three categories of validation approaches for automated scoring procedures, including approaches focusing on the relationship among scores given by different scorers (human and computer), approaches focusing on the relationship between test scores and external measures of the construct being assessed, and approaches focusing on the scoring process. Most studies have examined the relationship between human and computer generated scores, typically indicating that the relationship between the scores produced by computer and humans is similar to the relationship between the scores produced by two humans, indicating the potential interchangeability of human and automated scoring. There have been few studies

however that focus on the latter two categories. In particular, validation studies focusing on the scoring process for automated scoring procedures are limited. As Bennett (2006) has argued, automated scoring procedures should be grounded in a theory of domain proficiency, using experts to delineate proficiency in a domain rather than having them as a criterion to be predicted. Both construct irrelevant variance and construct underrepresentation may affect the validity of the scores obtained by automated scoring systems (Powers, Burstein, Chodorow, Fowles & Kukich, 2001). With respect to construct irrelevant variance, automated scoring procedures may be influenced by irrelevant features of students' writing and assign a higher or lower score than deserved. In addition, they may not fully represent the construct of good writing which can affect the score assigned (Powers et al., 2001).

Studies have been conducted that require experts to evaluate the relevance of the computer-generated features of the target construct, identify extraneous and missing features, and evaluate the appropriateness of the weights assigned to the features (Ben-Simon & Bennett, 2007). Ben-Simon and Bennett (2007) found that the dimensions that experts in writing believe are most important in the assessment of writing are not necessarily the same as those obtained by automated scoring procedures that statistically optimize weights of the dimensions. As an example, experts indicated that approximately 65% of the essay scores in the study should be based on organization, development and topical analysis, while empirical weights gave approximately 21% of the emphasis to these dimensions. The opposite pattern occurred for the dimensions related to grammar, usage, mechanics, style and essay length, with a much lower emphasis assigned by experts and a higher emphasis given by the automated scoring procedure. As indicated by Ben-Simon and Bennett (2007), the parameters of automated scoring procedures can be adjusted to be more consistent with those that experts believe are features of good writing; however these adjustments may not be based on the criteria specified in the scoring rubric implemented in the study but rather the criteria used by the scorers in assigning scores. The authors indicated that the rubric employed in their study was missing key features of good writing, leaving experts to apply some of their own criteria in the scoring process. This result illustrates the importance of linking the cognitive demands of the tasks to the criteria specified in the scoring rubric regardless if responses are to be scored by human raters or automated scoring procedures. The authors further suggested that current theories of writing cognition should be used in assessment design so as to ensure that a more theoretical, coherent model for identifying scoring dimensions and features is reflected in the criteria of the rubrics.

Typically, the agreement between the scores that are assigned by human raters and those assigned by the automated scoring procedure is very high. There is some recent research, however, that indicates scores assigned by human raters and by automatic scoring procedures may differ to some extent depending upon student demographics. Bridgeman, Trapani and Attali (2009) examined whether there

were systematic differences in the performance of subgroups using an automated scoring procedure versus human scoring for an eleventh grade English state assessment. The prompt required students to support an opinion on a proposed topic within a 45 minute class period. The essays were scored holistically using a 6 point scale. The results indicated that on average, both Asian American and Hispanic students received higher scores from the automated scoring procedure than from human raters, whereas African American students scored similarly across the two scoring methods. Under the assumption that Asian American and Hispanic subgroups have a higher proportion of students with English as a second language, the authors suggested that this finding may not be due to minority status, but instead it may be related to having English as a second language. This may be reasonable given that the African American subgroup performed similarly across the two scoring methods. In their conclusions, they suggest that "although we treat human scores at the gold standard, we are reluctant to label discrepancies from the human score as bias because it is not necessarily the case that the human score is a better indicator of writing ability than the e-rater score (Bennett & Behar, 1997)" (Bridgeman, Trapani & Attali, 2009, p. 17). As suggested by the authors, additional research needs to examine features that contribute to differential subgroup results for human and automated scores, especially for students for which English is a second language. An understanding of the features of automated scoring systems that led to differential subgroup patterns will inform future designs of these systems.

Automated scoring systems need to be capable of flagging bad faith essays because of the possibility of examinees trying to trick the systems into providing scores that are not warranted. Advances in the design of the more recent versions of automated scoring systems have led to accurate identification of bad faith essays. Bad faith essays include essays that are off topic and are written to a different prompt, essays that repeat the prompt, essays that consist of multiple repeated text, and essays that are a mix of a genuine response and a repetition of the prompt. Studies have been conducted that demonstrate the capability of automated scoring procedures in detecting bad faith essays. In a early study, Powers and his colleagues (2001) examined the extent to which an early version of e-rater could be tricked into assigning either too high or too low of a score. Writing experts were asked to fabricate essays in response to the writing prompts in the Graduate Record Examination (GRE) that would trick e-rater into assigning scores that were either higher or lower than deserved. The writing experts were instructed on how e-rater scores student essays, and were asked to write one essay for which e-rater would score higher than human readers and one essay for which e-rater would score lower than human readers. E-rater scores on these fabricated essays were then compared with the scores of two human readers. The predictions that e-rater would score higher than the human readers were upheld for 87% of the cases. Some of the essays that were scored higher by e-rater as compared with the human raters consisted of repeated paragraphs with or

without a rewording of the first sentence in each paragraph. E-rater also provided higher scores than human readers for essays that did not provide a critical analysis, but focused on the features that e-rater attends to such as relevant content words and complex sentence structures. An important result is that only 42% of the cases were upheld when the predictions were that e-rater would score lower than the human raters (Powers et al., 2001). Thus, the experts were less able to trick e-rater to provide a lower score than human raters. It should be noted that e-rater has been revised substantially and there have been numerous versions of e-rater since this study. Further, to detect off-topic essays, which may occur when students are trying to fool the system, a content vector analysis program is used along with the more recent versions of e-rater (Higgins, Burstein & Attali, 2006).

In an evaluation of IntellicMetric for use with the Graduate Management Test (GMAT), Rudner, Garcia, and Welch (2006) examined its ability to detect common cheating techniques. For three Analysis of an Issue prompts and three Analysis of an Argument prompts, approximately 13 essays for each prompt were fabricated, resulting in 78 fabricated essays. The fabricated essays were evaluated with 500 validation essays for each prompt. Five of the fabricated essays were off-topic and written to a response to a different prompt but of the same type (Issues or Arguments prompt type), five essays were off-topic and written to a response to a different prompt of a different type, one essay was a repetition of the entire prompt, one essay consisted of multiple repeated text, and one essay consisted of half a genuine response and have a repetition of the prompt. Their results indicated that the system was successful at identifying fabricated essays that were a copy of the prompt, consisted of multiple repeated text, and consisted of the prompt and partly a genuine response. For each detected essay, the system provided specific warning flags for plagiarism, copying the prompt and nonsensical writing. The system was not successful at detecting off-topic responses; however, as the authors indicated this version of the system did not include a routine to flag off-topic essays.

The current versions of automated scoring systems for essays have shown strength in not only having high rates of agreement with human raters in assigning scores, but also in detecting bad faith essays. Automated scoring procedures for computerized short constructed-response items and innovative item types have also been used effectively for large-scale assessment programs. Further, various features of students' performances can be captured with automated scoring procedures which is ideal for computerized innovative tasks that reflect multiple dimensions within a content domain. Typically, most of the work and costs in designing automated scoring systems occur prior to the operational administration of the assessments, allowing for timely scoring and reporting of the results.

Evaluating the Validity and Fairness of Performance Assessment

Assessments are used in conjunction with other information to make important inferences about proficiency at the student, school, and state level, and therefore it is essential to obtain evidence about the appropriateness of those inferences and any resulting decisions. In evaluating the worth and quality of any assessment, including performance assessments, evidence to support the validity of the score inferences is at the forefront. Validity pertains to the meaningfulness, appropriateness, and usefulness of test scores (Kane, 2006; Messick, 1989). The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) state that "validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). This requires specifying the purposes and the uses of the assessment, designing the assessment to fit these intentions, and providing evidence to support the proposed uses of the assessment and the intended score inferences. As an example, if the purpose of a performance assessment is to assess complex thinking skills so as to make inferences about students' problem solving and reasoning one of the important validity studies would be to examine the cognitive skills and processes underlying task performance for support of those intended score inferences. The alignment between the content knowledge and cognitive skills underlying task responses and those underlying the targeted construct domain needs to be made explicit because typically the goal is to generalize assessment score interpretations to the broader construct domain (Messick, 1989). Fundamental to the validation of test use and score interpretation is also the evaluation of both intended and unintended consequences of the use of an assessment (Kane, 2006; Messick, 1989). Because performance assessments are intended to improve teaching and student learning it is essential to obtain evidence of such positive consequences as well as any evidence of negative consequences (Messick, 1994).

As previously discussed, there are two sources of potential threats to the validity of score inferences – construct underrepresentation and construct irrelevant variance (Messick, 1989). Construct underrepresentation occurs when the assessment does not fully capture the targeted construct, and therefore the score inferences may not be generalizable to the larger domain of interest. Issues related to whether the content of the assessment is representative of the targeted domain will be discussed later in this section. Construct-irrelevant variance occurs when one or more irrelevant constructs is being assessed in addition to the intended construct. Sources of construct irrelevant variance for performance assessments may include, but are not limited to, task wording and context, response mode, and raters' attention to irrelevant features of responses or performances. As an example, in designing a performance assessment that measures students' mathematical problem solving and reasoning tasks should be set in contexts that are familiar to the population of students. If one or more subgroups of students are unfamiliar with a particular problem context and it affects their performance, the validity and fairness of

the score interpretations for those students is hindered. Similarly, if a mathematics performance assessment requires a high level of reading ability and students who have very similar mathematical proficiency perform differently due to differences in their reading ability, the assessment is measuring in part a construct that is not the target, namely, reading proficiency. This is of particular concern for English Language Learners (ELLs). Abedi and his colleagues (Abedi, Lord, & Plummer, 1997; Abedi & Lord, 2001) have identified a number of linguistic features that slow down the reader, increasing the chances of misinterpretation. In one study, they used their linguistic modification approach in that mathematics items were modified to reduce the complexity of sentence structures and unfamiliar vocabulary was replaced with familiar vocabulary (Abedi & Lord, 2001). The mathematics scores of both ELL students and non-ELL students in low and average-level mathematics classes improved significantly when the linguistic modification approach was used. In another study, Abedi and his colleagues (Abedi, Lord, Hofstetter, & Baker, 2000) found that out of four different accommodation strategies for ELLs, only the linguistically modified English form narrowed the gap between English Language Learner and other students. Thus, the linguistic modification approach can be used in the design of performance assessments to help ensure a valid and fair assessment of not only English Language Learners, but other students who may have difficulty with reading.

When students are asked to explain their reasoning on mathematics and science assessments, the writing ability of the student could be a source of construct-irrelevant variance. To help minimize the impact of writing ability on math and science assessments, scoring rubrics need to clearly delineate the relevant criteria. Construct-irrelevant variance may also occur when raters score student responses to performance tasks according to features that do not reflect the scoring criteria and are irrelevant to the construct being assessed (Messick, 1994). This can also be addressed by clearly articulated scoring rubrics and effective training of the raters.

Validity criteria that have been suggested for examining the quality of performance assessments include, but is not limited to, content representation, cognitive complexity, meaningfulness, transfer and generalizability, fairness, and consequences (Linn, Baker, & Dunbar, 1991; Messick, 1994). The discussion that follows is organized around these validity criteria. These criteria are closely intertwined to some of the sources of validity evidence proposed by the *Standards for Educational and Psychological Measurement.* (AERA, APA, and NCME, 1999): evidence based on test content, response processes, internal structure, relations to other variables, and consequences of testing.

Evaluating Content Representativeness

An analysis between the content of the assessment and the construct it is intended to measure provides important validity evidence (AERA, APA, NCME, 1999). Test content refers to the skills,

knowledge and processes that are intended to be assessed by tasks as well as the task formats and scoring procedures. Performance tasks can be designed so as to emulate the skills and processes reflected in the targeted construct. For many large scale assessment programs, it is important to ensure that the ability to generalize from a student's score on a performance assessment to the broader domain of interest is not limited by having too small of a number of tasks on the performance assessments. Although the performance tasks may be assessing students understanding of some concepts or set of concepts at a deeper level, the content of the domain may not be well represented by a relatively small subset of performance tasks. This can be addressed by including other item formats that can appropriately assess certain skills, and using performance task to assess complex thinking skills that cannot be assessed by the other item formats. For some high-stakes large scale assessments, including state assessment and accountability systems, performance tasks are used in conjunction with multiple-choice items to ensure that the assessment represents the content domain and to allow for inferences about individual student performance to the broader domain.

Methods are currently being investigated that will allow for accurate student level scores derived from mathematics and language arts performance assessments that are administered on different occasions throughout the year (Bennett et al., 2009). This will not only allow for content representation across the performance assessments, but the assessments can be administered in closer proximity to the relevant instruction and information from any one administration can be used to inform future instructional efforts. If school level scores are of interest primarily, matrix sampling procedures can be used to ensure content representation on the performance assessment as was done on the Maryland State Performance Assessment Program (Maryland State Board of Education, 1995).

The coherency and representativeness among the assessment tasks, scoring rubrics and procedures, and the target domain is another aspect of validity evidence for score interpretations. It is important to ensure that the cognitive skills and content of the target domain is systematically represented in the tasks and scoring procedures. The method used to transform performance to a score can provide evidence for the validity of the score interpretation. Both logical and empirical evidence can support the validity of the method used for transforming performance to a score.

For performance demonstrations such as a high school project, we are not interested in generalizing the student performance on the demonstration to the broader domain so the content domain does not need to be represented fully. The content and skills being assessed by the performance demonstration should be meaningful and relevant within the content domain. Performance demonstrations provide the opportunity for students to show what they know and can do on a real world task, similar to a driver's license test.

Evaluating Cognitive Complexity

One of the most attractive aspects of performance assessments is that they can be designed to assess complex thinking and problem solving skills. As Linn and his colleagues (1991) have cautioned, however, it should not be assumed that a performance assessment measures complex thinking skills, evidence is needed to examine the extent to which tasks and scoring rubrics are capturing the intended cognitive skills and processes. The alignment between the cognitive processes underlying task responses and the construct domain needs to be made explicit because typically the goal is to generalize scores interpretations to nonassessment construct domain interpretations (Messick, 1989). The validity of the score interpretations will be affected by the extent to which the design of performance assessments is guided by cognitive theories of student achievement and learning within academic disciplines. Further, the use of task templates will allow for the explicit delineation of the cognitive skills required to perform particular task types.

Several methods have been used to examine whether tasks are assessing the intended cognitive skills and processes (Messick, 1989), and they are particularly appropriate for performance assessments that are designed to tap complex thinking skills. These methods include *protocol analysis, analysis of reasons,* and *analysis of errors.* In *protocol analysis,* students are asked to think aloud as they solve a problem or describe retrospectively how they solve the problem. In the method of *analysis of reasons,* students are asked to provide rationales, typically written, to their responses to the tasks. The method of *analysis of errors* requires an examination of procedures, concepts or representations of the problems in order to make inferences about students' misconceptions or errors in their understanding. As an example, in the design of a science performance assessment Shavelson and Ruiz-Primo (1998) used Baxter and Glaser's (1998) analytic framework that reflects a content-process space depicting the necessary content knowledge and process skills for successful performance. Using protocol analysis Shavelson and Ruiz-Primo (1998) compared expert and novice reasoning on the science performance tasks that were content-rich and process-open. Their results from the protocol analysis confirmed some of their hypotheses regarding the different reasoning skills that that tasks were intended to elicit from examinees. Further, the results elucidated that complexity of experts' reasoning as compared to the novices and informed the design of the tasks and interpretation of the scores.

Evaluating Meaningfulness and Transparency

An important validity criterion for performance assessments is their meaningfulness (Linn et al., 1991) which refers to the extent to which students, teachers and other interested parties find value in the tasks at hand. Meaningfulness is inherent in the idea that performance assessments are intended to measure more directly the types of reasoning and problem solving skills that are valued by educators. A related criteria is transparency (Frederiksen & Collins, 1989), that is students and teacher need to know

what is being assessed, by what methods, the criteria used to evaluate performance, and what constitutes quality performance. It is important to ensure that all students are familiar with the task format and scoring criteria for both large-scale assessments and classroom assessments. Teachers can use performance tasks with their students, and engage them in discussions about what the tasks are assessing and the nature of the criteria used for evaluating student work. Teachers can also engage students in using scoring rubrics to evaluate their own work and the work of their peers.

Evaluating the Generalizability of Score Inferences

For many large scale assessments, the intent is to draw inferences about student achievement in the domain of interest based on scores derived from the assessment. A potential threat to the validity of score interpretations therefore is the extent to which the scores from the performance assessments can be generalized to the broader construct domain (Linn, Baker & Dunbar, 1991). It should be noted however, that this is not the intent for performance demonstrations as discussed previously.

Generalizability theory provides both a conceptual and statistical framework to examine the extent to which scored derived from an assessment can be generalized to the domain of interest (Brennan, 1996, 2000, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). It is particularly relevant in evaluating performance assessments that assess complex thinking skills because it examines multiple sources of errors that can limit the generalizability of the scores, such as error due to tasks, raters and occasions. Error due to tasks occurs because there are only a small number of tasks typically included in a performance assessment. As explained by Haertel and Linn (1996), students' individual reactions to specific tasks tend to average out on multiple-choice tests because of the relatively large number of items, but such individual reactions to specific items have more of an effect on scores from performance assessments that are composed of relatively few items. Thus, it is important to consider the sampling of tasks and by increasing the number of tasks on an assessment, the validity and generalizability of the assessment results is enhanced. Further, this concern with task specificity is consistent with research in cognition and learning that underscores the context-specific nature of problem solving and reasoning in subject matter domains (Greeno, 1989). The use of multiple item formats, including performance tasks, can improve on the generalizability of the scores.

Error due to raters can also affect the generalizability of the scores in that raters may differ in their evaluation of the quality of students' responses to a particular performance task and across performance tasks. Raters may differ in their leniency resulting in rater mean differences or they may differ in their judgments about whether one student's response is better than another student's response resulting in an interaction between the student and rater facets (Hieronymus & Hoover, 1987; Lane, Liu, Ankenmann, & Stone, 1996; Shavelson, Baxter, & Gao, 1993). Typically, occasion is an important hidden source of

error because performance assessments are only given on one occasion and occasion is not typically considered in generalizability studies (Cronbach, Linn, Brennan, & Haertel, 1997).

Generalizability theory estimates variance components for the object of measurement (e.g., student, class, school) and for the sources of error in measurement such as task and rater. The estimated variance components provide information about the relative contribution of each source of measurement error. The variance estimates are then used to design measurement procedures that allow for more accurate score interpretations. As an example, the researcher can examine the effects of increasing the number of items or number of raters, or both, on the generalizability of the scores. Generalizability coefficients are estimated to examine the extent to which the scores generalize to the larger construct domain for relative or absolute decisions, or both.

Generalizability studies have shown that error due to raters for science hands-on performance tasks (e.g., Shavelson et al., 1993) and mathematics constructed response items (Lane et al., 1996) tends to be smaller than for writing assessments (Dunbar et al., 1991). To help achieve consistency among raters attention is needed in the design of well articulated scoring rubrics, selection and training of raters, and evaluation of rater performance prior to and throughout operational scoring of student responses (Lane & Stone, 2006; Linn, 1993; Mehrens, 1992). Researchers have shown that task-sampling variability as compared to rater-sampling variability in students' scores is a greater source of measurement error in science, mathematics, and writing performance assessments (Baxter, Shavelson, Herman, Brown, & Valdadez, 1993; Gao, Shavelson, & Baxter, 1994; Hieronymus & Hoover, 1987; Lane et al., 1994, 1996; Shavelson et al., 1993). In other words, increasing the number of tasks in an assessment has a greater effect on the generalizability of the scores than increasing the number of raters scoring student responses.

Shavelson and his colleagues (1993) reported that task-sampling variability was the major source of measurement error using data from mathematics and science performance assessments. The results of their generalizability studies on a math assessment and two science assessments indicated that the person x task variance component accounted for the largest percent of total score variation, approximately 49%, 48%, and 82%, respectively. This indicates that students were responding differently across the performance tasks. The variance components that included raters (i.e., rater effect, person x rater interaction, and task x rater interaction) were either zero or negligible, indicating that sampling variability due to raters contributed little to no measurement error. They reported that to reach a .80 generalizability coefficient 15 tasks were needed for the math assessment, 8 for the state science assessment and 23 for the other hands-on science performance assessment. Lane and her colleagues (1996) found similar results with a mathematics performance assessment that consisted of constructed-response items requiring students to show their solution processes and explain their reasoning. The results indicated that error due

to raters was negligible, whereas error due to tasks was more substantial indicating that there was differential student performance across tasks. Generalizability studies for each form of the mathematics assessment indicated that between 42% and 62% of the total score variation was accounted for by the person x task variance component. Again, persons were responding differently across tasks due to task specificity. The variances due to the rater effect, person x rater interaction, and rater x task interaction were negligible. When the number of tasks was equal to 9 the generalizability coefficients ranged from .71 to .84. They also examined the generalizability of school-level scores for each form. The coefficients for absolute decisions (e.g., standards-based decisions) ranged from .80 to .97 when the number of tasks was equal to 36 using a matrix sampling design, providing evidence that the assessment allowed for accurate generalizability of grade level scores for the schools.

Shavelson and his colleagues (Shavelson et al., 1993; Shavelson, Ruiz-Primo, and Wiley, 1999) provided evidence that the large task sampling variability in science performance assessments was due to variability in both the person x task interaction and the person x task x occasion interaction. They conducted a generalizability study using data from a science performance assessment (Shavelson et al., 1993). The person x task variance component accounted for 32% of the total variability whereas, the person x task x occasion variance component accounted for 59% of the total variability. The latter suggests that students performed differently on each task from occasion to occasion. Shavelson and his colleagues (1999) provided additional support for the large effects due to occasion. In their generalizability study, the person x task variance component accounted for 26% of the total variability and the person x task x occasion variance component accounted for 31% of the total variability, indicating that there was a tendency for students to change their approach to each task from occasion to occasion. The variance component for the person x occasion effect was close to zero. In summary, "even though students approached the tasks differently each time they were tested, the aggregate level of their performance, averaged over the tasks, did not vary from one occasion to another" (Shavelson et al., 1999, p. 64-65).

In summary, the results from generalizability studies indicate that scoring rubrics and the procedures used to train raters can be designed so as to minimize rater error. Further, the use of well designed automated scoring systems allows for consistent application of the scoring rubrics in evaluating student work. Also, increasing the number of performance tasks will increase the generalizability of the scores. Likewise, including other item formats on performance assessments will aid in the generalizability of scores to the broader content domain.

Fairness of Assessments

The evaluation of the fairness of an assessment is inherently related to all sources of validity evidence. Bias can be conceptualized "as differential validity of a given interpretation of a test score for any definable, relevant subgroup of test takers" (Cole & Moss, 1999, p. 205). A fair assessment therefore requires evidence to support the meaningfulness, appropriateness, and usefulness of the test score inferences for all relevant subgroups of examinees. Validity evidence for assessments that are intended for students from various cultural, ethnic and linguistic backgrounds needs to be collected continuously and systematically as the assessment is being developed, administered and refined. The linguistic demands on items can be simplified to help ensure that English Language Learners are able to access the task as well as other students. As Abedi and his colleagues (2001) have demonstrated through their language modification approach, simplifying the linguistic demands on items can narrow the gap between English Language Learners and other students. The contexts used in mathematics tasks can be evaluated to ensure that they are familiar to various subgroups and will not negatively affect the performance on the task for one or more subgroups. The amount of writing required on mathematics, reading, and science assessments, for example, can be examined to help ensure that writing ability will not unduly influence the ability of the students to demonstrate what they know and can do on these assessments. Scoring rubrics can be designed to ensure that the relevant math, reading, or science skills are the focus, and not students' writing ability. The use of other response formats, such as graphic organizers, on reading assessments may alleviate the concerns of writing ability cofounding student performance on reading assessments (O'Reilly & Sheehan, in press).

Some proponents of performance assessments in the 1980's had hoped that subgroup differences that were exhibited on multiple-choice tests would be smaller or alleviated by using performance assessments. However, as stated by Linn and his colleagues (1991), differences among subgroups most likely occur because of differences in learning opportunities, familiarity, and motivation, and are not necessarily due to item format. Research that has examined subgroup differences has focused on both the impact of an assessment on subgroups by examining mean differences or differential group performance on individual items when groups are matched with respect to ability, that is, differential item functioning (Lane & Stone, 2006). Differential item functioning (DIF) methods are commonly used for examining whether individual test items are measuring a similar construct for different groups of examinees (e.g., gender and ethnic groups) of similar ability. Differential item functioning occurs when there is differential performance on an item for subgroups of students of approximately equal ability. The presence of DIF may suggest that inferences based on the test score may be less valid for a particular group or groups. Although researchers have argued that performance assessments offer the potential for more equitable assessments, performance assessments may measure construct-irrelevant features that contribute to DIF.

Gender or ethnic bias could be introduced by the typical contextualized nature of performance tasks or the amount of writing and reading required. In addition, the use of raters to score responses to performance assessments could introduce another possible source of differential item functioning (see for example, Gyagenda & Engelhard, in press). Results from DIF studies can be used to inform the design of assessment tasks and scoring rubrics so as to help minimize and potential bias.

Some researchers have supplemented differential item functioning methods with cognitive analyses of student performances designed to uncover reasons why items behave differently across subgroups of students of approximately equal ability. In a study to detect DIF in a mathematics performance assessment consisting of constructed-response items that required students to show their solution processes and explain their reasoning, using the *analyses of reasons* method, Lane, Wang and Magone (1996) examined differences in students' solution strategies, mathematical explanations and mathematical errors as a potential source of differential item functioning. They reported that for those items that exhibited DIF and favored females, females performed better than their matched males because females tended to provide more comprehensive conceptual explanations and were more complete in displaying their solution strategies. They suggest that increasing the opportunities in instruction for students to provide explanations and show their solution strategies may help alleviate these differences. Ericikan (2002) examined differential item response performances among different language groups. In her research, she conducted linguistic comparisons across different language test versions to identify potential sources of differential item functioning. Her results suggest that care is needed in the writing of items so as to minimize linguistic demands of items. As Wilson (2005) has suggested, the inclusion of DIF parameters into measurement models would allow for a direct measurement of different construct effects such as using different solution strategies and different types of explanations or to capture linguistic differences.

Some research studies have shown both gender and ethnic mean differences on performance assessments that measure complex thinking skill. As an example, ethnic and gender differences in persuasive writing were observed by Gabrielson, Gordon and Engelhard (1995). Their results indicated that high school female students wrote higher quality persuasive essays than male students, and white students wrote essays of higher quality than black students. The scores for conventions and sentence formation were more affected by gender and ethnic characteristics than the scores in content and organization and style which were consistent with results from Engelhard et al. (1994). These differences may be more reflective of differences in learning opportunities and motivation than true differences in ability, again suggesting the need for instruction to provide similar opportunities for all students.

More recently, studies have used advances in statistical models to examine subgroup differences so as to better control for student demographic variables and school level variables. One study examined the extent to which potentially heavy linguistic demands of a performance assessment might interfere with the performance of students who have English as a second language (Goldschmidt, Martinez, Niemi, and Baker, 2007). The results obtained by Goldschmidt and his colleagues (2007) revealed that subgroup differences on student written essays to a writing prompt were less affected by student background variables than a language arts commercially developed test consisting of multiple choice items and some constructed-response items. The performance gaps between white students, English only students and traditionally disadvantaged students (e.g., English language learners (ELLs)) were smaller for the writing performance assessment than the commercially developed test (Goldschmidt et al., 2007). Thus, the performance of students on the writing assessment used in this study was less affected by student demographic variables than their performance on the commercially developed test. As the authors indicate, although these are promising results, additional research is needed to determine if they can be replicated in other settings and with other subgroups. In particular, students in this study had opportunities in instruction to craft written essays, and such learning opportunities may have led to the results because of the alignment between instructional opportunities and the writing performance assessment.

Consequential Evidence

The evaluation of both intended and unintended consequences of any assessment is fundamental to the validation of score interpretation and use (Messick, 1989, 1995). Because a major goal of performance assessments is to improve teaching and student learning, it is essential to obtain evidence of such positive consequences and any potentially negative consequences (Messick, 1994). As Linn (1993) stated, the need to obtain evidence about consequences is "especially compelling for performance-based assessments… because particular intended consequences are an explicit part of the assessment systems' rationale" (p. 6). Further, adverse consequences bearing on issues of fairness are particularly relevant because it should not be assumed that a contextualized performance task is equally appropriate for all students because "contextual features that engage and motivate one student and facilitate his or her effective task performances may alienate and confuse another student and bias or distort task performance may alienate and confuse another student and bias or distort task performance" (Messick, 1994, p. 19). This concern can be addressed by a thoughtful design process in which fairness issues are addressed, including expert analyses of the tasks and rubrics as well as analyses of student thinking as they solve performance tasks with special attention to examining potential subgroup differences and features of tasks that may contribute to these differences.

Large-scale performance assessments that measure complex thinking skills have been shown to have a positive impact on instruction and student learning (Lane, Parke, & Stone, 2002; Stecher, Barron, Chun & Ross, 2000; Stein & Lane, 1996; Stone & Lane, 2003). In a study examining the consequences of Washington's state assessment, Stecher and his colleagues (Stecher et al. 2000) indicated that approximately two-thirds of 4th and 7th grade teachers reported that the state standards and the state assessment short-answer and extended-response items were influential in promoting better instruction and student learning. An important aspect of consequential evidence for performance assessments is the examination of the relationship between changes in instructional practice and improved student performance on the assessments. A series of studies examined the relationship between changes in instructional practice and improved performance on the Maryland State Performance Assessment Program, MSPAP, which was comprised entirely of performance tasks that were integrated across content domains (Lane, Parke, & Stone, 2002; Parke, Lane, & Stone, 2006; Stone & Lane, 2003). The results revealed that teacher reported reform-oriented instructional features accounted for differences in school performance on MSPAP in reading, writing, mathematics and science, and they accounted for differences in the rate of change in MSPAP school performance in reading and writing. The former suggests that schools in which teachers reported that their instruction over the years reflected more reform-oriented problem types and learning outcomes similar to those assessed by MSPAP had higher levels of school performance on MSPAP than schools in which teachers reported that their instruction reflected less reform-oriented problem types and learning outcomes. The latter suggests that increased reported use of reform-oriented performance tasks in writing and reading and a focus on the reading and writing learning outcomes in instruction was associated with greater rates of change in MSPAP school performance over a 5 year period. Support for these results in the mathematics domain was provided by a study conducted by Linn, Baker, and Betebenner (2002). They demonstrated that the slope of the trend lines for the math assessments on both NAEP and MSAP were similar, suggesting that the performance gains in Maryland are not specific to the content and format of either test but due to deepened mathematical understanding on the part of the students.

When using test scores to make inferences regarding the quality of education, contextual information is needed to inform the inferences and actions (Haertel, 1999). Stone and Lane (1993) indicated that a school contextual variable, percent free or reduced lunch which is typically used as a proxy for SES, was significantly related to school-level performance on MSPAP in mathematics, reading, writing, science and social studies. That is, schools with a higher percentage of free or reduced lunch tended to perform poorer on MSPAP. There was no significant relationship, however, between percent free or reduced lunch and growth on MSPAP at the school-level in four of the five subject areas – mathematics, writing,

science and social studies. This result indicates that improved school performance on performance assessments like MSPAP is not affected by contextual variables such as SES (as measured by percent free or reduced lunch). In other words, school level growth on the science, math, writing and social studies performance assessment was *not* related to the percent of students who were eligible for free or reduced lunches within the school.

Instructional Sensitivity

An assessment concept that can help inform the consequential aspect of validity is instructional sensitivity. Instructional sensitivity refers to the degree to which tasks are sensitive to improvements in instruction (Popham, 2003; Black & William, 2007). Performance assessments are considered to be vehicles that can help shape sound instructional practice by modeling to teachers what is important to teach and to students what is important to learn. In this regard it is important to evaluate the extent to which improved performance on an assessment is linked with improved instructional practices. To accomplish this, the assessments need to be sensitive to improvements of instruction. Assessments that may not be sensitive to well-designed instruction may be measuring something outside of instruction such as irrelevant constructs or learning that may occur outside of the school.

Two methods have been used to examine whether assessments are instructional sensitive: Studies have either examined whether students have had the opportunity to learn (OTL) the material or they have examined the extent to which differences in instruction affect performance on the assessment. In a study using a model-based approach to assessment design (Baker, 2007), it was found that student performance on a language arts performance assessment were sensitive to different types of language instruction and were able to capture improvement in instruction (Niemi, Wang, Steinberg, Baker, & Wang, 2007). This study examined the effects of 3 different types of instruction (literary analysis, organization of writing, and teacher-selected instruction) on student responses to an essay about conflict in literary work. The results indicated that students who received instruction on literary analysis were significantly more able to analyze and describe conflict in literature that students in the other two instructional groups, and students who had direct instruction on organization of writing performed significantly better on measures of writing coherency and organization. These results provide evidence that performance assessments can be instructional sensitivity with respect to different types of instruction, and suggest the need to ensure alignment and coherency among curriculum, instruction and assessment.

Evaluation of Additional Psychometric Characteristics of Performance Assessments

This section briefly discusses additional psychometric issues in the design of performance assessments. First, a brief presentation on the measurement models that have been developed for

performance assessments and extended constructed- response items will be provided. Measurement models that account for rater effects will also be introduced. These types of models have been used successfully in large-scale assessment programs to account for rater error in the scores obtained when evaluating performance assessments, allowing for more valid score interpretations. This will be followed with a brief discussion on issues related to linking performance assessments.

Measurement Models and Performance Assessments

Item Response Theory (IRT) models are typically used to scale assessments that consist of performance tasks only and assessments that consist of both performance tasks and multiple-choice items. IRT involves a class of mathematical models that are used to estimate test performance based on characteristics of the items and characteristics of the examinees that are presumed to underlie performance. The models use one or more ability parameters and various item parameters to predict item responses (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985). These parameters in conjunction with a mathematical function are used to model the probability of a score response as a function of ability.

The more commonly applied models assume one underlying ability dimension determines item performance (Allen & Yen, 1979), and accommodate ordinal response scales that are typical of performance assessments. They include the graded response model (Samejima, 1969; 1996), the partial credit model (Masters, 1982), and the generalized partial credit model (Muraki, 1992). As an example, Lane, Stone, Ankenmann, and Liu (1995) demonstrated the use of the graded response model with a mathematics performance assessment and Allen, Johnson, Mislevy, and Thomas (1994) discussed the application of the generalized partial credit model to NAEP which consists of multiple-choice items and constructed-response items.

Performance assessment data may be best modeled by multidimensional item response theory (MIRT) which allows for the estimation of student proficiency on more than one skill area (Reckase, 1997). The application of MIRT models to assessments that are intended to measure student proficiency on multiple skills can provide a set of scores that profile student proficiency across the skills. These scores can then be used to guide the instructional process so as to narrow gaps in student understanding. Profiles can be developed at the student level or the group level (e.g., class) to inform instruction and student learning. MIRT models are particularly relevant for performance assessments because these assessments are able to capture complex performances that draw on multiple dimensions within the content domain, such as procedural skills, conceptual skills and reasoning skills.

Modeling Rater Effects Using IRT Models

As previously discussed, performance assessments require either human scorers or automated scoring procedures in evaluating student work. When human scorers are used, performance assessments are considered to be "rater-mediated' since they do not provide direct information about the domain of interest but mediated information through interpretations by raters (Engelhard, 2002). Engelhard (2002) provided a conceptual model for performance assessments in which the obtained score is not only dependent on the domain of interest (e.g., writing ability), but also on rater severity, difficulty of the task, and the structure of the rating scale (e.g., analytic versus holistic, number of score levels). Test developers exert control over the task difficulty and the nature of the rating scale; however, a number of potential sources of construct-irrelevant variance are introduced into the rating process by the raters, including differential interpretation of score scales, differential assignment of ratings to males and females, halo effects, and bias in rater interpretation of task difficulty (Engelhard, 2002). These sources of construct-irrelevant variance can affect the validity and fairness of score interpretations.

Models have been developed that account for rater variability in scoring performance assessments. As an example, Patz and his colleagues (Patz, 1996; Patz, Junker, & Johnson, 2002; and Patz, Junker, Johnson, & Mariano, 2002) developed a hierarchical rating model to account for the dependencies between rater judgments. A parameter was introduced into the model that could be considered an "ideal rating" or expected score for an individual, and raters could vary with respect to how close their rating is to this ideal rating. This variability reflects random error (e.g., lack of consistency) as well as systematic error (e.g., rater tendencies such as leniency). As discussed by Bejar, Williamson, and Mislevy (2006) this modeling of rater variability may reflect an accurate modeling of rater cognition in that under operational scoring conditions raters may try to predict the score an expert rater would assign based on the scoring rubric and benchmark papers. In addition, covariates can be introduced into the model to predict rater behaviors such as rater features (e.g., hours of scoring) and item features. The modeling of rater variability is a way to account for error in the scores obtained when evaluating performance assessments, and allows for more valid interpretations of the scores.

Equating and Linking

Equating helps ensure comparability of interpretations of assessment results from assessment forms administered at one time or over time; however, equating an assessment that consists of only performance tasks is complex (Kolen, 2004). One way to equate forms so that they can be used interchangeably is to have a common set of items, typically called anchor items, on each of the forms. The anchor items are then used to adjust for any differences in difficulty across forms. An important issue that needs to be addressed in using performance tasks as anchor items in the equating procedure is that rater teams could change their scoring standards over time and the application of standard equating practices would lead to

bias in the equating process and consequently inaccurate scores (Bock, 1995; Kim, Walker & McHale, 2008a; Tate, 1999). As a solution to this problem, Tate (1999, 2000) suggested an initial linking study in which any changes in rater severity and discrimination across years could be identified. This would allow for an accurate assessment of across- rater team ability differences and equating of the tests. To accomplish the equating, a large representative sample of anchor item papers (i.e., trend papers) from Year 1 are rescored by raters in Year 2. These raters in Year 2 are the same raters who score the new constructed responses in Year 2. These trend papers now have a set of scores from the old raters in Year 1 and a set of scores from the new raters in Year 2. This allows for examining the extent to which the two rater teams across years differ in severity in assigning scores, and then adjustments can be made to ensure the two tests are on the same scale. Tate contends that instead of having item parameters, there are item/rating team parameters that reflect the notion that if the rating team changes across years, any differences due to the change in rating teams will be reflected in the item parameters. Another way to conceptualize this is that the item parameters are confounded by rater team effects so the rating team needs to be considered in the equating.

The effectiveness of this IRT linking method using trend score papers was established by Tate and his colleague (Tate, 2003; Kamata & Tate, 2005). The use of trend score papers in non-IRT equating methods has also proven effective by Kim and colleagues (2008a; 2008b). They compared the effectiveness of equating for a design that required anchor items and a design that did not require anchor items with and without trend score papers. The design that does not incorporate anchor items alleviates the concern of content representativeness of anchor items. Their results indicated that both designs using trend score papers were more effective in equating the scores as compared to those designs that did not use the trend score papers. More importantly, their results indicate that changes in rater severity can be examined and the equating of test forms across years should adjust for differences in rater severity if the trend scoring indicates that a rater shift has occurred (Kim et al., 2008a, 2008b). Trend scoring should be implemented for any assessment program that uses constructed-response items for equating across years to control for equating bias caused by a scoring shift over years. Kim and colleagues (2008b) point out that the trend scoring method requires additional rating of student papers which in turn increases cost, and it may be a bit cumbersome to implement. The use of image and online scoring methods however can ease the complexities of the implementation of the trend scoring method.

Conclusion

Performance assessments have been an integral part of educational systems in many countries however they have not been fully utilized in this country. There is evidence that the format of the

assessment affects the type of thinking and reasoning skills that are used by students, with performance assessments being better suited to assessing high level, complex thinking skills (e.g., Martinez & Katz, 1996). Recent advances in the design and scoring of performance assessments support their increased use in large-scale assessment programs. In particular, computer simulations allow for the design of meaningful, real world tasks that require students to problem solve and reason. Scores can be generated for computer simulation tasks across a number of dimensions and reported in a timely manner given the advances in automated scoring systems. Automated scoring systems have also proven effective in the evaluation of student essays to writing prompts and short constructed-response items. Advances in the design of automated scoring systems will continue to support the increased use of computers for assessment design and delivery, allowing for a more integrative, comprehensive approach to assessment design.

Well-specified content standards that reflect high level thinking and reasoning skills can guide the design of performance assessments so as to ensure the alignment among curriculum, instruction and assessment. Various task design strategies have proven useful in helping ensure the validity and fairness of performance assessment results. The language modification approach used by Abedi and Lord (2001) that minimizes the complexity of linguistic demands on mathematics items has led to improved performance for English Language Learners as well as other students. The use of computer based reading comprehension items where students use graphic displays to demonstrate their understanding will help minimize the extent to which students' writing ability affects their scores on a reading comprehension assessment (O'Reilly & Sheehan, in press). Task templates can be designed so as to ensure tasks embody the intended cognitive demands, and are not measuring one or more irrelevant constructs. Task templates also have the potential to increase the production of tasks, especially for computer-based simulation tasks.

When human raters are used, well articulated scoring rubrics and rigorous training procedures for raters will minimize error introduced in the scores due to inconsistency within and among raters. Measurement models and procedures have been designed to model rater errors and inconsistencies so as to control for them in the estimation of student scores on performance assessments. Procedures and measurement models have also been developed to account for changes in rater performance over years and to adjust student scores due to these rater changes. This is important because any changes in student performance over the years should reflect actual changes in proficiency and learning, and not be a result of inconsistent scoring on the part of the raters. Further, if raters become more stringent over years, this could mask any improved student performance on the assessment.

The educational benefit of using performance assessments has been demonstrated by a number of researchers (Lane et al., 2002; Niemi et al., 2007; Stecher et al., 2000). When students are given the

opportunity to work on meaningful, real world tasks in instruction, students have demonstrated improved performance on performance assessments. Moreover, research has shown that growth on performance assessments at the school level is not related to SES variables. Sound educational practice begs for the alignment among curriculum, instruction and assessment, and there is ample evidence to support the use of performance assessments in both instruction and assessment to improve student learning for all students.

Figure 1
BEAR Assessment System Construct Map for the Matter Strand in Chemistry

	Matter Substrands	
Levels of Success	Visualizing Matter: Atomic and Molecular Views	Measuring Matter: Measurement and Model Refinement
5 - Integrating	bonding and relative reactivity	models and evidence
4 - Predicting	phase and composition	limitations of models
3 - Relating	properties and atomic views	measured amounts of models
2 - Representing	matter with chemical symbols	mass with a particulate view
1 - Describing	properties of matter	amounts of matter

Source: Adapted from Wilson (2005)

Figure 2
BioKids Assessment Tasks for “Formulating Scientific Explanations Using Evidence”

Figure 3
Holistic General Scoring Rubric for Mathematics Constructed-Response Items

Performance Criteria

4 *Mathematical Knowledge.* Shows understanding of the problem's mathematical concepts and principles; uses appropriate mathematical terminology and notations; executes algorithms completely and correctly.

Strategic Knowledge. Identifies all the important elements of the problem and shows understanding of the relationships among them; reflects an appropriate and systematic strategy for solving the problem; gives clear evidence of a solution process, and solution process is complete and systematic.

Communication Gives a complete response with a clear, unambiguous explanation and/or description; may include an appropriate and complete diagram; communicates effectively to the identified audience; presents strong supporting arguments which are logically sound and complete; may include examples and counter-examples.

3 *Mathematical Knowledge.* Shows nearly complete understanding of the problem's mathematical concepts and principles; uses nearly correct mathematical terminology and notations; executes algorithms completely; and computations are generally correct but may contain minor errors.

Strategic Knowledge. Identifies the most important elements of the problem and shows general understanding of the relationships among them; and gives clear evidence of a solution process, and solution process is complete or nearly complete, and systematic.

Communication. Gives a fairly complete response with reasonably clear explanations or descriptions; may include a nearly complete, appropriate diagram; generally communicates effectively to the identified audience; presents strong supporting arguments which are logically sound but may contain some minor gaps.

2 *Mathematical Knowledge.* Shows understanding of some of the problem's mathematical concepts and principles; and may contain computational errors.

Strategic Knowledge. Identifies some important elements of the problem but shows only limited understanding of the relationships among them; and gives some evidence of a solution process, but solution process may be incomplete or somewhat unsystematic.

Communication Makes significant progress towards completion of the problem, but the explanation or description may be somewhat ambiguous or unclear; may include a diagram which is flawed or unclear; communication may be somewhat vague or difficult to interpret; and arguments may be incomplete or may be based on a logically unsound premise.

1 *Mathematical Knowledge.* Shows very limited understanding of some of the problem's mathematical concepts and principles; may misuse or fail to use mathematical terms; and may make major computational errors.

Strategic Knowledge. Fails to identify important elements or places too much emphasis on unimportant elements; may reflect an inappropriate strategy for solving the problem; gives incomplete evidence of a solution process; solution process may be missing, difficult to identify, or completely unsystematic.

Communication Has some satisfactory elements but may fail to complete or may omit significant parts of the problem; explanation or description may be missing or difficult to follow; may include a diagram, which incorrectly represents the problem situation, or diagram may be unclear and difficult to interpret.

0 Shows no understanding of the problem's mathematical concepts and principles.

Source: Adapted from Lane (1993)

Figure 4
Bear Assessment System Scoring Guide for the Matter Strand in Chemistry

Level	Descriptor	Criteria
0	Irrelevant or Blank Response	Response contains no relevant information
1	Describe the properties of matter	Rely on macroscopic observation and logic skills. No use of atomic model. Uses common sense and no correct chemistry concepts. 1- makes one or more macroscopic observation and/or lists chemical terms without meaning 1 uses macroscopic observation AND comparative logic skills to get a classification, BUT shows no indication of using chemistry concepts 1+ makes simple microscopic observations and provides supporting examples, BUT chemical principle/rule cited incorrectly
2	Represent changes in matter with chemical symbols	Beginning to use definitions of chemistry to describe, label, and represent matter in terms of chemical composition. Use correct chemical symbols and terminology 2- Cites definitions/rules about matter somewhat correctly 2 Cites definition/rules about chemical composition 2+ Cites and uses definitions/rules about chemical composition of matter and its transformation
3	Relate	Relates one concept to another and develops models of explanation
4	Predicts how the properties of matter can be changed	Apply behavioral models of chemistry to predict transformation of matter
5	Explains the interactions between atoms and molecules	Integrates models of chemistry to understand empirical observations of matter

Source: Adapted from Wilson (2005)

Figure 5.
Knowledge Integration Scoring Rubric

Link Levels	Description
Complex	Elaborate 2 or more scientifically valid links among relevant ideas
Full	Elaborate 1 scientifically valid link between 2 relevant ideas
Partial	State relevant ideas but do not fully elaborate the link between relevant ideas
No	Make invalid ideas or have non-normative ideas

Source: Liu, Lee, Hofstetter, and Linn (2008).

Appendix

References

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

Abedi, 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.

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.

Baker, E.L. (2007). Model-based assessments to support learning and accountability: The evolution of CRESST's research on multiple-purpose measures. *Educational Assessment, 12*(3&4), 179-194.

Baker, E.L., O'Neil, H.F., & Linn, R.L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist, 48*(12), 1210-1218.

Baron, J.B. (1991). Strategies for the development of effective performance exercises. *Applied Measurement in Education, 4*(4), 305-318.

Baxter, G.P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice, 17*(3), 37-45.

Baxter, G.P., Shavelson, R.J., Herman, S.J., Brown, K.A. & Valdadez J.R., (1993). Mathematics performance assessment: Technical quality and diverse student impact. *Journal for Research in Mathematics Education, 24*, 190-216.

Bejar, I.I.,, Williamson, D.M., & Mislevy, R.J. (2006). Human Scoring (pp. 49-82). In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *automated scoring of complex tasks in computer-based testing.* Hillside, NJ: Erlbaum.

Ben-Simon, A. & Bennett, R.E. (2007). Toward more substantively meaningful essay scoring. *Journal of Technology, Learning and Assessment 6*(1), accessed at http://escholarship.bc.edu/jtla/.

Bennett, R.E. (2006). Moving the field forward: Some thoughts on validity and automated scoring (pp. 403-412). In D.M. Williamson, R.J. Mislevy, & I.I. Behar (Eds.), *Automated scoring of complex tasks in computer-based testing.* Hillside, NJ: Erlbaum.

Bennett, R.E. & Gitomer, D.H. (in press). Transforming K-12 Assessment: Integrating Accountability Testing, Formative Assessment and Professional Support. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century.* New York: Springer.

Bennett, R. E., Persky, H., Weiss, A.R., & Jenkins, F. (2007). Problem solving in technology-rich environments: A report from the NAEP Technology-Based Assessment Project (NCES 2007-466). Washington, DC: National Center for Education Statistics, U.S. Department of Education. Retrieved May 2, 2009 from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2007466.

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

Bock, R.D. (1995). Open-ended exercise in large-scale educational assessment. In L.B. Resnick and J.G. Wirt (Eds.), *Linking school and work: Roles for standards and assessment.* (pp. 305-338). San Francisco, CA: Jossey-Bass.

Brennan, R.L. (1996). Generalizability of performance assessments. In G.W. Phillips (Ed.), *Technical issues in large-scale performance assessment.* Washington, DC: National Center for Education Statistics (NCES 96-802).

Brennan, R.L. (2001). *Generalizability Theory.* New York: Springer-Verlag.

Brennan, R.L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24*, 339-353.

Bridgeman, B., Trapani, C. & Attali, Y. (2009). *Considering fairness and validity in evaluating automated scoring.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Burstein, J. (2003). The E-rater scoring engine: Automated essay scoring with Natural Language Processing. In. M.D. Shermis and J.C. Burstein (Eds.), *Automated Essay Scoring* (pp. 113-122). Mahwah, NJ: Lawrence Erlbaum Associates.

Brown, J.S., & Burton, R.R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2,* 155-192.

Chi, M.T.H., Feltovich, P.J., & Glaser, R. (1981). Categorization and representation of physics problems of experts and novices. *Cognitive Science, 5*, 121-152.

Chi, M.T.H., Glaser, R., & Farr, M. (Eds.) (1988). *The nature of expertise.* Hillsdale, NJ: Erlbaum.

Clauser, B.E. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement, 24*(4), 310-324.

Cole, N.S. & Moss, P.A. (1989). Bias in test use. In R.L. Linn (Ed.), *Educational Measurement*, (3rd ed., pp. 201-220). New York: American Council on Education and Macmillan.

Clyman, S.G., Melnick, D.E., & Clauser, B.E. (1995). Computer-based case simulations. In E.L. Mancall & P.G. Bashook (Eds.), Assessing clinical reasoning: The oral examination and alternative methods (pp. 139– 149). Evanston, IL: American Board of Medical Specialties.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles.* New York, NY: John Wiley.

Cronbach, L.J., Linn, R.L., Brennan, R.L., & Haertel, E.H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57*(3), 373-399.

Darling-Hammond, L., Ancess, J., & Falk, B. (1995). *Authentic assessment in action: Studies of school and students at work.* New York City: NY, Teachers' College Press.

Delaware Department of Education. (2000, November). *Delaware student testing Program Special Writing Study Report.* Retrieved from www.doe.k12.de.us/aab/report_special_writing%20study.pdf.

DeVore, R. N. (2002). *Considerations in the development of accounting simulations.* (Technical Report 13). NJ: AICPA.

Dunbar, S.B., Koretz, D.M., & Hoover, H.D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*(4), 289-304.

Educational Testing Service (2004, July 4). *Pretesting plan for SAT essay topics.* [internal communication].

Elliott, S. (2003). Intellimetric: From Here to Validity. In. M.D. Shermis and J.C. Burstein (Eds.), *Automated Essay Scoring* (pp.71-86). Mahwah, NJ: Lawrence Erlbaum Associates.

Embretson, S.E. (1985). *Test Design: Developments in Psychology and Psychometrics.* Orlando, FL: Academic Press.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261-287). Mahwah, NJ: Erlbaum.

Engelhard, Jr., G., Gordon, B., Walker, E.V., & Gabrielson, S. (1994). Writing tasks and gender: Influences on writing quality of black and white students. *Journal of Educational Research, 87,*197-209.

Ericikan, K. (2002) Disentangling sources of differential item functioning in multi-language assessments. *International Journal of Testing, 2*, 199-215.

Ericsson, K.A., & Simon, H.A. (1984). *Protocol analysis*: *Verbal reports as data*. Cambridge, MA: MIT Press.

Ericsson, K.A. & Smith, J. (1991). Prospects and limits of the empirical study of expertise: An introduction. In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp.1-38). Cambridge: Cambridge University Press.

Ferrara, S.F. (1987, April). *Practical considerations in equating a direct writing assessment required for high school graduation.* Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C.

Frederiksen, J.R. & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Gabreilson, S., Gordon, B., & Engelhard, G. (1995). The effects of task choice on the quality of writing obtained in a statewide assessment. *Applied Measurement in Education, 8*(4), 273-290.

Gao, X., Shavelson, R.J., & Baxter, G.P. (1994). Generalizability of large-scale performance assessments in science. Promises and problems. *Applied Measurement in Education, 7*, 323-334

Glaser, R., Lesgold, A., & Lajoie, S. (1987). Toward a cognitive theory for the measurement of achievement. In R.R. Ronning, J.A., Glover, J.C., Conoley, & J.C. Witt (Eds.), *The influence of cognitive psychology on testing* (pp. 41-85). Hillsdale, NJ: Erlbaum.

Goldberg, G.L. & Roswell, B.S. (2001). Are multiple measures meaningful?: Lessons learned from a statewide performance assessment. *Applied Measurement in Education, 14*(2), 125-150.

Goldschmidt, P., Martinez, J.F., Niemi, D., & Baker, E.L. Relationships among measures as empirical evidence of validity: Incorporating multiple indicators of achievement and school context. *Educational Assessment, 12*(3 & 4), 239-266.

Gotwals, A.W. & Songer, N.B. (2006). *Cognitive Predictions: BioKIDS Implementation of the PADI Assessment System.* (PADI Technical Report 10). Menlo Park, CA: SRI International.

Greeno, J.G. (1989). A perspective on thinking. *American Psychologist*, 44, 134-141.

Gyagenda, I.S., & Engelhard, G. (in press). Rater, domain, and gender influences on the assessed quality of student writing. In Garner, M., Engelhard, G., Wilson, & M., Fisher, W. (Eds.). Advances in Rasch Measurement, Volume One. JAM Press.

Haertel, E.H. (1999). Performance assessment and education reform. *Phi Delta Kappan, 80*(9), 662-667.

Haertel, E.H. & Linn, R.L. (1996). Comparability. In G.W. Phillips (Ed.), *Technical Issues in Large-Scale Performance Assessment* (NCES 96-802). Washington, DC: U.S. Department of Education.

Huot, B. (1990). The literature of direct writing assessments: Major concerns and prevailing trends. *Review of Educational Research, 60*(2), 237-263.

Kane, M. T. (2006). Validation. In B. Brennan (Ed.), *Educational Measurement*. American Council on Education & Praeger: Westport, CT.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17.

Kim, S., Walker, M.E., & McHale, F. (2008a, May). *Equating of mixed-format tests in large-scale assessments.* (ETS Research Report-08-26). Princeton, NJ: ETS.

Kim, S., Walker, M.E., & McHale, F. (2008b, October). *Comparisons among designs for equating constructed-response items.* (ETS Research Report-08-53). Princeton, NJ: ETS.

Lane, S. (in press). Issues in the design and scoring of performance assessments that assess complex thinking skills. In G. Schraw (Ed.). *Assessment of Higher Order Thinking Skills.*

Lane, S. (1993). The conceptual framework for the development of a mathematics performance assessment instrument. *Educational Measurement: Issues and Practice, 12*(3), 16-23.

Lane, S., Liu, M., Ankenmann, R.D., & Stone, C.A. (1996). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement, 33*(1), 71-92.

Lane, S., Parke, C.S., & Stone, C.A. (2002). The Impact of a State Performance-Based Assessment and Accountability Program on Mathematics Instruction and Student Learning: Evidence from Survey Data and School Performance. *Educational Assessment, 8*(4), 279-315.

Lane, S., Silver, E.A., Ankenmann, R.D., Cai, J., Finseth, C., Liu, M., Magone, M.E., Meel, D., Moskal, B., Parke, C.S., Stone, C.A., Wang, N., & Zhu, Y. (1995). *QUASAR Cognitive Assessment*

Instrument (QCAI). Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

Lane, S. & Stone, C.A. (2006). Performance Assessments. In B. Brennan (Ed.), *Educational Measurement*. American Council on Education & Praeger: Westport, CT.

Lane, S., Stone, C. A., Ankenmann, R. D., & Liu, M. (1995). Examination of the assumptions and properties of the graded item response model: An example using a mathematics performance assessment. *Applied Measurement in Education, 8*, 313-340.

Lane, S., Wang, N., & Magone, M. (1996). Gender related DIF on a middle school mathematics performance assessment. *Educational Measurement: Issues and Practice, 15*, 21-27,31.

Linn, R.L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15*, 1-16.

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1991). Complex performance assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Liu, O.L., Lee, H. Hofstetter, C., & Linn, M.C. (2008). Assessing knowledge integration in science: Constructs, measures, and evidence. *Educational Assessment, 13*(1), 33-55.

Lloyd-Jones, R. (1977). Primary trait scoring. In C.R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, and judging* (pp. 33-60). Urbana, IN: National Council for Teachers in Education.

Martinez, M.E. & Katz, I.R. (1996). Cognitive processing requirements of constructed figural response and multiple-choice items in architecture assessment. *Educational Assessment, 3*(1), 83-98.

Maryland State Board of Education (1995). *Maryland school performance report*: *State and school systems.* Baltimore, MD: author.

Maryland State Department of Education. (1990). Technical report: Maryland Writing Test, Level II. Baltimore, MD: Author. Retrieved from http://www.marces.org/mdarch/htm/M031987.HTM.

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

Messick, S. (1989) Validity. In R.L. Linn (Ed.), *Educational Measurement*, (3rd ed., pp. 13-104). New York: American Council on Education and Macmillan.

Messick, S. (1996). Validity of performance assessments. In G.W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (1-18). Washington, DC: National Center for Educational Statistics.

Miller, M.D. & Crocker, L. (1990). Validation methods for direct writing assessment. *Applied Measurement in Education, 3*(3), 285-296.

Mislevy, R.J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33,* 379-416.

Mislevy, R.J. & Haertel, G.D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6-20.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the Structure of Educational Assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3-62.

Mislevy, R. J., Steinberg, L. S., Breyer, F.J., Almond, R.G., & Johnson, L. (2002). Making sense of data from complex assessments. *Applied Measurement in Education, 15*(4), 363-390.

Mullis, I.V.S. (1984). Scoring direct writing assessments: What are the alternatives? *Educational Measurement: Issues and Practice, 3*(1), 16-18.

Myford, C.M., & Mislevy, R.J. (1995). *Monitoring and improving a portfolio assessment system*. (Center for Performance Assessment Research Report). Princeton, NJ: Educational Testing Service.

National Council on Education Standards and Testing (1992). *Raising standards for American Education.* Washington, DC: Author.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment.* Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds). Board on Testing and Assessment, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council (2006). *Systems for State science assessment*. M.R. Wilson & M.W. Bertenthal (Eds.). Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Nelson, N. & Calfee, R.C. (1998). The reading-writing connection. In N. Nelson & R. C. Calfee (Eds.), *The reading-writing connection*. University Chicago Press.

Niemi, D., Wang, J., Steinberg, D.H., Baker, E.L., & Wang, H. (2007). Instructional sensitivity of a complex language arts performance assessment. *Educational Assessment, 12*(3 & 4), 215-238.

O'Reilly, T., & Sheehan, K. M. (in press). Cognitively based assessment of, for and as learning: a framework for assessing reading competency (RR-xx-xx). Princeton, NJ: Educational Testing Service.

Parke, C. S., Lane, S., & Stone, C. A. (2006). Impact of a state performance assessment program in reading and writing. *Educational Research and Evaluation, 12*(3), 239-269.

Popham, W.J. (2003). Living (or dying) with your NCLB tests. *School Administrator, 60*(11), 10-14.

Patz, R.J., Junker, B.W., Johnson, M.S., & Mariano, L.T. (2002). The hierarchical rater model for rated test items and is application to large-scale educational assessment data. *Journal of educational and Behavioral Statistics, 27*(4), 341-384.

Powers, D. E., Burstein, J.C., Chodorow, M.S., Fowles, M. E., & Kukich, K. (1998). *Qualifying essay readers for an online scoring network (OSN).* (RR 98-20). Princeton, NJ: ETS.

Reckase, M.D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21,* 25-36.

Resnick, L.B. & Resnick, D.P. (1982). Assessing the thinking curriculum: New tools for educational reform. In B. G. Gifford & M.C. O'Conner (Eds.)., *Changing assessment: Alternative views of aptitude, achievement and instruction* (pp. 37-55), Boston: Kluwer Academics.

Shavelson, R.J., Baxter, G.P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*(3), 215-232.

Shavelson, R.J., Ruiz-Primo, M.A., & Wiley, E.W. (1999). Note on sources of sampling variability. *Journal of Educational Measurement, 36*(1), 61-71.

Simon, H.A., & Chase, W.G. (1973). Skill in chess. *American Scientist, 61,* 394-403.

Smith, C.L., Wiser, M. Anderson, C.W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement, 14*(1&2), 1-98.

Stecher, B., Barron, S., Chun, T., & Ross, K. (2000, August). *The effects of the Washington state education reform in schools and classrooms* (CSE Tech. Rep. NO. 525). Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing.

Stein, M.K. & Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching and learning in a reform mathematics project. *Educational Research and Evaluation, 2*(1), 50-80.

Tate, R. L. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal of Educational Measurement, 36,* 336-346.

Tate, R.L., (2000). Performance of a proposed method for the linking of mixed format tests with constructed-response and multiple-choice items. *Journal of Educational Measurement, 36,* 336-346.

U.S. Department of Education (2005). The Nation's Report Card. Washington, DC: Author Retrieved from http://nationsreportcard.gov/science_2005/s0116.asp.

Vendlinski, T.P., Baker, E.L., & Niemi, D. (2008). *Templates and objects in authoring problem-solving assessments.* (CRESST Tech. Rep. No. 735). Los Angeles: University of California, National Center Research on Evaluation, Standards, and Student Testing (CRESST).

Williamson, D.M., Behar, I.I., & Mislevy, R.J. (2006). Automated scoring of complex tasks in computer-based testing: An introduction. In D.M. Williamson, I.I. Bejar, & R.J. Mislevy (Eds.) *Automated*

scoring of complex tasks in computer-based testing. (pp. 1-14). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Erlbaum.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13,* 181-208.

Wolfe, E. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing, 4*(1), 83-106.

Yang, Y., Buchendahl, C.W., Juszkiewicz, P.J., Bhola, D.S. A review of strategies for validating computer-automated scoring. *Applied Measurement in Education, 15*(4), 391-412.

Do not cite or quote without permission

The Cost of New Higher Quality Assessments:

A Comprehensive Analysis of the Potential Costs for Future State Assessments

Assessment Solutions Group

DRAFT -- January, 2010 – DRAFT

Acknowledgments

This report was written by Barry Topol, John Olson, and Ed Roeber, Partners and Senior Consultants in the Assessment Solutions Group. Data analysis was conducted by Patrick Hennon, who also created the tables and charts for the report. The authors would like to thank Linda Darling-Hammond, Frank Marshall Adamson, and Tom Toch for their invaluable input and advice. The authors also would like to thank the Nellie Mae Education Foundation for their financial support of this study, the Ford Foundation, the Sandler Foundation, and the Hewlett Foundation, which supported a broader set of studies of performance assessment under the aegis of the Stanford Center for Opportunity Policy in Education, and the Council for Chief State School Officers, within which this work is situated.

I. Executive Summary

[Note: this section will be written last, after the other 4 sections have been reviewed and discussed with the sponsors of this work]

DRAFT

II. Overview, Purpose, and Background

The purpose of this study is to determine, as precisely as possible in advance, the amount of money a typical state would incur to implement a high-quality assessment system including performance components in comparison to the amount currently being spent on their state assessment under three new conditions: the economies available from collaborating with other states in a consortium constructing common assessment items and tasks; the economies available from using technology for assessment development, distribution, and scoring; and specific conditions for teacher involvement in moderated scoring. The data from the study will be used to inform states, policymakers, and other key decision makers how much new, higher quality assessment systems could cost under various conditions, and what the impact of some cost mitigation strategies might be. The design of a high-quality assessment was developed by staff at the Assessment Solutions Group (ASG), in particular, Ed Roeber, Michigan State University, based on the work that Linda Darling-Hammond and her colleagues at Stanford University conducted in collaboration with the Council of Chief State School Officers to lay out a vision of what assessment systems that support higher-quality learning might look like, and to summarize research on lessons regarding performance assessment over the last several decades. This work is summarized in a corresponding collection of papers.[1]

As part of the process for designing and developing the new assessments that have been proposed, it was determined that costs for various types of assessment approaches needed to be modeled to provide an accurate estimate on what the overall costs (development and ongoing administration) would be to states. This report summarizes the background, assumptions, methodology, and results of the analyses conducted by ASG, the organization contracted to do this work. As noted, ASG developed the designs for a typical current state assessment and a new high-quality assessment, and analyzed the costs for these assessments using its cost model system in order to summarize the costs and compare detailed information for the two. For the purposes of this study, the estimates that are provided are illustrative and not intended to be the only resolutions of the questions regarding how to best implement a high quality assessment system. Different calculations could be obtained depending on the specific assessment design and/or vendor solution selected.

Background

Among the many driving forces impacting state assessment, the issues of increased amounts of testing, cost, lack of state funds, and assessment quality are at the forefront. Two new initiatives, the Common Core Standards project and the Race to the Top common assessments competition (discussed in more detail later in this section), are helping focus attention on important steps for improving state assessments. The ongoing work by Stanford, the Council of Chief State School Officers (CCSSO), the National Governors Association (NGA), and others to address these issues also are great steps in this direction.

[1] Add citations to papers by Jamal Abedi; Linda Darling-Hammond; Suzanne Lane; Stuart Kahl & Raymond Pecheone; Larry Picus et al.; and Brian Stecher.

However, given the current financial situation in most states, new assessment designs need to be as cost-effective and efficient as possible. States are now trying to do more with less – increased amounts of testing but with decreased funding.

The level of statewide assessment activity occurring in the United States jumped dramatically during the past two decades. In the early 1990's, fewer than 30 states had some type of statewide assessment activity and this usually consisted of only one statewide assessment program component. The adoption of the Improving America's Schools Act (IASA) in 1994 began the trend to increased statewide assessment activity, since it required that all states create academic content standards in the areas of mathematics and reading/English language arts, as well as assessments at one elementary, one middle school, and one high school grade. The Individuals with Disabilities Education Act of 1997 (IDEA-97) added the requirement that all students with disabilities participate in statewide assessments, while the No Child Left Behind (NCLB) Act of 2001 expanded the number of grades (to grades 3-8 plus one high school grade) and content areas assessed (adding science assessment no later than 2007 in at least one elementary, middle school, and high school grade). Both IASA and NCLB also required states to assess the English language proficiency of English language learners (ELLs).

The result is that the amount of statewide assessment in each state has increased dramatically. In states that pioneered statewide assessment programs, such as Michigan during the 1960's, the state did not add any new assessment components until these Federal laws went into effect in the 1990's. It now has six different assessment programs, each covering different grades and/or subgroups of students, for different assessment purposes – a more than 500% increase in the size of the state assessment. The amount of change in other states has been comparable, with many adding statewide assessments as a state accountability policy lever for the first time in their states' history in addition to expanding assessments to meet Federal requirements.

Clearly, a substantial number of testing programs are being operated at the state level across the nation.

Synonymous with the considerable and rapid expansion of statewide assessment efforts is the equally dramatic increases in the costs for the assessment programs. Whereas once states' assessment costs were a minor part of the state education agencies' budgets, now the costs are substantially higher (and much more noticeable to policymakers and the public). The required state assessments that once cost just a few million dollars can now run to as much as $100 million per year in a large state. Even though a portion of these costs is paid from Federal funds, the overall state portion of the costs of testing has risen dramatically in recent years.

In the past decade, the total amount of testing-related costs has increased dramatically. Across the U.S., various studies have estimated assessment activities now cost anywhere from $950 million to $1.3 billion annually, and are increasing. In the conference report accompanying the adoption of NCLB, Congress mandated that the U.S. Government Accounting Office (GAO) conduct a study to estimate the costs to states of complying with NCLB between 2002 and 2008 (GAO, 2003). The GAO chose to cost out three scenarios. The first was an all multiple-choice (MC) item format for all required state tests. The second

was a scenario in which states used a combination of MC and short constructed-response (CR) items that were currently in use in 2002. The third scenario listed the costs if states were to use a combination of MC and extended CR items. The actual appropriations from the Federal government to states were at slightly above the first level over the six years from 2002 to 2008 (totaling $400+ million annually). Some states chose MC only programs, although most used a combination of MC and CR item formats. Thus, while states did receive some Federal support for the added costs of testing, they also had to appropriate additional state dollars to support mandated statewide assessment.

States committed to more extensive performance assessment – such as Connecticut – which included extended writing tasks, science investigations, and other intellectually challenging tasks in its assessments – were unable to afford a large share of the costs of their assessments when NCLB required annual every student be tested. Connecticut sued the US Department of Education for the costs of maintaining its rich assessment program under NCLB, and, in the course of negotiations, was advised by the Department to revert to multiple choice testing (citation to be added).

A few other studies also have been done over the years to look at total state assessment costs and one conclusion is that there is a paucity of recent and accurate cost figures for statewide assessment programs. Although there are a variety of estimates, and overall figures for Federal expenditures are readily obtainable, the amount that each state spends on its statewide assessment activities is not systematically collected, nor is it analyzed in any appreciable depth.

Also, although states spend a significant amount of money on their statewide assessments, many do not have accurate methods to objectively estimate the appropriate costs for their assessment programs. Furthermore, most states do not have access to good information as to what the costs should be for individual components or special features they may wish to include in an assessment, thereby making it extremely difficult to determine the relative value of one component/feature vs. another when constructing an assessment. Given the size and scope of the contracts, states need good information on costs that will help them create assessment designs that are as efficient as possible. This need becomes even more acute as states try to redesign their assessment programs to reflect higher quality designs that both improve instruction and student learning.

In addition, because the lack of funding has often resulted in states adopting less complex assessments with largely MC items or cutting back on the number of CR items so that only a few are used, the current approaches being used in states need to be re-evaluated. Many testing experts are encouraging states to assess their students at higher and deeper levels so that problem solving and higher order thinking skills can be better measured and reported. Therefore, it is important to closely look at the current designs being used in states and compare those to new designs that incorporate other approaches that can have more validity for improving instruction and assessing student learning, such as more use of Short CR and Extended CR items and use of innovative performance measures.

DRAFT

Common Core Standards and Race to the Top Common Assessment Initiatives

The development of the common core standards and Race to the Top (RTTT) funding for assessment and other educational reforms represent two important initiatives in upgrading the educational system in the United States. The common core standards is a joint project spearheaded by CCSSO and NGA to develop a common set of content standards for the states to benchmark the academic standards of the best and most rigorous educational systems in the world. These standards will be used to focus the curriculum on the rigorous skills students will need to succeed in the 21st century and help states in terms of improving student education and assessment. The common assessment, aligned to the common core standards, will make assessment an integral part of curriculum and instruction to actually improve student learning. Through a RTTT competition, funds from the USED will be used for new innovations in education, including assessment, and various consortia of states will be able to use these funds to develop a next generation of higher-quality assessments in reading/language arts and mathematics that can be used as part of their state assessment program in the future. The USED is interested in supporting one or more consortia of states that work toward jointly developing and implementing common, high-quality assessments aligned with a consortium's common set of K-12 standards that are internationally benchmarked and that build toward college and career readiness by the time of high school completion. New, innovative assessment designs are being considered that will both help students learn and teachers develop effective teaching and intervention strategies. These new assessments will likely include new item types such as performance tasks and performance events, as well as the use of more constructed response items than current assessments.

However, without any systemic changes in the way assessments are procured, developed and administered, the cost of these new assessments could exceed the cost of current assessments by a significant amount. If these costs are not anticipated and controlled, they could spell the end of such innovative approaches to assessment. As more details of this new initiative are being unveiled, few analysts have evaluated how much it will actually cost to develop, administer, and maintain these new, innovative assessments under different assessment conditions. Therefore, as noted earlier, this study provides information that can help determine the amount of money a typical state would incur to implement a high-quality assessment system in comparison to the amount currently being spent on their state assessment.

Additional Issues Affecting Current State Assessments

Some of the other issues concerning the design of current state assessments are that they are largely summative in their approach; the assessments are not always instructionally sensitive, balanced, or innovative; and little teacher instructional strategies or other useful information are provided that can positively impact students within the school year.

One of the purposes of these set of papers is to suggest better approaches to large-scale assessment that are cost effective and make it a valuable part of the curriculum. The ongoing work by Stanford, CCSSO, NGA, and others to design more innovative, high-quality assessments is an opportunity to improve some

of the current approaches that are being used. Also, the information provided by a detailed analysis of the costs for these types of assessments can help in the deliberations. ASG's expertise in designing and costing state assessments can be of much benefit to policy makers, states, and state consortia for the design more efficient and effective assessment solutions.

A key premise of this report is that the total cost of improved state assessments could be significantly more than current assessments if changes are not made in the way assessments are procured and delivered. Furthermore, assessment cost has been a "black box," especially since the advent of NCLB, and most states are not aware or informed of costs of many different features and functionalities. Thus, states are not able to make educated trade-offs or other decisions concerning changes to their assessment. In addition, it is likely that states are not efficient as they could be in their current assessment systems.

The ASG Cost Model

The ASG ***Assessment Cost Model*** is a variable input, metric-based output model. That is, specific assessment program variables are input to the model and applied against cost factors, metrics, and/or databases to derive assessment cost. Several hundred variables associated with the functions and activities required to develop and administer an assessment (e.g., item & test development, production and manufacturing, logistics, editing and scoring, reporting, psychometrics, program management, quality assurance, information technology etc.) are contained in the model which allows ASG to build up the cost of ***any*** assessment from the ground up as opposed to making generalized estimates of the cost of an assessment based on broad industry parameters. The ASG cost model has the ability to conduct a detailed study of the costs for all types of assessment components to not only determine the cost of the assessment but to also identify ways to improve the cost effectiveness and efficiency of a state assessment. The model generates a plethora of detailed cost information that can be used to evaluate assessment and assessment component cost. Cost reports by function, area, cost type etc. are generated and key metrics are presented to better understand assessment cost, as well as allowing for a comparison to model *"efficient cost"* data.

Without an assessment cost model, it is possible that a state, or consortia of states, will spend a significant amount of money on a new assessment, but not have accurate methods to objectively estimate the appropriate costs for the assessment program. Furthermore, in designing a new assessment, it is important to have access to good information as to what the costs should be for individual components or special features that may be included in the assessment, thereby making it possible to determine the relative value of one component/feature vs. another when constructing the assessment. The data from a rigorous cost analysis can help a state, or consortia of states, estimate the individual and total costs of a future assessment. It is hoped that states, as well as the USED, will then be able to budget more efficiently and effectively because they will have a better understanding of component costs and different options.

In this study, ASG used its proprietary assessment cost modeling software to determine the costs of a "typical" current state assessment and the costs for various new innovative assessments. After estimating costs for a current assessment, the cost model was used to determine the cost of a new, high quality assessment for a "typical" single state purchasing the assessment for its own use. The model was then

used to create different design, development, and delivery strategies in order to reduce the cost of the assessment. The resulting reduced assessment costs were then compared against the cost of the current assessment.

Benefits of ASG Methodology for Helping States Analyze Assessment Costs

In general, states do not have access to good information as to what the cost should be for individual components or special requirements they may wish to include in an assessment, nor do they have accurate methods to estimate all aspects of the costs for their assessment programs. States also may have difficulties preparing requests for proposals (RFPs) and comparing vendor proposals, and are never quite sure whether assessment contractors are proposing "apples-to-apples" programs. In a nutshell, states cannot always tell whether prices quoted for an assessment are too high, too low, or about right.

As various designs are proposed for new assessments in the future, states, or consortia of states, need methods that allow them to understand what they will be paying for – from the total bottom-line price down to each assessment component. This will allow states to make decisions about how to tailor the design of the assessments to most effectively and most efficiently assess students (i.e., meeting all Federal and state assessment requirements) while still providing the achievement information needed at the state and local district levels. If they are able to better understand their assessment costs, states will be able to better design and implement programs that not only meet Federal and state requirements, but more importantly, will be affordable. This may permit states to spend less on testing and more on helping local districts better use the assessment results to improve student learning, and thus achievement – a primary goal of large-scale assessment.

The results from this study help provide data for an apples-to-apples comparison of current "typical state assessment" and new high-quality assessment costs that are created with the same model using similar assumptions. The information reported from the cost model yields independent, objective, and accurate estimates of incremental costs for states and provides fair comparisons of various approaches to developing and delivering the new assessment. In addition, the information can help states that may want to upgrade their assessment system and/or cost out various cost reduction strategies. In the following sections of this report, details are provided on the assumptions used in the model, the methodology used to analyze the data, and the results from the series of cost analyses that were conducted.

DRAFT

III. Methodology and Key Assumptions

In this section, the methods used to analyze costs in this study are described. Information on the assumptions for the various models of state assessment programs that were evaluated are listed below, and details are provided on the definitions and scenarios used to run the cost models and compare the data.

ASG used a straightforward approach to determine the cost of a typical state assessment, for developing and implementing a new high quality assessment system, and to calculate the resulting incremental cost to adopt the new system. The following assessment models were created and analyses conducted:

1. Define and price a representative assessment program for a moderately large state - the goal of this analysis was to calculate the assessment costs associated with a typical current assessment program for reading/language arts and mathematics. This included the costs for the state assessment program run by the state, as well as an interim benchmark assessment program procured by the state for local district use.

- A typical state and typical assessment system were defined and noted as the current assessment system. The current system served as the baseline model for the calculation of both the development and ongoing administration costs. Cost calculations and comparisons (total and incremental) to other scenarios were made relative to this baseline model. Pricing for the current assessment system was developed using the Assessment Solutions Group (ASG) cost model to calculate the summative component and pricing of existing interim assessment products and ASG assumptions to calculate the interim assessment component.

2. Design and price a high quality, future assessment program for the same moderately large state - In the same typical state as used in number 1 above, several scenarios were created around the more innovative approaches to assessment that could be used in the future state assessment program. The costs for representative scenarios were determined.

3. Develop cost reduction strategies based on consortia of various numbers of states implementing the high quality assessment design – In these cases, one or more of the scenarios developed in number 2 above were used to develop costs for a consortium of 10, 20, and 30 states working together. Such cost estimates illustrate the cost savings for groups of states working together to create the assessments that are needed. Several analyses based on the state consortia model (A below) were created to examine further cost efficiency scenarios and are notated as follows:

A. Participating in a state assessment consortium to share development and overhead costs. State consortia sizes of 10, 20, and 30 states were analyzed.

B. Moving to online delivery of the assessment - Online assessment (OLA) delivery eliminates much of the cost of pencil and paper systems and many states have stated that they want to use an OLA in the future (if they have not already implemented one).

C. Using teachers to score performance event (PEs) and performance task (PTs) items – Two different models were examined, one (C1) assuming a professional development model with no additional teacher compensation beyond that supported by the state or district for normal professional development days and the other (C2) assuming an $125/day stipend to teachers.

D. Using distributed scoring for constructed response items - A scenario was run assuming a 50/50 mix of site based and distributed scoring for the constructed response items. Distributed scoring allows the person scoring the item to work from his or her home while accessing a centralized scoring and monitoring platform. Distributed scoring was also assumed, in all cases, for the scoring of performance event and performance task items.

E. Adopting automated scoring for some constructed response items - Automated systems are being developed and placed in service using computers to score essay type responses via the use of artificial intelligence (AI) engines. ASG examined one scenario, at a low per response price point, to determine the impact on assessment cost.

F. Developing a customized interim benchmark assessment system – The cost of developing an interim benchmark system with similar item types and structure as the high quality system was modeled. (Cases 3A-3E focus on the development and administration of the summative assessment system.)

   F(2) – Using state consortia to develop and make available different options for the administration of an interim assessment system.

Each of these models of assessment programs is explained more fully below. The results of the analyses and costs associated with each model are shown in the next section of this report. For the purposes of this study, the estimates that are provided are illustrative and not intended to be the only resolutions of the questions regarding how to best implement a high quality assessment system. Different calculations could be obtained depending on the specific assessment design and/or vendor solution selected.

Model 1 - Comprehensive Assessment Program for a Moderately Large State

The typical state assessment program now in use in the United States has a number of characteristics in common. These commonalities are driven by Federal requirements for such programs. The No Child Left Behind (NCLB) Act requires that states assess English language arts/reading and mathematics at grades 3-8 plus one high school grade, and to assess science at one elementary, middle school and high school grade. There are assessment requirements for English language learners (namely, annual assessment with an English proficiency test) and for students with disabilities (provision of assessment accommodations and an alternate assessment for such students unable to participate in the regular assessment program).

For purposes of this cost study, a moderately large state was selected to determine what the state should be paying for its assessment services – test development and test administration.

The typical summative assessment was defined as one administered at the end of the school year with 50 multiple choice questions and 2 extended constructed response items in mathematics and reading and 10 multiple choice questions and 1 extended constructed response item in writing. Writing was included because a substantial number of states have writing as part of their ELA assessment. Science was not included because the common core standards being created for use by the states include only English language arts and mathematics. Summative assessment assumptions are shown in Exhibit 1.

Exhibit 1 – Summative Assessment Assumptions

Summative Assessment Assumption	Description
Test Years	Year 0 (full field test) + 3 operational years
Grades/Students Assessed	Grades 3-8 and 10; 125,000 students per grade
Domains Assessed	Mathematics and English Language Arts (Reading & Writing)
Delivery Method	Pencil and paper for summative assessment
Number of Unique Test Forms	2 plus a breach form (breach form developed in year 1 but not printed)
Color	2 color (no items presented in color)
Item Release Rate	25% each year
Field Test Methodology	Full field test in year 0; embedded years 1-3
Travel and Meetings	Standard setting, bias review, sensitivity review (standard meetings)
Shipping Method	Ground transportation
Scanning Description	Scannable answer documents in grades 4-8, 10; scannable books in grade 3
Scoring of Constructed Response Items	Vendor scored, 20% read behind rate, 90% exact and adjacent agreement required
Reporting	End of year reports (state, district, school demographic, etc) electronically delivered with the exception of the parent report which is printed and mailed
Vendor Gross Margin	35%*

* Vendor gross margin higher than current industry average. All other things being equal, it is expected that prices will rise in the next few years.

Interim Benchmark Assessment

Interim assessments have become more common among local school systems that are concerned about whether all students will meet the accountability requirements of NCLB. The goal is to use assessments periodically to determine student progress towards mastering the knowledge and skills expected of students when they are assessed on the annual state assessment program instruments. There are several ways in which these interim assessments may be implemented and used.

Off-the-Shelf versus Custom Developed. One of the first choices for districts and states that wish to use interim benchmark assessments is whether to select one of the commercial products on the market or to develop their own instruments. The advantage of using an available set of assessments is that they are readily available and ready to use; they can be implemented easily in the school district. On the other hand, such off-the-shelf products may not measure the skills the district (or the state) considers to be most important, and this mismatch might not permit educators to receive good information on such outcomes.

An alternative that some districts and states are using is to create their own interim benchmark assessments. The advantage of this approach is that the assessments can better measure the skills considered to be important by the district (or the state) and in addition, can use a broader range of assessment types (constructed response items, performance events or performance tasks) not commonly found on off-the-shelf interim assessment products.

Paper-Based or Online Testing. Another choice that districts and states face is whether the interim benchmark assessment program is delivered online or via paper-based assessments. The advantage of online assessments is the ease of test administration and the speed of return of results. The major disadvantage is whether the school has the necessary computer infrastructure to permit assessing large numbers of students in a brief period of time.

Paper-based assessments are useful when a broader array of assessments – those for which online administration would be challenging – are used. In addition, they are helpful when the number of students to be assessed is small or if custom-developed interim benchmark assessments were developed and the cost of entering these into online assessment systems is viewed as too high.

State Education Agency Role in Interim Benchmark Assessment. There are several ways in which state education agencies might assist local districts that wish to use interim benchmark assessments. At one extreme are states that purchase a single interim benchmark assessment system (or that custom-develop one) for all districts in the state. Another potential role would be to provide mandate the use of such a system locally, but not provide a single system for the entire state. This may or may not come with the necessary resources for operating such a system. Third, states may provide assessment materials previously used in the assessment program for local district use in “stocking” their interim benchmark assessments. Finally, states might simply acknowledge that districts are using such a system and permit them to use Federal, state, or local funds to pay the necessary costs of developing and using such a system.

Given the above, there are currently a wide variety of options and associated pricing for interim assessments. If we were to select one such option it would be an online test with 40 MC and 0 CR items delivered three times a year at an all in price around $8 a student.[2]

Options for States Working Together. When states form consortia to develop and implement large-scale, summative assessments at the state level, they may also wish to consider how they could work together to provide interim benchmark assessments to their local school districts. There are at least four different options for this to occur:

A. The full consortium buys/leases a complete system (items and online delivery system) from a vendor and this system is provided to local districts to use as they see fit.

B. The consortium purchases/leases an online assessment system from a vendor, but the consortium loads its own assessment (which it has developed) into the system and provides the system for local district use.

C. The consortium develops its own assessments and administers these and the state assessments using the same online system that they have either created or leased.

D. The consortium develops its own assessments and provides these to the local school districts to use as they see fit – to load into any online system that they have and/or to use as paper-based assessments.

In order to assist states working in consortium to understand the costs of these options, cost estimates were prepared to show what each would cost per student and per state.

Exhibit 1B shows the interim assessment assumptions.

Interim Assessment Assumption	Description
Test Years	Years 1 - 3
Grades/ Students Assessed	Grades 3-8 and 10; 125,000 students per grade
Domains Assessed	Mathematics and English Language Arts
Delivery Method	Online (multiple annual administration) with a paper and pencil option
Number of Unique Test Forms	2
Color	Online system would have color capability
Item Release Rate	TBD
Field Test Methodology	Items are field tested
Reporting	Reporting is automated and available within hours

[2] Current and projected interim assessment pricing based on catalog pricing and interviews with Pearson, CTB/McGraw Hill and NWEA

Exhibit 1C shows the interim assessment per student pricing assumptions.

Interim Assessment System - Options Per Student Cost Assumptions				
States	Purchase System and Content	Purchase System Only - Add Owned Content	Purchase Summative and Interim System - Add Owned Content	Develop Content and Provide to Districts Upon Request
10	$8.00	$6.00	$2.00	Var.
20	$7.00	$5.00	$1.50	Var.
30	$6.00	$4.00	$1.00	Var.

Baseline Cost Determination

As stated above, the ASG cost model was used to develop the appropriate price for the summative assessment defined earlier for the typical state, calculated for Years 0, 1, 2, and 3. The assumptions above were used to develop the costs options for the interim assessment system. The combined costs for the various options can then be set as the baseline for comparing current and future assessment system costs.

Model 2 – High Quality Assessments

After the costs for the baseline assessment program were determined, the next step was to develop a scenario for the design of a high quality large-scale assessment program. This is not an easy task, since there are a variety of ideas about how large-scale assessments could be changed and improved. However, in order to develop cost estimates for such a program, a design was developed and specific quantities of different types of assessment items were determined. The high quality assessment designs include a greater mix of constructed response items than is commonly seen on such assessments today, as well as new innovative item types defined as performance events and performance tasks. Many testing experts view these types of items doing a much better job of testing a student's problem solving and higher order critical thinking skills than current multiple choice items. Table 1 shows a description of the types of items assumed to be used in such a program.

A summary of the new assessment designs is shown in Exhibit 2.

DRAFT

Table 1- High Quality Assessment Item Types and Examples

Multiple Choice – This is an on-demand item in which students select the correct answer from among four choices given to them.

Example Who wrote the play "Romeo and Juliet?"

A William Shakespeare
B Thornton Wilder
C William Blake
D Thomas Smythe

Short Constructed-Response (SCR) – This is an on-demand written exercise in which the student produces a response that ranges from a word or a few numbers to a few sentences or a few numbers.

Example Describe in one paragraph the basic plot of "Romeo and Juliet."

Extended Constructed-Response (ECR) – This is an on-demand written exercise in which the student produces a response that ranges from one paragraph to a couple of pages in response to a prompt. The essay is typically scored on a 0-4 or 0-6 basis for one or more dimensions.

Example Write an essay in which you describe the conflict(s) inherent in the play Romeo and Juliet. Then describe how such conflicts could occur in modern day America. Describe at least two ways in which this play could describe modern America.

Performance Event (PE) – This is an on-demand activity that students complete in a class period in school. It may involve a written activity, but often may involve students actually doing something, being observed and rated by the teacher. The PE will be scored on one or more dimensions, each typically on a 0-4 or 0-6 scale

Example The student is asked in an English class to sketch the set for a production of Shakespeare's "Romeo and Juliet" to illustrate how play were staged in England in Shakespeare's time. You have 30 minutes to complete this exercise.

Performance Task (PT) – This is an activity that students will work on in class and outside of class for periods ranging from a couple of days to several weeks. Typically, because these are such complex tasks, they may result in a paper, a completed project, and/or presentation. The PT

DRAFT

may involve multiple parts that could be scored holistically or separately. The PT will be scored on one or more dimensions, each typically on a 0-4 or 0-6 scale.

Example: Develop a paper, drawings, and a presentation to compare how plays written at the time of Shakespeare were staged versus how a play such as "Our Town" might be produced and staged. Consider changes in how plays were written, the venues where they were staged, the manner in which the audience would "interact" with the players, and the net effect on those who attended the production. Your paper should be at least five pages in length, include at a minimum two drawings, and your presentation should be at least ten minutes in length.

Exhibit 2 – Summative and interim assessment test designs.

Summative Assessment Design

Summative Assessment	**Item Counts**				
Mathematics	Multiple Choice	Short Constructed Response	Extended Constructed Response	Performance Event	Performance Task
Current Typical Assessment	50	0	2	0	0
High Quality Assessment	25	2 (1 in grade 3)	2 (0 in gr. 3, 1 in gr. 4)	2	2 (0 in gr. 3, 1 in gr. 4)
Summative Assessment	**Item Counts**				
English Language Arts	Multiple Choice	Short Constructed Response	Extended Constructed Response	Performance Event	Performance Task
Current Typical Assessment (Reading)	50	0	2	0	0
Current Typical Assessment (Writing)*	10	0	1	0	0
High Quality Assessment (Reading)	25	2 (1 in gr. 3 & 4)	2 (1 in gr. 3 & 4)	2	1
High Quality Assessment (Writing)*	10	2 (1 in gr. 3 & 4)	2 (1 in gr. 3 & 4)	2	0

*Administered in grades 4, 7 and 10

Interim Assessment Design

Interim Assessment	Item Counts				
Mathematics	Multiple Choice	Short Constructed Response	Extended Constructed Response	Performance Event	Performance Task
Current Typical Assessment**	40	0	0	0	0
High Quality Assessment**	25	2	1 (0 in grade 3)	1	1 (0 in grade 3)
Interim Assessment	**Item Counts**				
English Language Arts	Multiple Choice	Short Constructed Response	Extended Constructed Response	Performance Event	Performance Task
Current Typical Assessment**	40	0	0	0	0
High Quality Assessment**	25	2	1	1	1

**Administered 3 times a year

The full "procurement cost" of implementing the new assessment system was calculated which consists of the initial development of the items and forms (a Year 0 incremental expense), as well as the ongoing annual cost of administering the assessment program. Calculations for the cost per functional area (development, production, IT, program management, quality assurance, warehousing/logistics, scoring, etc.), cost per student, cost per grade, cost per domain and key metrics for each functional area and assessment as a whole were also generated. The model was used to calculate Year 0, Year 1, Year 2 and Year 3 assessment costs so the same data set was available to compare the future high quality assessment cost data to the baseline current assessment data.

Based on the results of previous studies, our estimates assume that the average time it takes an experienced person to score a performance event and performance task is 3 and 6 minutes respectively. [3] It is likely that these times could vary depending on the nature of the items actually used in a high quality assessment. The estimates selected are at the lower end of the range of times that have been previously measured in scoring performance items and should serve as a starting point in the discussion of the development of new high quality assessments. The estimates are based on average times reported by scorers working under the direction of an assessment organization, taking into account both students who write elaborate responses to the performance events and tasks, and those students who provide little or no response. The latter can constitute upwards of one-third of the "responses" to these items. The scoring

[3] Baron, 1984; Breland, Camp, Jones, Morris and Rock, 1987; Hymes, 1991; US Congress Office of Technology Assessment, 1992; Stevenson, 1990; Hill and Reidy, 1993.

time for each type of item is dependent on the number of parts that are written into each performance event or task, as well as how elaborate it is anticipated that student responses will be.

Model 3: Cost Reduction Strategies/Cases

Once the current and high quality assessment system costs were calculated for a given typical state, strategies were developed to reduce the cost of the new assessment system. The following models were examined and assessment system costs calculated. Note, each model and case was calculated independently of the other cases. A final calculation was made to determine the impact on assessment cost if all of the cost reduction strategies modeled were implemented.

A. Participation in a consortium of states to develop and administer a new high quality assessment – Participation in state consortia of 10, 20 and 30 states was examined. Participating in a consortium of states allows the members to spread the fixed costs of development and vendor overhead functions (IT, QA, etc.) over the entire group. Additionally, a consortium of states should be able to negotiate better pricing from both online and test development vendors. ASG assumed a lower vendor gross margin in the consortia cases.

B. Use of technology in delivering the assessment – The use of online testing was examined and the impact on assessment cost calculated. Online test development and administration systems can be significantly less expensive than using paper and pencil to administer the assessment. While costs to a state for the purchase of additional PCs was not modeled in our calculations, it would not be difficult for a state to calculate the cost benefit analysis of moving to online testing using the data in this study. Additionally, several strategies exist that a state can use to mitigate the impact of high student to PC ratios on the required testing window.

C. Use of teachers to score Performance Events and Performance Tasks - Different scenarios were run assuming teachers would score the performance events and performance tasks. In one scenario (C1) teacher scoring of performance events and performance tasks was treated as part of teacher professional development and therefore, the cost of scoring these items was not included in the calculation of total future assessment system costs. In another scenario (C2) it was assumed that teachers would be paid a $125 a day stipend to score the items.

D. Use of distributed scoring to mark the constructed response items - Distributed scoring of constructed response items is cheaper than centralized on site scoring because it avoids the fixed facilities, computer and scoring center management costs. ASG used a 50/50 a mix of site based and distributed scoring because both methods of scoring CR items are typically implemented in order to get the total number of readers required and, to a lesser extent, for the vendor to get a feel for the issues that arise in scoring particular responses. Note that for comparability purposes, ASG assumed distributed scoring for PEs and PTs in all cases.

E. Use of AI technology in scoring constructed response items – A major assessment cost is the labor required to score the constructed response items. A variety of systems have been or are

being developed and placed in service by vendors to automatically score student essay responses using Artificial Intelligence engines. Based on ASG's research[4], today these systems cost between $.50 and $3.00 per question with the bulk of the pricing by vendors at the higher end of the range. It is assumed that as time passes and systems continue to mature, pricing should become more affordable. A scenario was run, for the 30 state consortia, at $.50 per item to score a student response and $6,000 per item system training fee, to determine the impact of computer based scoring on the cost of scoring the constructed response items.

F. Development of a customized interim benchmark assessment system - In the case of a possible future interim assessment system, it was assumed that the state or state consortia would incur the initial development cost of creating the new interim assessments that would line up with the high quality summative assessment but would use an existing system to deliver the interim assessments at the same price as is available today. It was also assumed that teachers would score the constructed response questions as a normal part of the curriculum.

In Case F(2) different options for implementing and administering an interim assessment system were examined.

Additionally, it is quite possible that a state could pay less than commercial prices today for an interim system, particularly if the system is procured by a consortium of states or the same system is used to deliver both the summative and interim assessments. ASG made assumptions on future system prices based on discussions with current interim assessment system providers[5].

Finally, it should be noted that having a comprehensive, balanced assessment system with classroom based assessment components occurring during the year takes on some of the information purposes that are otherwise carried by interim assessments and thus has the potential to provide some economy in the system.

Exhibit 4 provides an overview of the various model cases, number of states testing, and online testing methods, as well as online pricing and vendor profit margin assumptions.

The data generated in the various cases are important in understanding the costs of converting to a new assessment system and how to mitigate the additional costs of implementing a high quality assessment system. The use of online technology, teacher scoring of performance items and participation in a consortium of states to procure and administer the assessment are important elements in maintaining affordable assessments in the future.

[4] Pricing estimates based on interviews with Vantage Learning, Internet Testing Services, Measurement Inc., and ASG research on other systems.

[5] Pearson, NWEA, and CTB/McGraw Hill

Advantages of Using this Methodology

The methodology outlined above and used in this study for calculating the baseline costs of a typical current assessment and the incremental costs of moving to a new, higher quality assessment system has several advantages. First, using the same model to calculate both the current and high quality assessment system costs provides an "apples to apples" comparison of the *incremental cost* of moving to a new assessment system. While the baseline assessment may not be structured exactly the same as that used in a particular state, it provides a good approximation of a typical state's assessment cost. Since the same model and assumptions are used to make the calculations for both the current and future assessment systems, the incremental cost of moving to a new assessment system has a high degree of validity. Next, the methodology avoids the problem of using a particular state's assessment cost as a baseline as any given state's assessment costs may not be representative of typical costs due to the particulars of that state's assessment program and/or the operational methodology in which the assessment is delivered. The methodology also eliminates differences in vendor pricing and operational assessment delivery practices as potential sources of bias and error. Finally, the methodology provides for further apples to apples cost comparisons when new, lower cost approaches to developing and administering assessments are developed and analyzed.

DRAFT

Model Cases

Note: High Quality Assessment (HQA) cases are independent of each other

Pricing and Profit Variables	1. Current State	2. New HQA	3A. High Quality Assessment with State Consortia Cases			3B. High Quality Assessment with Online Assessment Cases				3C. Teacher Scoring of PEs and PTs w/ and w/out stipend***			3D. HQA with Distributed Scoriing of CR Items			3E. HQA w/ Online Scoring of CR Items+	3F. Develop New Interim Assessment	3F(2). Develop New Interim Assessment Admin Options
Number of States	1	1	10	20	30	1	10	20	30	10	20	30	10	20	30	30	any	10-20-30
Vendor Gross Profit %*	35%	35%	30%	28%	25%	35%	30%	28%	25%	30%	28%	25%	30%	28%	25%	25%	n/a	n/a
Online Per Student Price**	-	-	-	-	-	$4.00	$3.50	$2.50	$1.50	$3.50	$2.50	$1.50	$3.50	$2.50	$1.50	$0.50	n/a	var
One-time Online Fixed Cost ($000)	-	-	-	-	-	$250	$300	$325	$350	$300	$325	$350	$300	$325	$350	$1,900	n/a	$0

* for development and admin functions in cases with online components

** Prime contractor overhead of 3%-4% is added on top of this number

*** Vendor system charge on teacher scoring cases is roughly 25% of total teacher stipend (whether paid or not)

+ 15% human read behind rate assumed. $.50 is the per **score** price.
Constructed Response (CR) items do not include performance events (PEs) and performance tasks (PTs). PEs and PTs already assumed to be scored using distributed scoring.
Note: cases 3B, 3C and 3D assume online system used to deliver the assessment and score all MC items, hence pricing for the online system component does not change

DRAFT

This page intentionally left blank.

DRAFT

IV. Results of the Cost Analyses

As described in the previous section, several different assessment designs were analyzed in the ASG cost model. For each of those models, the ASG cost estimation system was used to derive representative costs. This will permit assessment designers to determine how much high quality assessment systems might cost an individual state (working alone) and how these costs might be reduced through consortia of states working together, as well as by using a variety of cost savings strategies.

Summaries of the costs for each of these options are provided below.

1. Representative comprehensive assessment program for a moderately large state

The costs for the current assessment program for a "typical" state were estimated for four fiscal years – a base year (labeled Year 0) which is necessary to prepare and field test the needed assessment materials for use – and three additional years of operational testing, labeled Years 1, 2 and 3. For each year, the anticipated costs of operating the typical assessment program in a single state were calculated. Costs were calculated for the following: total cost; cost per student; cost by function; and, cost by content area and grade. Each of these is described for this assessment model. The presumption was made that only a limited number of constructed response items and no performance assessments (events or tasks) are used in this program, since many states have had to cut back or eliminate constructed response items due to budget cuts.

Table 1 summarizes the total yearly cost for this traditional comprehensive assessment program. These (and other) costs are based on the specifications shown in the previous section.

Table 1. Total Assessment Cost by Year

Year	Year 0	Year 1	Year 2	Year 3	Total Cost
Cost	$3,936,258	$16,533,386	$15,566,449	$16,189,107	$52,325,199

Year 0 shown in Table 1 includes development costs for the assessment program. Note that the costs for Year 1, which are higher than those for Years 2 and 3, include development, but not printing, of a breach form in addition to the operational forms. As can be seen, the cost of operating even a conventional comprehensive assessment program, one with limited use of constructed response items and no performance assessment, is substantial. However, since 2002, states have received support from federal funding associated with NCLB in order to afford these costs, with state funds being used to pay for the remaining costs.

Another way to examine these costs is on a per-pupil basis. The per-pupil cost is also considerable, as is shown in Table 2.

Table 2. Per-Pupil Cost of Assessment, Per Year

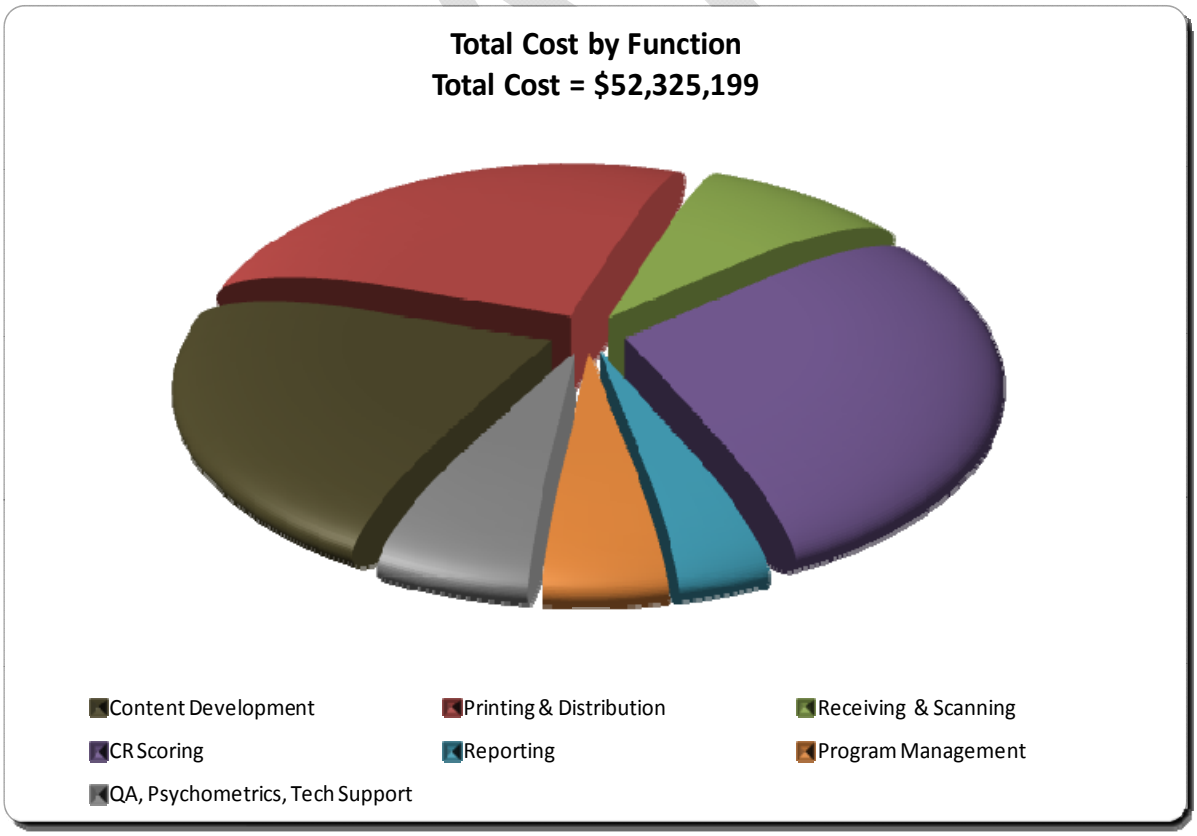
Year	Year 0	Year 1	Year 2	Year 3	Average Cost
Cost	$4.50	$19.01	$17.79	$18.50	$19.93

Note that average cost includes Year 0 expenditures and averages all costs over 3 years.

Table 2 indicates that there are development costs required for a high stakes assessment program even before it is administered. This is typically the case. The program is not too costly per pupil because only a limited number of expensive types of items (constructed response) and no performance assessments are used.

By examining the costs for this assessment model by function, it is possible to see which aspects of the assessment program are most and least costly. The former might provide areas to examine for cost savings. Figure 1 shows the costs for each portion of the assessment program.

Figure 1. Cost by Function



As can be seen, the most expensive portions of this assessment program are content development and constructed response scoring. Other expensive portions of this program include printing, distribution, and scanning.

A final way of looking at cost is by content area and grade, as is indicated in Table 3. This type of costing shows the different costs associated with the English language arts (reading and writing) assessments, as well as the mathematics assessments, at each grade level. Grade level costs will vary because, for example at third grade, a scannable test booklet is used, while an answer folder (scannable answer sheets) is used at the other grades.

Table 3. Cost by Content Area and Grade

Grade	Mathematics	Reading	Writing	Total
3	$1.37	$1.43	-	$5.80
4	1.10	1.19	$1.10	7.40
5	1.16	1.23	-	7.39
6	1.12	1.24	-	8.36
7	1.10	1.14	1.11	10.36
8	1.07	1.21	-	10.28
10	1.09	1.12	1.13	13.35
Total	$8.02	$8.57	$3.35	$19.93

As can be seen, while there are some differences between grade level and subject area costs, these are not substantial, due in part to the minimal use of written response items in the conventional assessment program at each grade and subject area.

The model generated costs for a typical current assessment for a moderately large state were mostly as expected, as summarized below:

- Costs and costs by function are in line and typical with what would be expected in a large scale assessment.

- Costs for production, manufacturing and warehousing are a bit lower than normal due to an efficient test design and the exclusion of science from the subject areas being tested. Adding a science assessment would yield costs for these functions in the $8.00 per student range, which is what would be expected in a typical large scale assessment.
- Development costs are about one-quarter of total costs and consistent with a typical assessment.
- CR scoring is a bit high but in the reasonable range at a bit less than one-third of total costs, which is the result of using all ECR items in the three exams.
- The writing assessment includes 10 MC questions vs. 25 for reading and mathematics. The lower development cost of the writing exam is offset by the higher scoring costs associated with the writing constructed response items.
- As noted earlier, the vendor margin is probably a bit higher than that experienced in the industry today but reflects ASG's views on the direction of future pricing.

2. An innovative, high-quality assessment program for the same moderately large state

The costs of a new high-quality assessment (HQA) program for the same "typical" state, with the same numbers of students assessed, were also estimated for the same four fiscal years – a base year (labeled Year 0) which is necessary to prepare the needed assessment materials for use – and three additional years of operational testing, labeled Years 1, 2 and 3. For each year, the anticipated cost of operating the high quality assessment program in the single state was calculated. Costs were also calculated for the following: total cost; cost per student; cost by function; and, cost by content area and grade. Each of these is described for this assessment model.

The high-quality model assessment program differed from the conventional one shown above mainly in terms of the number and type of constructed response items, as well as performance events and performance tasks. Also, with the addition of more CR and performance items, fewer MC items were used. Such a program involves considerably more scoring of student responses than in a traditional program, and it is anticipated that such a program will be substantially more expensive than the conventional program. Note that, in this case, it was assumed that all scoring activities were performed by the vendor. In new systems, it is possible that teachers within a state are part of a moderated scoring system, in part to support professional learning. We examine this possibility in later scenarios.

Table 4 summarizes the total yearly cost for the representative comprehensive assessment program. These (and other) costs are based on the specifications shown in the previous section.

Table 4. Total Assessment Cost by Year

Year	Year 0	Year 1	Year 2	Year 3	Total Cost
Cost	$7,813,641	$45,562,943	$45,473,513	$47,292,454	$146,142,551

As can be seen, the cost of operating the high-quality comprehensive assessment program, one with substantial use of constructed response items and performance assessment, is substantially larger than the conventional program - $146 million over three-plus years versus approximately $52 million (see Table 1).

This is also reflected in the per-pupil cost, which is shown in Table 5.

Table 5. Per-Pupil Cost of Assessment, Per Year

Year	Year 0	Year 1	Year 2	Year 3	Average Cost
Cost	$8.93	$52.07	$51.97	$54.05	$55.67

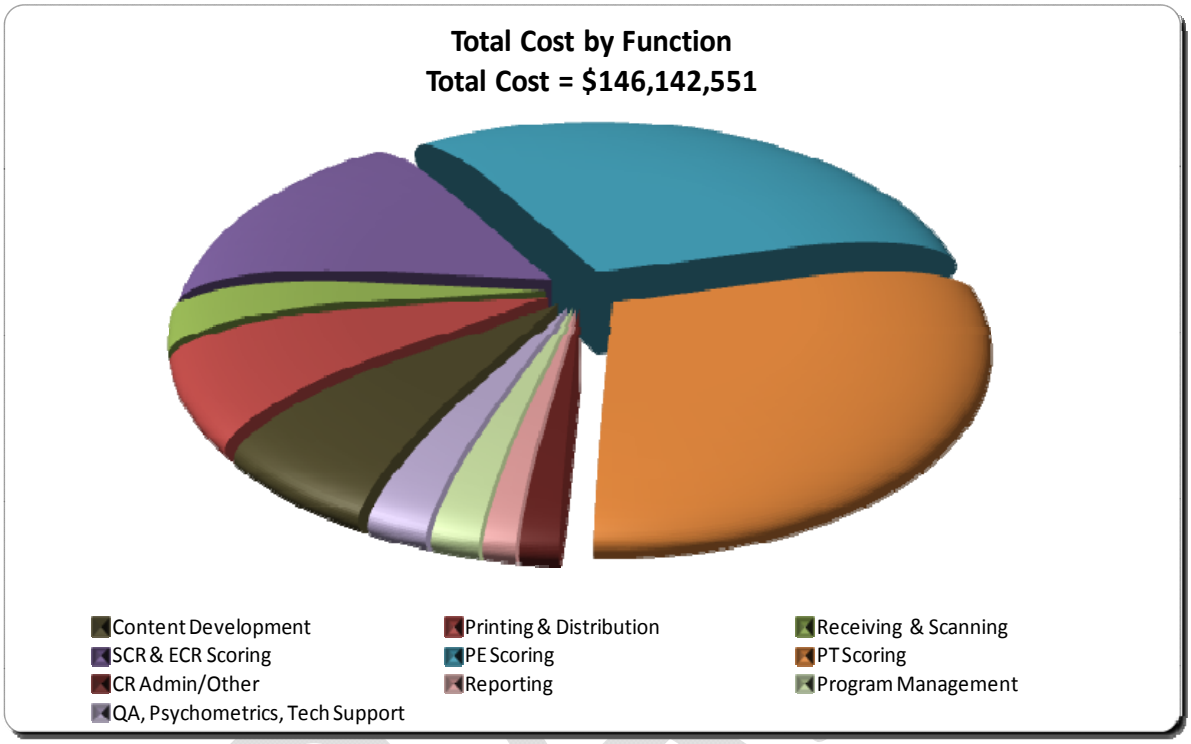
Table 5 indicates that the high-quality assessment program is more costly per pupil because the larger number of expensive types of items (constructed response and performance assessments) used in it. This is something that assessment program designers will need to consider as innovative approaches to assessment are considered in response to the common core development movement among the states.

Also, something interesting to note is that when the assessment design was changed to the HQA, roughly 12% was saved on development, printing and warehousing costs because the reduced number of MC items results in less development, less field testing and a smaller test book.

Another way of looking at the cost for this assessment model is cost by function. Figure 2 shows the costs for each portion of the assessment program. This allows the reader to see which functions are the most expensive in the high-quality assessment program. This serves to provide targets for cost saving measures, as explored in the set of cost-saving options that follow.

DRAFT

Figure 2. Cost by Function and Year



As can be seen, the single most expensive portion of this assessment program is scoring – of the short- and extended-constructed response items, the performance events, performance tasks, and administrative costs associated with scoring. The total of nearly $110 million for such scoring is more than three-quarters of the total cost of this innovative assessment program. This scoring cost is about $95 million more than in the traditional assessment program (shown in Figure 1 above). The other costly portion of this high-quality assessment program is content development, representing around five percent of the total cost, although content development is about $1 million less for the high-quality assessment model than the conventional one. As mentioned above, this occurs because by moving from 50 MC items to 25 MC items, the number of items to be developed and field tested is much less. The cost of procuring (not testing) a CR item - although about double a MC item (~$600 vs. ~$300) - does not offset the fewer MC items required to be developed. (Note, it costs $2,500 and $5,000 respectively to develop but not field test the PEs and PTs.)

A final way of looking at cost is by content area and grade, as is indicated in Table 6. This view of the assessment system shows the different costs associated with the English language arts (reading and writing) assessments, as well as the mathematics assessments, at each grade level.

Table 6. Cost by Content Area and Grade

Grade	Mathematics	Reading	Writing	Total
3	$1.93	$3.04	-	$4.97
4	2.98	2.99	2.46	8.43
5	3.97	3.24	-	7.21
6	3.93	3.25	-	7.18
7	3.96	3.18	3.22	10.36
8	3.93	3.24	-	7.17
10	3.93	3.17	3.26	10.36
Total	24.64	22.10	8.94	55.67

As mentioned above, grade level costs vary because at third grade, a scannable test booklet is used, while an answer folder is used at the other grades. In addition, there is a differential use of the innovative assessment types across the content areas. This will be reflected in the greater differences in costs between the content areas, as seen in Table 6 versus Table 3.

In conclusion, for the high quality assessment for the same state, some interesting things happen when comparing it to the typical assessment, as summarized below:

- Development costs actually decrease as the fewer MC items offset the additional cost of developing additional CR and PE/PT items.
- With fewer MC items, it was possible to eliminate a field test form and decrease the number of pages in the test book by 25% to 33%.
- Costs for production, manufacturing and warehousing costs are 12% lower than in Case 1. These savings are more than offset by the increase in CR scoring and addition of the scoring for the PEs and PTs.
- Scoring of CR and PE/PT items increases substantially and results in a much higher total and per pupil cost than Case 1, as expected.

3. <u>Cost Reduction Strategies</u>

As noted in the previous section, several potential cost-saving strategies were selected to determine if they could reduce the costs of the high-quality assessment model, perhaps so much so that it could be more

affordable to states. Each cost reduction strategy was examined separately, so its impact on total costs could be separately ascertained. However, it is anticipated that two or more of these strategies might be employed by states so as to minimize the costs of innovations in assessment design used by them.

3A. Participating in a state assessment consortium of 10, 20 or 30 states to share development and overhead costs

There are a number of costs associated with state assessment programs that are fixed or which do not vary greatly depending on the number of students assessed. The goal of a consortium is to avoid redundant activities so that these fixed costs can be shared among a larger number of states, thereby making the assessment more efficient. The total cost to any state would be less since these costs would be spread over a larger base – whether this is calculated in terms of number of states or number of students.

Partially offsetting the cost savings will be the expectation that some costs increase because certain activities for a group of states working together end up being more expensive than the same activities carried just in one state. For example, project management meetings called by a group of states in a consortium will involve greater travel costs, and may also involve increases in management costs (for administering the multi-state project). Hence, it is important to examine how costs would be affected by different sizes of consortia.

For sake of convenience, consortia that involve 10, 20, or 30 states working together were selected. The total number of students used for costing represented 20%, 40%, or 60% of the total number of students enrolled in public schools as reported by the U.S. Department of Education National Center for Education Statistics in 2009.

Data are reported in a similar manner as above for consortia of 10, 20 and 30 states with the costs for a single state shown again for convenience.

Table 7. Total Assessment Cost by Year and Consortium Size

No. States	Year 0	Year 1	Year 2	Year 3	Total Cost	Average Per State
1	$7,813,641	$45,562,943	$45,473,513	$47,292,454	$146,142,551	$146,142,551
10	$7,255,524	$220,534,504	227,580,504	236,683,724	692,054,256	$69,205,426
20	10,865,234	422,821,426	438,008,219	455,528,548	1,327,223,427	66,361,171
30	14,109,627	605,517,274	628,080,944	653,204,182	1,900,912,027	63,363,734

A couple of things can be noted from this table. First, the per-state cost charge goes down as more states join the consortium. This is to be expected as fixed costs are spread over more entities. Second, the per-state cost of all three consortia, is substantially lower than the cost of a state operating a comparable program by itself. For example, Table 4 showed that a single state would pay $146 million for a high-quality assessment program, while states working together can save substantially – between $75 million to more than $80 million per state over the three-plus years of operation. The costs for the innovative program when states work together come much closer to the cost of the conventional program - $52 million (see Table 1) versus $63-66 million (Table 7). Therefore, states working together in a consortium could save substantial money over working alone and could implement a much more innovative assessment program for not much more money.

The per-pupil costs for each size of consortium are shown in Table 8. This is a more direct method to compare costs, since it is the per pupil cost that would give an individual state a better idea of what such an innovative program would cost to operate.

Table 8. Per-Pupil Cost of Assessment, Per Year, by Consortium Size

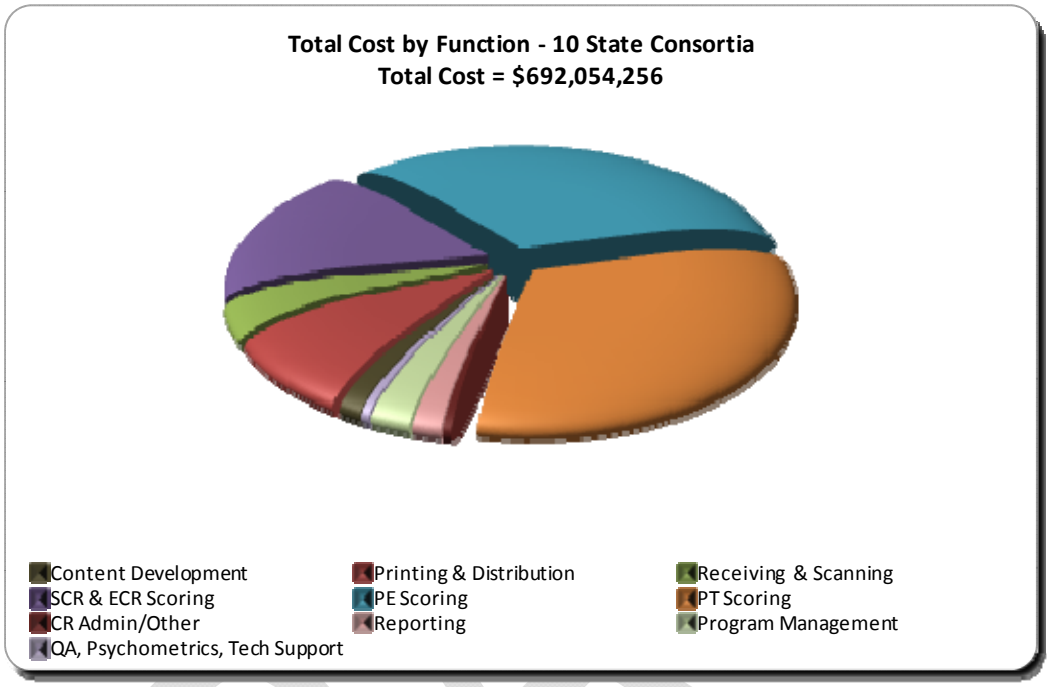
Size	Year 0	Year 1	Year 2	Year 3	Average Per Pupil
1 State	$8.93	$52.07	$51.97	$54.05	$55.67
10 States	1.33	40.54	41.83	43.51	42.41
20 States	1.00	38.86	40.26	41.87	40.66
30 States	0.86	37.10	38.49	40.03	38.83

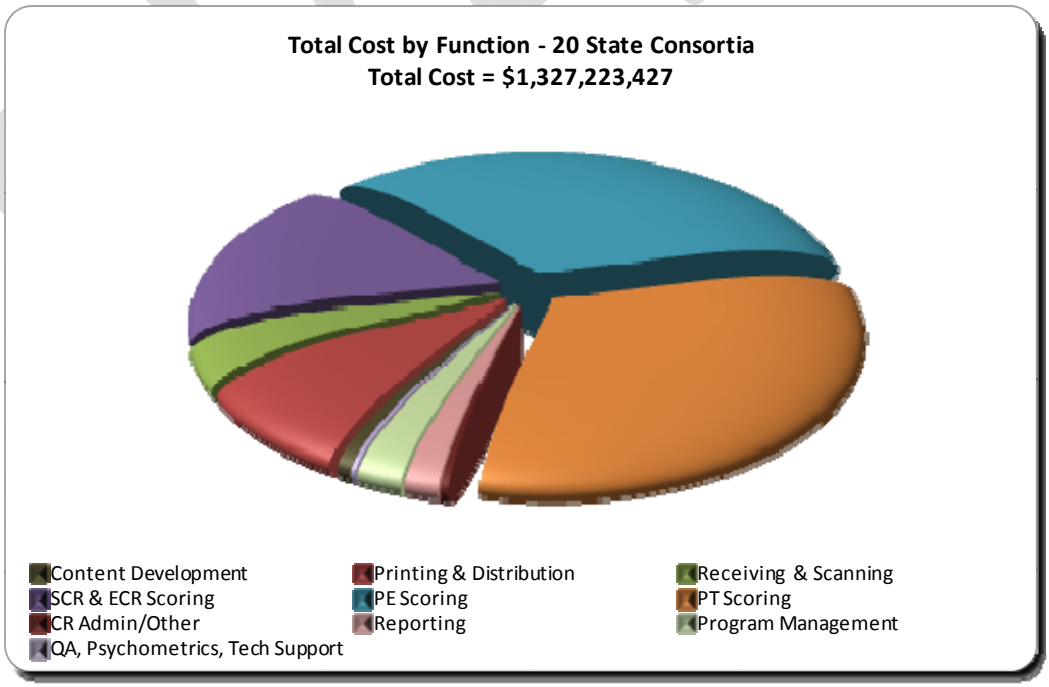
As can be seen, the net effect of a larger group of states working together is a savings of about $3.50 per student for working in a consortium of 30 states versus 10 states, and substantially greater savings (in the range of $10-$15 per student) than for a state to tackle this work on its own (see Table 6 for comparison purposes).

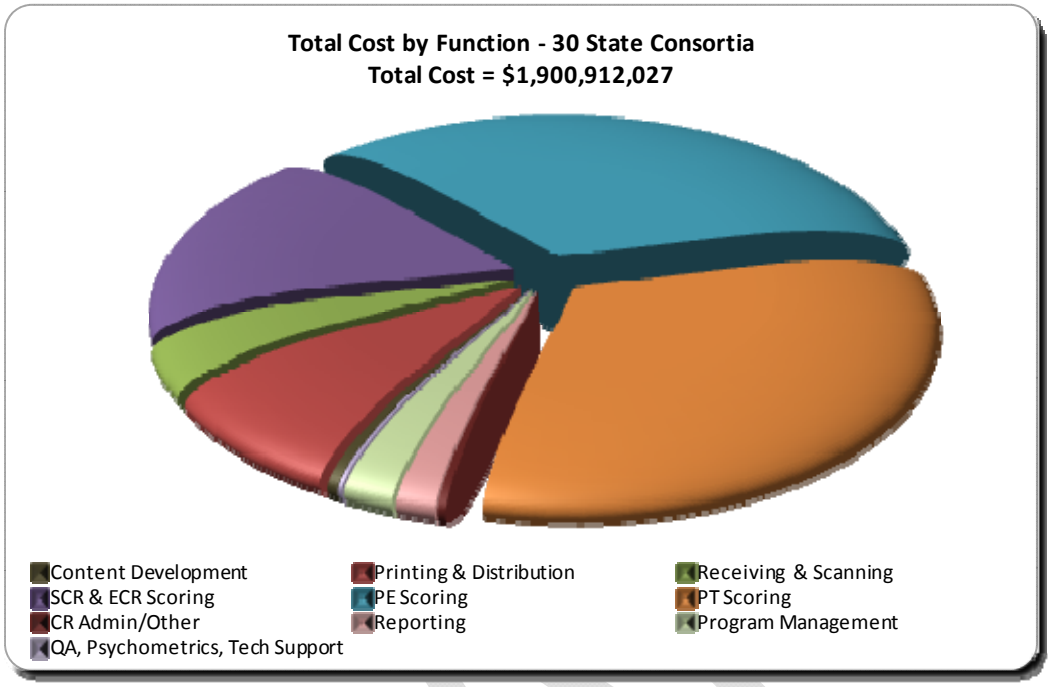
Costs for the three consortium sizes by function are shown in Figure 3 so that comparisons can be made on the amount of possible savings that can result by working with groups of states.

DRAFT

Figure 3. Total Cost by Function and Consortium Size







The pie charts show which functions are fixed or relatively fixed and those that are variable based on the number of students assessed. For example, the content development costs are almost flat (actually, go down slightly as the size of the consortium increases). Scoring costs, on the other hand, are directly proportional to the number of students assessed, so states working together in a consortium will have minimal impact on the costs of this function. This is largely due to price as the assumption was made that a consortia of 10 states will get a 5% discount, 20 states get a 7% discount and 30 states get a 9% discount from the base case for scoring costs. There are some additional costs that, if shared, would result in savings to states, for example, QA, psychometrics, and tech support.

Table 9 shows the cost per student for each content area for the single state and different consortia sizes.

Table 9. Cost by Content Area, Grade, and Consortium Size

Consortium Size	Mathematics	Reading	Writing	Total
1 State	24.64	22.10	8.94	55.67
10 States	18.93	16.58	6.91	42.41
20 States	18.15	15.88	6.63	40.66
30 States	17.34	15.16	6.34	38.83

In conclusion, for the high quality assessment developed and procured by state consortia, a few interesting things were noted, as summarized below:

- While the student count has increased by a factor of roughly 6.2 times, expense items increased by less than this amount, including those expenses that vary with the number of students. This is mostly due to improved efficiencies in printing, distribution and scoring (highly variable cost functions).
- Significant efficiencies were seen in the scoring of CR and PE/PT items. This is because of three factors: a) margin is lower by 5%+ as a consortium of states is able to negotiate better vendor pricing than a single state, b) the same number of students are needed to be field tested as consortia size increases in order to get valid results on these items, and c) the training component of the scoring is essentially a fixed cost.
- The QA, IT and psychometrics functions expenses did not increase much with the increase in the number of students.
- In total, per pupil costs decline from $55.67 to $42.41 in the 10 state consortia, a reduction of roughly 24%. The reduction is 36% for the 30 state consortia.
- Two operational forms and a breach form (that is not printed) of the exam were developed. Our assumption is that states will be comfortable using the same assessment form and will work through the security issues involved in such a situation rather than developing separate forms for their individual use. Developing and using more than 2 operational forms would increase development and production costs significantly.

Overall, it is clear that states working together in developing and implementing a common, high quality assessment program can do so at substantial cost savings to each participating state. That said, however, states working together will not be able to save enough from this action alone to bring the costs of innovative assessments in line with current expenditures on assessment.

DRAFT

Note that in actual operation, states might choose to share fixed costs on a per-state basis (e.g., a consortium of 10 states would divide the fixed costs by 10) while the variable costs might be shared on a per-student basis. This could result in slightly different "per state" costs than are shown here.

3B. Moving to online delivery of the assessment to reduce production and shipment costs

The next cost option explored was the online delivery of the assessment by computer. The assumption was made that all students in each state participating in a consortium would be assessed in this manner, with the only exception being a few students with disabilities for whom a paper-based assessment would be an appropriate accommodation.

For costing purposes, the same high-quality assessment program was used, and the assumption was made that the scoring would be carried out by the vendor using their trained human scorers. Table 10 shows the costs associated with this assessment delivery method.

Table 10. Assessment Cost by Delivery Method and Consortium Size

Consortium Size	Total Assessment Cost	Online Per-Pupil Cost	Paper Per-Pupil Cost (3A)
10 States	$663,287,152	$40.64	$42.41
20 States	1,240,224,116	38.00	40.66
30 States	1,730,018,897	35.34	38.83

As can be seen, the cost of assessing students online is less than paper-based testing, as shown in the comparison of costs between this table and Tables 7 and 8. The net per-pupil savings is about $2.25 to $3.50 per student.

In conclusion, for consortia with online administration, some savings were seen from the move to online assessment, as summarized below:

- The savings are a bit less than expected as moving to the HQA eliminated 12% of per student pencil and paper costs and therefore less potential savings when moving to online test administration.

- Online system costs (including a prime contractor management fee) of $3.72 per student in the 10 state consortia, $2.63 in the 20 state consortia, and $1.58 in the 30 state consortia were assumed.

- The savings for moving to online assessment are still significant as a percentage of total non CR and PE/PT scoring costs.

Therefore, moving to online assessment within the context of an assessment consortium of any size will save states some money, but the savings are not substantial (where an assessment includes vendor scoring of performance items) – certainly, not larger than the cost savings of simply working together.

3C. Using teachers to score performance event (PEs) and performance task (PTs) items

Another option is for local educators to score the performance events, and performance tasks. This approach has been used successfully in some testing programs and by a variety of countries. The first table, Table 11, shows the costs when teachers are doing this as a professional development activity. A number of studies have noted the strong learning benefits of teachers' involvement in scoring, and thus one could argue that this activity could be undertaken as part of state professional development budgets. In this case, the costs for teachers' time in carrying out the scoring tasks are not associated with the assessment contract and not included in the cost of using an assessment vendor. Table 12 shows (just for a consortium of 10 states) what the assessment would cost if teachers were paid a $125 stipend per day for scoring the performance events and performance tasks. These costs were provided to look at the case where payments to teachers would be necessary for them to participate in scoring the assessments items requiring human scoring.

Table 11. Assessment Costs When Teacher Scoring as PD Activity is Used

Consortium Size	Total Assessment Cost	Teacher PD Scoring Per Pupil Cost
10 States	$305,198,877	$18.70
20 States	520,475,313	15.95
30 States	713,554,967	14.57

This table shows that the cost of the assessment program would be substantially lower if states' teachers are used for scoring and if the cost of this scoring were associated with the professional development budget, not that of the assessment program. These costs decline somewhat in consortia that involve more states.

Table 12. Assessment Costs When Teachers are Paid $125 Stipend Per Day

Consortium Size		Total Assessment Cost Including Teacher Stipend		Per Pupil – Teacher Stipend Cost
10 States		$508,635,610		$31.17
30 States		$1,258,768,591		$25.71

This table shows that the costs of the assessment programs for consortia of 10 and 30 states are higher than the previous scenario, but much less than the costs for vendor scoring of the performance items (Table 8) if teachers are paid a stipend of $125 per day to score the performance events and performance tasks.

In conclusion, for the approach of using teachers to score PE and PT items, significant cost savings were found, as summarized below:

- Using teachers to score PE/PT items as part of their compensation (professional development) makes a huge difference, since the cost of scoring these items is "free", which reduced costs to $18.70 per student in the 10 state consortia and $14.57 in the 30 state consortia.
- Paying teachers a $125/day stipend still saves significant costs and results in an assessment cost of $31.17 per student at the 10 state consortia size and $25.71 at the 30 state consortia size.
- Key assumptions made in the teacher scoring case were 1) the scoring would be distributed so there are no facilities, management overhead costs and 2) the vendor would accept a lower OH and profit margin on this work (10% OH and 15% margin) based on the $125 teacher stipend amount.

These analyses show that if the states are able to have teachers score the assessment items that require it without having to pay them a stipend because the costs are considered to be a professional development cost, they can save considerable amount of money (while giving their teachers the positive experience of learning to score the assessments and thus improving their understanding of student learning). With very large consortia, the $25.71 per student cost of the assessment (assuming a teacher stipend of $125/day) is at the mid to high end of the range of today's high stakes assessments.

3D. <u>Using distributed scoring for constructed response items</u>

The next potential cost saving strategy is to use distributed scoring for scoring the constructed response items. This arrangement is used to minimize costs since scorers are working from home and the assessment contractor does not need to provide a bricks-and-mortar facility with its own computer equipment to score these items. Scorers provide their own equipment and connect to the contractor and score the items digitally online. This scoring is carried out in a secure manner.

Table 13 shows the impact of distributed scoring used in consortia involving 10, 20 or 30 states. It assumes that half of the student responses are contractor scored in one of its facilities, and the other half are scored through distributed scoring off site.

Table 13. Assessment Cost Using 50-50 Distributed Scoring by Consortium Size

Consortium Size	Total 50-50 Assessment Cost	50-50 Assessment Cost (per pupil)	All On-Site Contractor Cost (per pupil) (3A)
10 States	$679,802,413	$41.65	$42.41
20 States	1,303,450,391	39.93	40.66
30 States	1,866,192,174	38.12	38.83

As can be seen, there are slight cost savings for using distributed scoring when compared to scoring by contractor staff fully on-site at the contractor site(s), as seen in the final column (taken from Table 10).

In conclusion, the approach of using distributed scoring at a 50/50 on-site vs. distributed mix had less impact on assessment cost than expected due to the relatively low percentage of scores and costs represented by the CRs once the PE/PT scoring costs were included. The cost data for this option are summarized below:

- Generally, distributed scoring is expected to be about 25%-30% less expensive than on-site scoring for those items that are scored using the distributed model.
- Since a 50/50 mix was used, an 11% reduction in the total CR scoring costs was seen.

Thus, given the small magnitude of this cost option, it may not be worth pursuing further in a high quality assessment design with performance items. However, if a new assessment were to include more constructed response items (than currently used) and no performance items, distributed scoring can make sense.

3E. <u>Automated scoring for some constructed response items</u>

The final cost-saving option for the large-scale summative assessments is to use computerized artificial intelligence software to score student responses to constructed response items (excluding PE/PT items). Such software has been studied extensively in recent years and has been found to produce relatively comparable results to hand scoring by humans. It is more effective with some sorts of student written work (such as extended constructed response items or essays) than others. Thus, a mixture of hand scoring and computer scoring was costed. (Note – to make such assessment scoring truly efficient, the essays to be scored should be entered into the computer via online assessment.)

Table 14 shows the financial impact of using computer-based scoring with a consortium of 30 states.

Table 14. Assessment Cost Using a Mixture of Computer and Human Scoring of Written Response Items

Consortium Size	Total Assessment Cost	Per-Pupil Cost
30 States	$1,855,328,550	$37.90

As can be seen above, there is some cost savings from using computer artificial intelligence software to score a portion of the written-response assessments.

For the use of AI scoring of CR items, the data are summarized below:

- An "engine tuning" cost of $6,000 per item was assumed to load the responses into and train the scoring engine.
- Even at $0.50 per response for scoring, which is the bottom end of the price range, significant cost savings were not obtained vs. using human scorers. A different AI vendor (who did not supply a quote) noted that 'it is a myth that current costs for online scoring of constructed response questions is less expensive than human scoring.'
- It is worth noting that online scoring is significantly faster than human scoring. Whereas human scoring of a typical state assessment takes several weeks, online scoring can be accomplished in a day or two, once the system has been set up.
- Costs for AI scoring need to decrease further to make this option practical. However, this is a very valuable future potential savings source and "we" need to stay on top of it.

The minimal cost savings found for use of artificial intelligence software to score some of the students' written work is surprising. At the time of the issuance of this draft version of the paper, ASG is continuing its research in this area to determine if solutions from other vendors and organizations, that have taken different approaches to AI scoring, can be implemented in the high stakes testing arena. In any event, as systems mature and prices come down (particularly the costs of "training" the system) these systems will undoubtedly warrant further investigation.

3F. <u>Development of a customized interim benchmark assessment system with similar item types and structure as the high quality system</u>

One additional scenario was examined – that of developing a customized set of assessment tasks to be used by local school districts as interim benchmark assessments. The costs shown in Table 15 include only the cost of developing the assessments, not the costs of administering, scoring and reporting them. These functions are not necessary if scoring is locally managed and results are used for classroom

information and local instructional development purposes. The consortia are shown for consortia of different sizes.

Table 15. Interim Benchmark Assessment Development Costs

Consortium Size	Total Assessment Cost	Per-Pupil Cost
10 States	$4,623,736	$0.85
20 States	$4,974,680	0.46
30 States	$5,329,608	0.33

This table shows that the cost for a consortium of states to develop and interim benchmark assessment bank of tasks is very inexpensive. Of course, to these costs, the states would need to add any assessment administration, scoring, and reporting costs if the state(s) chose to carry out these activities for local school districts (and absorb the costs).

In conclusion, the development of a customized interim benchmark assessment set of tasks would be inexpensive for a group of states to carry out.

3 F(2) Interim Assessment Administration Options

The second portion of the interim system analysis conducted by ASG was to determine the cost to provide an interim benchmark system for use among the states participating in a consortium. When states form consortia to develop and implement large-scale assessments at the state level, they may also wish to consider how they could provide interim benchmark assessments to their local school districts. For cost purposes, consortia of 10, 20 and 30 states were used. Four options for interim benchmark assessments at the state level were examined. These are:

A. The full consortium buys/leases a complete system (items and online delivery system) from a vendor and this system is provided to local districts to use as they see fit.

B. The consortium purchases/leases an online assessment system from a vendor, but the consortium loads its own assessment (which it has developed) into the system and provides the system for local district use.

C. The consortium develops its own assessments and administers these and the state assessments using the same online system that they have either created or leased.

D. The consortium develops its own assessments and provides these to the local school districts to use as they see fit – to load into any online system that they have and/or to use as paper-based assessments.

In order to assist states working in consortium to understand the costs of these options, cost estimates were prepared to show what options A-C would cost per student and per state. Each is shown below.

A. Consortium Buys/Leases System and Provides to School Districts

Table 16 shows the costs for selecting an existing system to deliver interim benchmark assessments, including the use of the vendor's assessment items.

Table 16. Consortium-Provided Interim Benchmark System Cost

	Total Cost	Average Per State
10 States	$130,559,184	$4,351,973
20 States	$220,478,572	$3,807,976
30 States	$293,758,164	$3,263,980

This table shows that a state-provided interim benchmark assessment system is not inexpensive. However, offsetting the cost is the speed with which an existing system can be installed and used.

B. Consortium Buys/Leases Online System, But Builds Own Assessments

The next scenario shows the costs for an interim benchmark system in which the consortium of states leases the online assessment system, but the assessments are provided by the consortium itself, not the vendor. Table 17 shows the costs for this system.

Table 17. Consortium-Leased Online System with Consortium-Developed Items

	Total Cost	Average Per State
10 States	$97,919,388	$3,263,980
20 States	$163,198,980	$2,719,983
30 States	$195,838,776	$2,175,986

This table shows that if states created their own assessments, there would be a modest cost savings to the states. Whether the savings are large enough for states to elect this option would be a matter for the consortium to determine.

C. Consortium Builds its Own Assessments and Uses the State Assessment System to Deliver the Assessments

This option is for the states to create the interim benchmark assessments on their own, and then to use the same online assessment engine used for the state assessment program to deliver the interim assessments periodically. These costs are shown in Table 18.

Table 18. Consortium Developed Assessment Delivered by Online State Assessment System

	Total Cost	**Average Per State**
10 States	$32,639,796	$1,087,993
20 States	$48,959,694	$815,995
30 States	$48,959,694	$543,997

As can be seen, the added cost to administer the interim assessments using the same system as used for the state assessment program is not as great as using a separate system. This is one way that a consortium of states could save money.

D. Consortium Builds Own Assessments and Local Districts Use As They Desire

The fourth and final option is for a consortium to build their own assessments and turn these over to local districts to use whenever and however they desire. This would mean some districts might choose to use them electronically, some might use them in paper-based systems, and others choose not to use them at all. This is the least expensive of the four options, since no test administration, scoring, and reporting services are provided. However, it is also the most flexible option, since local districts are not obliged to use the interim benchmark assessments at a particular time or manner.

Due to the wide variety of options (state providing camera ready proofs, state providing digitized data, state providing hard copy assessments, etc.) we did not price a scenario for this option

3 A-E. Use of All Cost-Saving Measures Together

In the end, the use of all of the assessment cost options that could affect the overall state assessment costs for a consortium was also computed. This information is shown in Table 19 for the summative assessment.

Table 19. Assessment Cost Using All Cost Savings Measures

Size	Year 0	Year 1	Year 2	Year 3	Total	Per-Pupil
30 States	$6,266,215	$160,576,640	$161,427,363	$163,969,135	$492,239,352	$10.05

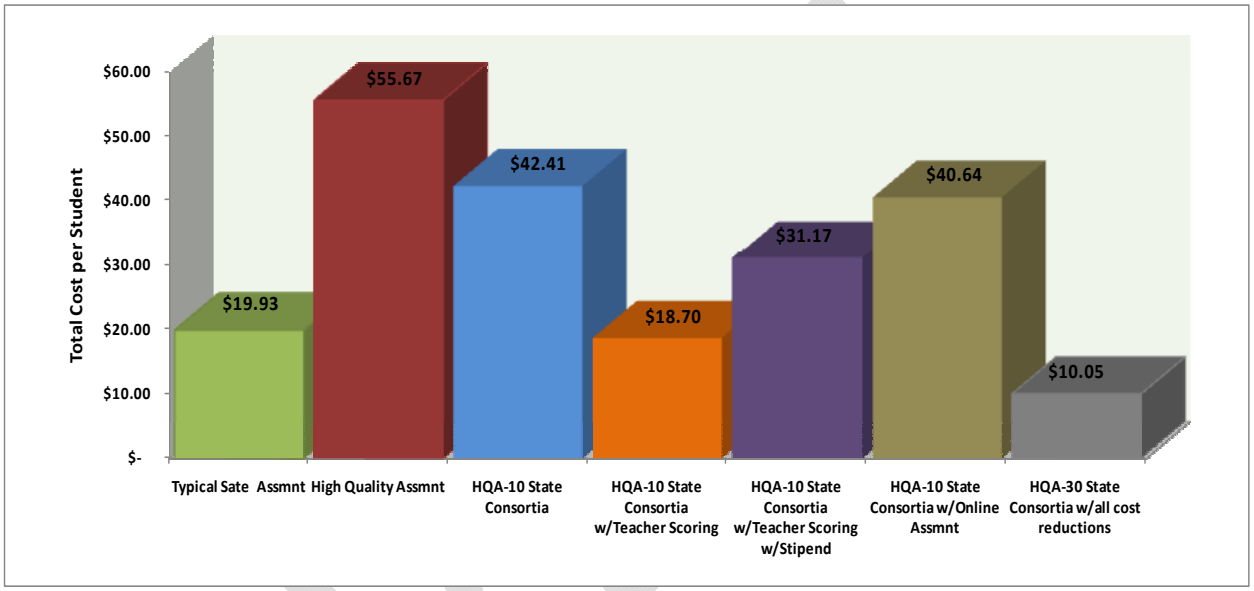
Table 19 shows that if a consortia of states were to adopt all of the cost reductions strategies described above, the per pupil cost of the summative assessment would be $10.05 vs. $19.93 for a single state's current summative assessment system. Adding the costs of an interim assessment system to this total would still allow a consortium of states to offer a rich, performance based system of assessment for less than most states are spending today for tests that offer much less utility for supporting intellectually challenging learning and information for instruction.

DRAFT

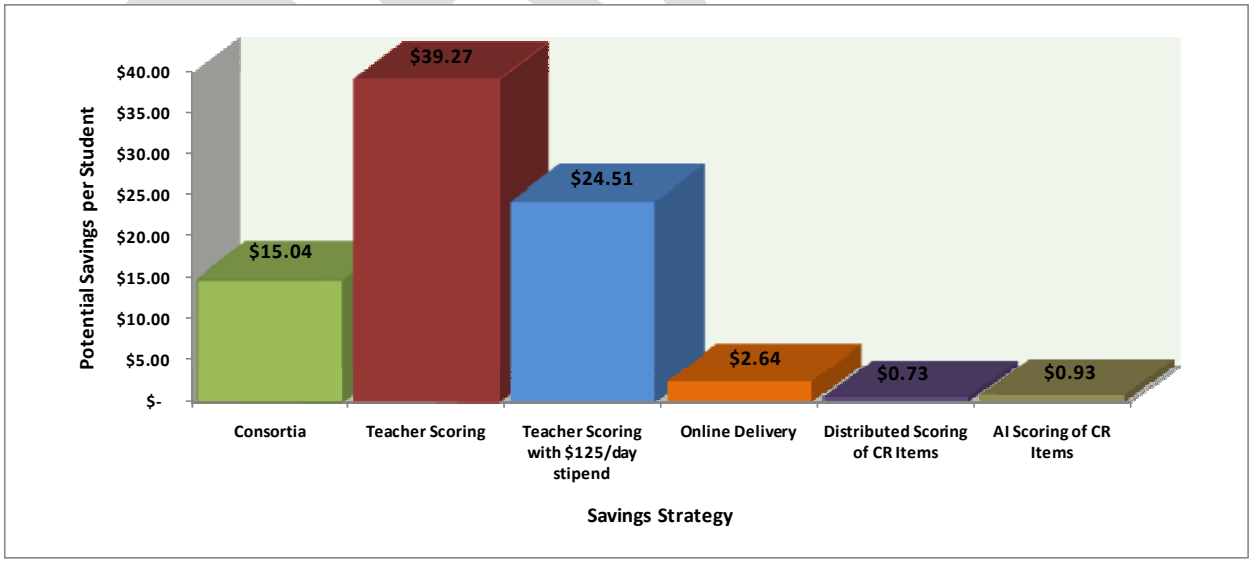
V. Summary, Conclusions, Discussion, and Recommendations

Overall Summary

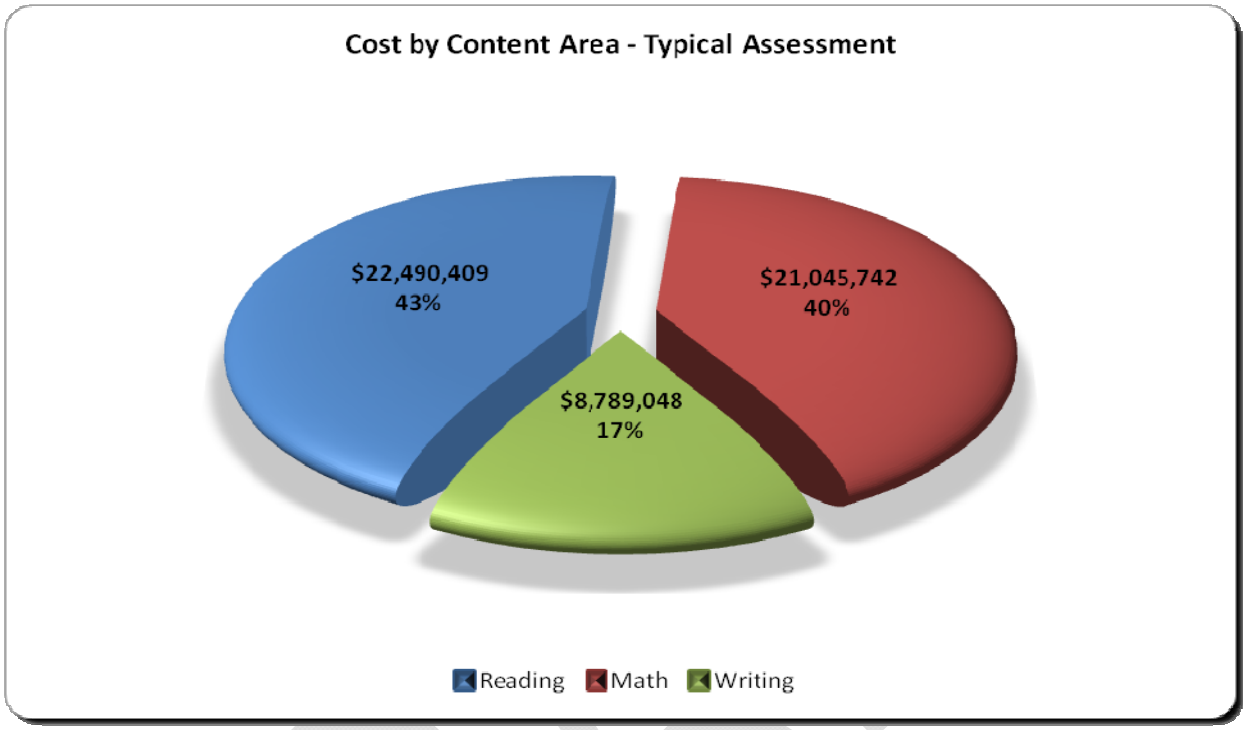
In this study, a series of analyses were conducted to model the costs of various assessment designs and approaches for implementation. The following chart shows the cost per student for seven of the different models that were analyzed.

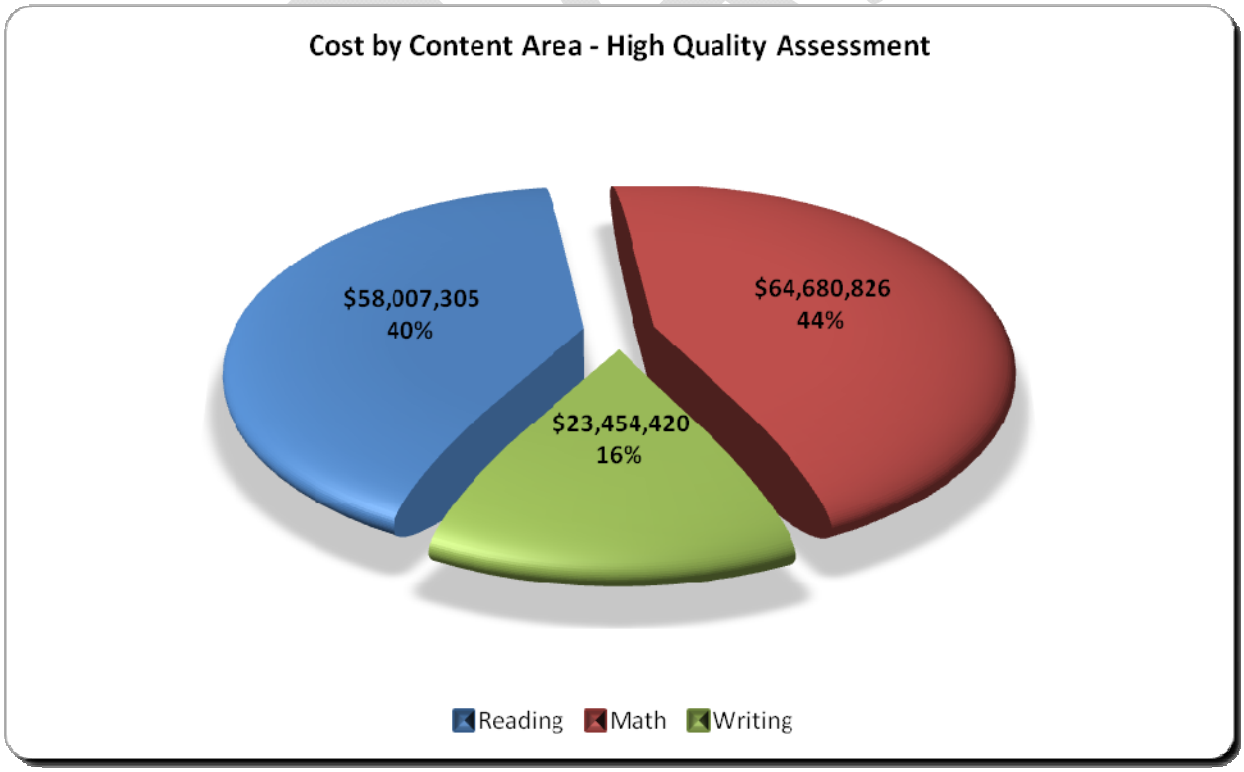


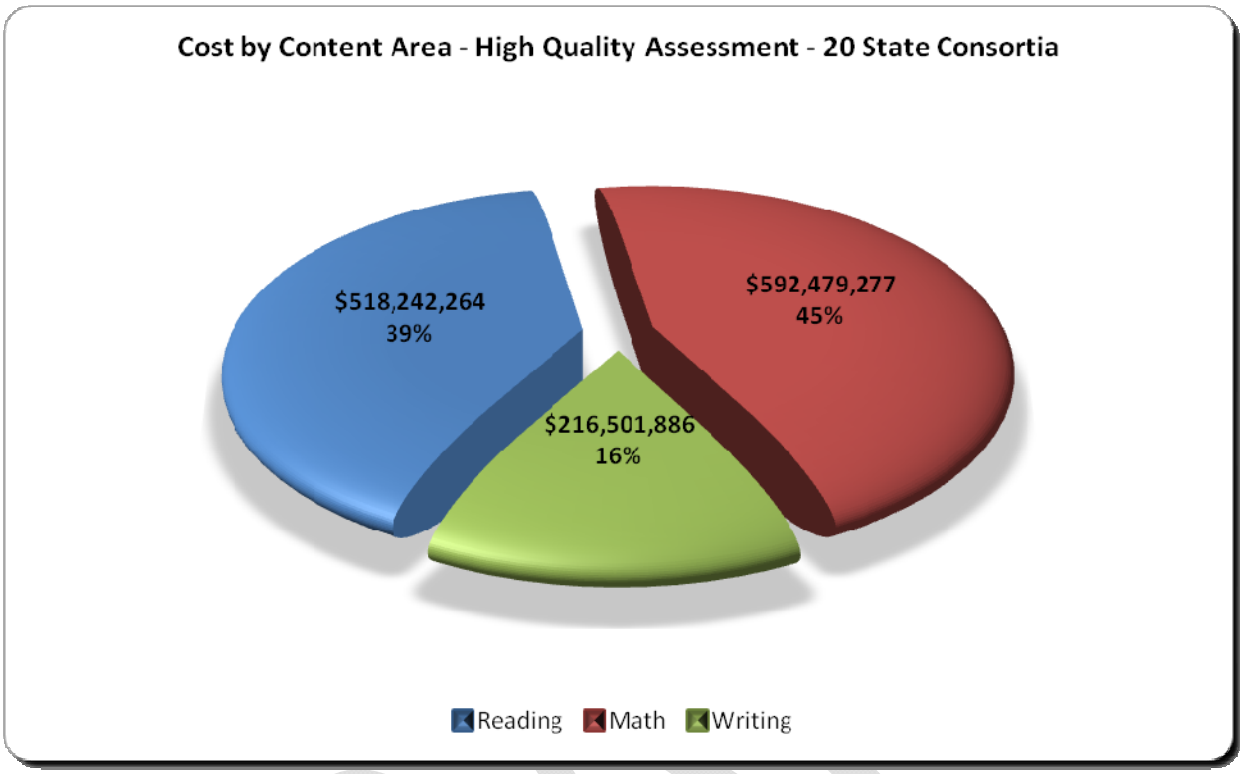
The next chart shows the cost savings for six of the possible options that were analyzed.



The following charts show costs by content area for the different assessment models.







Conclusions and Discussion

Based on the findings from the analyses presented in Section IV, the following conclusions can be made from this study.

Need for Innovative Assessment Approaches

- State assessments need to be improved so that they do a better job of measuring the critical skills students will need in the 21st century, are integrated into the curriculum, help students learn, and provide teachers opportunities to develop new instructional strategies. An interim assessment system is an important part of any balanced assessment system and should include the same item types as the summative system.
- There are many worthwhile ideas about how new, high-quality, innovative assessment systems might be designed and constructed. This study looked at one such model, as well as possible variations on approaches that could be used by states, and analyzed ways to make that model as

cost efficient as possible. As states decide to work together to design and implement new, high quality assessments, these states would be wise to examine the costs for their consortium design, so as to make certain that they have designed the most efficient and cost effective program possible.

Costs for Implementation

- It was found that the development cost of a new high-quality assessment is relatively inexpensive relative to the total cost of the assessment. **However, a key factor in the sustainability of new improved assessments and whether or not states can adopt and use them will be the ongoing administration costs that need to be managed.**
- Of the cost reduction strategies that were examined, teacher scoring (with and without stipend) and formation of consortia were the two most important means of controlling overall assessment cost. While there are issues to be tackled regarding the implementation of teacher scoring (including determining the opportunity cost of forgone teacher time) and the formation of state consortia, both strategies should be a part of the plans for any future assessment system.
- Current systems exist for interim assessments and the costs of these systems can likely be reduced substantially if procured by a consortia of states with new content already developed (assuming teachers score the performance items). Combining the purchase of an interim assessment system with a summative system provides the largest interim assessment cost savings opportunity.

State Consortia

- In order to reduce costs across states, it will be important to have states participate in assessment consortia to share the overhead associated with development and management of assessments. Larger consortia are more cost effective, although the majority of cost savings relative to a single state case can be seen at the 10 state consortia size.
- Implementing a high-quality assessment system with performance items is affordable, with teacher scoring of performance items, at a price comparable to today's assessments when procured by a consortium of states.

Scoring

- In order to implement and afford a high-quality assessment system that includes a variety of performance items, it will be essential to have teachers involved in the scoring process. Financially, and logistically, the scoring model currently used in most states (vendor does all scoring) could be a challenge for states and/or state consortia in the future. The total amount of money for outside scoring and the sheer number of scorers that would be required to mark all the answers could be difficult, if not impossible, to find and manage. Many countries with high performing educational systems involve their teachers in the scoring of performance items and integrate that part of the process in the curriculum as a professional development activity. State consortia that purchase assessments and pay teachers a stipend to score these responses will be able to develop and administer an assessment at the high end of the range of prices seen today.

Other Cost Reduction Strategies

- The use of online technology (i.e., online assessments) should be encouraged as it also has the potential to reduce assessment cost and improve quality. The size of the cost reductions that were calculated assuming implementation of an online assessment was not as large as was expected. However, this is somewhat related to the assessment design (only 3 subject areas, an efficient design, small test books) and would undoubtedly be larger given different design parameters. The procurement of PCs to improve the student to PC ratio should be encouraged at all levels of the educational system.
- Ultimately, the use of artificial intelligence systems to score essay type responses holds tremendous potential value for the future affordability of high-quality assessments. Today, AI scoring systems are often too expensive for states and the cost of scoring CR questions by this method is about the same as using human scorers. As systems mature and costs come down, AI scoring systems offer tremendous advantages in the delivery of results from performance assessments.

Finally, in our opinion, the RTTT and Common Core Assessments are key initiatives in improving education, and states have an opportunity to receive some much needed resources and assistance to help them make important improvements to their assessment programs. The research conducted for this study and results reported in this paper demonstrate that under the right conditions, states can dramatically improve their assessment systems at an affordable cost. However, states must be careful to design an efficient assessment system and understand its ongoing administration costs, as well as future state budget allocations prior to committing to an innovative, high-quality assessment and implementing it in their state. States also will need to think through the various management issues when forming and working with a state consortia as well as using teachers to score performance items. Professional help in all these areas is highly recommended.

Final Recommendations

Based on the data from the cost models and the conclusions listed above, ASG makes the following recommendations:

- Developing and implementing a high-quality assessment will likely cost more than most current state assessments, but it can be affordable for states if they look carefully at the design, find a balance in the number of CR items, PEs and PTs that are used, and consider various cost reduction strategies.

- States should strongly consider being part of a large consortium where certain costs can be shared across many states, such as for item development and project management.

- States should consider using a scoring model that has teachers scoring the performance items as part of their professional development via a distributed scoring system. Having all scoring done

by the testing vendor is likely to be cost prohibitive for most states. Paying teachers a stipend also helps reduce costs, but not as much as using a PD approach. In either case, there are benefits for teacher learning and instruction, as well as cost savings, associated with teacher scoring.

- States may want to consider moving to online assessment, as it can be more cost-effective than printing test booklets and shipping them to schools. Although the initial savings may not be as large as thought when implementing a HQA design, in the long run, online assessment will save states both money and time. Many states feel that current PC to student ratios of 4 or 5 to 1 make it difficult to implement online administration of assessments. Policies to help states procure additional PCs and bandwidth for schools should be encouraged.

- Ultimately automated scoring of essay responses should lower scoring costs for these items significantly and further enable the implementation of high quality assessments at reasonable prices. AI scoring should be encouraged and its progress monitored.

- States should consider examining the costs for their future assessments in more detail and look at different options that make the assessment both higher in quality and more efficient. For example, states may want to design an assessment that has many more CR items, no PTs, and uses an AI scoring engine to score all items. There are many variations on the possible designs that could be used by states, and all have different cost implications.

- State consortia interested in implementing a higher quality assessment need to make sure they can afford the ongoing administration costs of the assessment. It is recommended that state consortia go about the process of developing and costing a new assessment in a thoughtful manner and use a comprehensive costing model to analyze and determine the price in advance of any new assessment system they would like to implement.

Possible Future Research, Additional Analyses, and Other Studies of State Assessment Costs

Obviously, there are many ways to design a high quality assessment system. The designs selected for this study were based on input from a variety of assessment experts. In some ways, the cost models done for this study are just a beginning, and provide valuable information to begin a more in-depth discussion of what potential costs could be for various approaches. It is hoped that additional cost models will be done based on other ways to design an innovative assessment system. Analyses can include many other kinds of assumptions and variables. These could include variations in the number of CR items, different percentages of items released each year, and changes in program components and features. Information from these types of analyses can be helpful for a state, or consortium of states, to further reduce their costs while maintaining other core components of the assessment system.

In this study, costs were analyzed only for reading/language arts and mathematics assessments. These areas were selected because of their relationship to NCLB and the current plans for common core

standards and common assessments. In the future, it would be useful to run cost models of other content areas since many states also assess their students in science, social studies, the arts, and other disciplines. In addition, cost analyses of alternate assessments for students with disabilities and English language proficiency assessments for English language learners would be useful in order to determine how to design these assessments more efficiently, as well as reduce costs. It may well be possible for consortia of states to work together to also create and implement these additional assessments. In addition, some states may also want to analyze costs for their end of course tests or other high school examinations.

Ultimately, the possibilities for productive assessment will be enhanced by these kinds of analyses. As we show here, by taking advantage of collaboration, technologies, and judicious design decisions, states can offer a rich, performance-based system of assessment that supports high-quality instruction for less than most states are spending today.

DRAFT

References

Assessment Solutions Group. *"Cost Estimate for the New Kentucky Assessment System."* Prepared for the Kentucky Department of Education. Author: 2009.

Darling-Hammond, Linda. "*Developing Assessment Systems that Support High-Quality Learning*." Stanford University: Author. 2010.

Education Sector. "*Margins of Error: The Education Testing Industry in the No Child Left Behind Era."* Washington, DC: Author. 2006.

General Accounting Office. "*Title I – Characteristics of Tests Will Influence Expenses; Information Sharing May Help State Realize Efficiencies."* Washington, DC: Author. 2003.

Jackson, J. Mark and Eric Bassett. "*The State of the K-12 State Assessment Market*." Boston, MA: Eduventures. 2005.

[Note: Add other ASG study references here, plus the references noted earlier to cost estimates that inform the assumptions in this document.]

DRAFT

Appendices

About the Assessment Solutions Group

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

Uses of the ASG Cost Model and Additional Services

The ASG cost model also can be used to evaluate state assessments for efficient designs. The model can estimate costs for current programs as well as possible new ones states may want to develop in the future. In addition, costing services can assist states in structuring their RFP documents and evaluating vendor responses in order to ensure that proposed assessment programs can be evaluated on a comparable basis, are reasonably priced, **and** include sufficient quality safeguards. In addition, ASG can provide temporary staffing to state departments of education and other education related institutions. Specialists in assessment program design, test development, assessment administration, program management, budgeting and accounting, technical support, psychometrics and analysis and IT work can be made available to help meet staffing needs.

For More Information

For more information about ASG, go to www.assessmentgroup.org

To contact the authors, please send email to:

- Barry Topol, btopol@assessmentgroup.org
- John Olson, jolson@assessmentgroup.org
- Ed Roeber, eroeber@assessmentgroup.org

DRAFT

UNITED STATES DEPARTMENT OF EDUCATION
RACE TO THE TOP ASSESSMENT COMPETITION
PUBLIC AND EXPERT INPUT MEETINGS
HIGH SCHOOL
Friday, November 13, 2009

Testimony of Nicholas C. Donohue

MR. DONOHUE:
Good afternoon. My name is Nicholas Donohue. I'm the president and CEO of the Nellie Mae Education Foundation. Welcome to New England. We are the largest regional philanthropy dedicated to education in the region, and we're involved in a variety of interesting investments around assessment at the state and regional level.

I also bring with me my experience as a chief state school officer from the State of New Hampshire, at the inception of the No Child Left Behind, and as an original founder of the New England Comprehensive Assistant Program known as NECAP. My input comes in two short parts.

One is that I encourage you to take the time and energy to really dig into the learnings from NECAP and other state collaboratives in ways that aren't possible at a hearing like this. They are technical, political issues that affect any collective effort, and they will -- they would be good to learn from.

Second when awarding grants to consortia I would recommend a set of principles to find the expectations you might have. One is that you define a set of core purposes and expect the collaborative to respond to these. Two, expect them to pay attention to variety as a principle, design principle. Three that you expect them to actually deliver design principles that define and support the independent components of the system, a multilevel system, but also of the integration of that system. And I'll say something that in a moment. Four, you obviously expect them to model implementation on the ground in ways that are interesting and practical. Five, you specifically expect them to deliver cost -- learnings about cost and cost effectiveness of a varied system, and six, that they attend to familiar principles around quality and fairness and equity.

On the issue of core principles, I'd like to suggest that the department should state clearly a set of principles that collaboratives much achieve. Those principles should include that the assessment systems that they design and model should allow meaningful international comparisons. They should obviously equip students with the knowledge and habits in line with dispositions to succeed. These systems should promote improved schooling in ways that provide students with engaging educational opportunities. These systems should also provide parents and students and other stakeholders with the information necessary about relative performance.

A second principle, around attention to variety is rooted in the notion that if you have a variety of core purposes you need a variety of approaches. The previous speaker talked about the difficulty of achieving a wide variety of purposes; however, in a well-designed, intentionally designed system that included national, regional, state and local components, it is possible to bridge a variety of purposes. Variety also attends to our commitment such as around college-ready, how do we promote college-readiness without actually assessing it and it also supports international comparisons as we've heard.

A third principle is that you expect design specifications from these collaboratives that they actually tell you very concretely about what it takes to build independent components at the local, state and regional levels, and most importantly, that they tell you exactly what it takes to integrate these components, that they demonstrate specific findings around quality control mechanisms across state lines, that they describe specifically professional development systems that support the participation of educators, and significantly, that they describe the systems that promote public understanding and engagement in the development of the assessment systems.

The days of cobbling together parts of a system and pretending it's a system need to end. The intentional system design needs to be a priority. I mentioned modeling a system. This is obvious, I think, providing core principles and demanding the collaboratives demonstrate on the ground examples of how you would implement this, maybe in a subset of schools in and across a region would seem to be sufficient. Lastly, cost estimates need to be provided. The current conversations around cost are too much just about dollars. We all know there are other collateral costs related to assessment in terms of time and attention to learning, and these collaboratives could give you good information about what some of those collateral costs are, and a true picture of cost-effectiveness.

Any award ought to attend to all these principles. I will close by saying that the key is to provide guidance and specificity and purposes and allow flexibility in implementation. Thank you.

MS. WEISS: Thanks so much.

MR. DONOHUE: You're welcome.

Dear Secretary Duncan,

Thank for you for the opportunity to submit comment on the competitive Race to the Top Assessment Program to support one or more consortia of states that are working toward jointly developing and implementing common, high-quality assessments aligned with a consortium's common set of kindergarten-through-grade-12 (K–12) standards that are internationally benchmarked and that build toward college and career readiness by the time of high school completion. Our brief comments below and attached materials reference the major categories of questions the Department is seeking input on.

First, please find attached the testimony I provided to the Department at the public meeting in Boston, Massachusetts on November 13, 2009 on the Race to the Top Assessment competition. That testimony included some remarks on the Project and Consortium Management questions, as well as the General and Technical Assessment questions. I also provided testimony on these two categories of questions at the public meeting in Washington, DC on January 13, 2010.

Second, please find attached a document titled *Principles for Student Assessment* that speak directly to the questions on General and Technical Assessments the Department posted for the public meeting being held today in Washington, DC.

Third, please find a link to the executive summary (www.ccebos.org/PA_executive_summary_1.10.pdf and full report (www.ccebos.org/Performance_Assessment_Review_1.10.pdf) from the Center for Collaborative Education that we supported, titled *Including Performance Assessments in Accountability Systems: A Review of Scale-up Efforts*. This report addresses the questions on General and Technical Assessments. We understand the purpose of this competition is to develop high-quality assessments and accountability systems are not directly in this purview. However, one key purpose of assessments is for accountability and public reporting so we feel this report can inform the Department's thinking.

In closing, I share that the Nellie Mae Education Foundation is in "alliance" with the Ford Foundation, the Hewlett Foundation, and the Sandler Foundation in our thinking on these assessment issues. NMEF is supporting our New England states in a variety of assessment projects to move our region forward to a balanced assessment system that is aligned with the principles we submit to you as well as the suggestions for a federal RFP that our four foundations are submitting via the Hewlett Foundation.

Respectfully yours,

Nicholas C. Donohue
President & CEO

Nellie Mae Education Foundation
1250 Hancock Street, Suite 205N
Quincy, MA 02168
781-348-4201
ndonohue@nmefdn.org

Nellie Mae Education Foundation

Opening Doors to Tomorrow

General Assessment Input

Race to The Top Assessment Program

Submitted by Nicholas C. Donohue
President and CEO, Nellie Mae Education Foundation

Principles for Student Assessment

- Rooted in a clear set of attainable purposes.

  It is essential and possible for any system of assessment to reconcile and include the varied purposes of: 1) instructional improvement, 2) accountability, and 3) public transparency.

  The core purposes of a system of assessment should be to: 1) allow meaningful international comparisons, 2) equip students with the knowledge, habits of mind, dispositions and skills to succeed in 21st Century American work and life; 3) promote and improve educational opportunities that provide students with engaging educational opportunities that promote critical thinking, employ real world learning opportunities and are designed with attention to youth development principles; and 4) provide, parents, students, other stakeholders with information necessary to judge whether districts and schools are successful in ensuring these outcomes for all students, especially the underserved.

- Based on current research and knowledge.

  All assessments, whether for large-scale accountability or day-to-day instruction, should be based on current knowledge about the nature of learning and validated for their specific purposes. It is essential that assessments accurately measure what they purport to measure (validity) and that when used in various and multiple settings that similar responses will be judged similarly (reliability.) States should conduct frequent evaluations of assessments to ensure their quality as well as to determine how impact on teaching and student learning. The federal government should maintain rigorous quality control expectations regarding the development of these instruments.

- Includes a variety of approaches to assessment – on demand multiple choice, open-ended constructed response, evidence-based performance assessments.

  Various assessment types have different strengths and weaknesses. Multiple-choice assessments provide highly reliable measures of student performance. However the depth and validity of what can be measured through these kinds of items is limited. Open ended, constructed response items provide a balance of reliability and it is possible to get a deeper

view of complex reasoning skills. Evidence-based, performance assessments that include complex tasks demanding complex representations of performance through demonstrations and multi-modal projects give the strongest evidence of whether learners have achieved higher order learning outcomes. And while the issues of building inter-rater reliability of these kinds of assessment opportunities are significant, they are no longer insurmountable. In addition, the evolution of standards to include more complex cognitive and skills processes demands this kind of approach.

Developmental learning theory suggests that the use of different kinds of assessments at different stages of learning may be more appropriate than rigid and persistent application of one assessment approach regardless of age and developmental stage. It is crucial that well thought out, mix of assessment opportunities is provided to achieve each of the various purposes of a unified assessment system.

- Defined by a complement of national, state and local components each itself defined in part by its relationship to the others resulting in a unified, multi-level system of assessment.

  In order to achieve the principle of achieving a variety of purposes stated above, any system of assessment must include variety of ways to collect a variety of information.

  The national component of a unified system should collect information to provide information about 1) the standing of the United States compared to other countries and 2) to compare and contrast relative state performance.

  The state systems component should provide information 1) to contribute to international comparisons on a national level 2)) to make useful judgments and public comparisons about relative district performance in order to a) provide useful information to parents and other stakeholders and b) determine appropriate responses to local performance positive and negative.

  Local systems should be designed to 1) contribute to state processes to compare and contrast local performance and respond to those varying degrees of comparative and 2) to provide feedback to educators and learners about the quality and improvements necessary regarding educational engagements.

  These three levels of work should be designed: 1) consist with the overall purposes of the system, 2) for strong equating, 3) cost effective, feasible implementation as 4) intimately related components to a unified system versus independent components of loosely joined set of systems.

  Every level of the system should include a different variety and predominance of on-demand multiple choice, open-ended constructed response and evidence-based performance oriented assessments. The relative prominence of these various approaches to assessment should vary based on the level (national, state, local) in question. This relative balance would allow the three levels of a unified system to contribute to the variety of purposes towards which it is designed as a whole and to the specific purposes for which any one level is intended.

  Nationally this principle implies the evolution of NAEP to 1) maintain and strengthen its open-ended response features and 2) strengthen and enhance the evidence-based,

performance orientation to attend to the knowledge and skills that define successful preparation today are both crucial.

Consistent with the primary purposes of this state level assessment component its primary focus should be on state specific standards with enough flexibility to provide sufficient information to allow for national state by state comparisons and international comparisons. In order to do this, at the state level, it will be necessary to develop varied state assessment components that strike a relative balance between open-ended response items, multiple choice and common state level performance tasks in that order of emphasis.

At the local level, the focus must be a balance between 1) state level standards and 2) local graduation requirements with an emphasis on the latter in order to achieve the relative purposes of local assessment described above. This implies a significant predominance of locally developed, curriculum embedded performance tasks – a subset of which can be used at the state level. This last feature will allow the use of qualitative data to make higher stakes, comparative determinations about school performance.

- Aligned with the standards that define desired student outcomes and the educational opportunities provided to learners.

  It is important that here is a close connection between these three key facets of an education system. Assessments should measure desired outcomes. Assessments should also be designed in ways that consistent enough with instructional opportunities to ensure that what is tested has been taught. Assessments can measure other things, but they must measure what learners are expected to learn (for the purpose of feedback) and what education systems are meant to teach (for purposes of accountability and transparency.)

  The strength of alignment among these three dimensions is important at any level of implementation – national, state or local – and for type of assessment – multiple choice, constructed response or evidence-based performance task.

- Measures of absolute achievement against outcomes and relative progress.

  It is now possible to use assessment data to determine relative progress toward standards. This is a valuable development as long as significant progress is defined as sufficient. For reasons of equity and high expectations, the focus should remain on the absolute measures of achievement complemented by progress measures to provide a fully indexed view of school and student performance.

- Continuously improve teaching and learning.

  It is possible and important to frame assessment development and monitoring processes as part of a multi-faceted high quality professional learning system. The examination of student work for the purposes of cut score identification, future item development, determinations of proficiency and other related assessment purposes are essential opportunities for educators to understand what students really know, what needs to change in terms of their educational opportunities and what educators need to do to attend to these improvements in practice.

- Carefully consider the student populations involved, address any associated validity and fairness issues, and are equally accessible to underserved learners.

Issues of test bias must be addressed at all levels of any system of assessment. Lack of attention to the science regarding these issues risks integrity and legitimacy of low, mid and high stakes determinations connected to them.

- Cost effective, responsibly defined and are properly resourced at the local, state, and federal levels to support these principles in practice.

  Investments in assessment should be investments in accountability and in learning. Still, there are limits to any investment and the rationale for budgeting in any system should be based on a cost benefit analysis. However, discussions about costs for assessment are too often limited to easily determine dollar costs. A true cost benefit analysis would include estimations of 1) time for learning lost to test preparation that does not itself increase learning, 2) the related cost of this lost time in terms of remedial needs and needs related to non-promotion and dropping out, and 3) the in-kind contributions made in terms of professional learning experiences for educators involved.

  When considered in this way – combining dollar costs with collateral costs (positive and negative) into an appropriate analysis to guide expenditure – cost analysis have more integrity and are more accurate. In addition, this approach supports the case for the introduction of a wider variety of assessment approaches including evidence-based, performance oriented approaches. Test preparation that includes content-rich, complex tasks provide invaluable instructional opportunities as the assessments themselves are rooted in and driven by the same principles that define quality student engagement.

Input for USED on General and Technical Assessment
January 20, 2010

David Foster
Chief Scientist, Kryterion, Inc.

Here is a brief description of my experience that I use when submitting proposals for presentations and other purposes.

> Dr. David Foster currently serves as Chief Scientist of Kryterion, a unique Internet test administration company. He has started two other successful companies, Galton Technologies in 1997 and Caveon in 2003. Earlier, he directed the test development efforts at Novell from 1990 to 1997, introducing many new innovations, including adaptive testing, testing in multiple languages, and simulations-based testing. Dr. Foster serves on several industry boards and councils. He has been President and Chairperson of the Association of Test Publishers. He currently sits on the Council for the International Test Commission, recently served on the American National Standards Institute Board of Directors. He founded the Performance Testing Council to further the use of tests that directly measure important job or educational skills. He has authored numerous research and general articles for industry periodicals and reference texts, and has presented extensively at industry conferences. He has been creating and improving computerized testing systems since 1982 and has been influential over the past 28 years in introducing many important industry innovations and initiatives. Most recently at Kryterion he has introduced technologies that use the Internet to deliver secure tests wherever and whenever they are needed.

For this input to USED, I will be sticking to what I know most about and to what I felt has been under represented in the discussions/input, which is technology use in testing, performance testing, and test security.

General Comments/Recommendations

I support the comments and advice of the experts you have enlisted as well as those of many others who have provided input. Here are a few of the more important ones I noted while reading all the input documents and listening to the experts:

1. Don't require means, just ends. Encourage competition and multiple approaches to solutions.
2. Require the use of technology for testing; traditional systems, including many CBT models are already outdated. Online models that are browser/platform independent, can be updated at anytime, are becoming ubiquitous, and are less expensive, provide the most promise for current and future assessments.
3. Support innovations in testing. New technology inherently encourages innovations in testing. It is important to be able to introduce and use innovations without being restricted by older psychometrics and outdated testing conventions. Testing experts need to find a way to

encourage innovation that enhances the reliability, validity and fairness of tests. In my experience innovations have always improved testing for examinees and all other stakeholders.

Personal Observations and Comments

1. Stick to your guns about wanting innovation, broad implementation, and fast timelines. All these things are possible. Caution and careful progress is what we have had in the past.
2. We live in a world that is changing rapidly technologically. Logical assessment innovations with obvious benefits should be used immediately even if there isn't much research support. If those innovations are taken advantage of at the moment, they will naturally drop away and be replaced by something newer, leaving you will have the same problem.
3. Look toward what Distance Education programs are doing. Invite their input. They are more familiar and capable with technology use, and have many issues in common with you. They are beginning to solve performance testing and security issues.
4. Look for advice from other organizations, even for-profit organizations that have pioneered many if not most online innovations in test administration, security, performance testing, item banking, etc. A few include QuestionMark, Internet Testing Systems, and my own company, Kryterion. These companies will likely be contractors to the consortia, but they should be used at this stage to provide input or evaluate input you have already received.
5. Some current CBT models are already outdated. For example, many test development and administration systems are aging and proprietary, dependent on specific platforms, slow in transferring data and in providing reports. That technology should be leapfrogged.
6. Somehow we all need to separate this discussion from our collective historical experience with assessment in education. Much of the input received so far has been based on this historical experience, making it difficult to see new solutions and possibilities. New paradigms have not been introduced. CBT is not a new paradigm, but simply a way of doing paper-and-pencil tests using computers.

Performance Assessment

- There is absolutely no doubt that at the item level, technology can be used to measure most content standards better.
- We tend to look at this improperly as an measurement experts. We tend to think of performance assessment as a unique model different from multiple choice testing. All assessments should be performance assessment even if those assessments are comprised of multiple-choice questions. This has been stated more tactfully a few times in input documents for the Race to the Top Program. Simply, assessments should be built to specific content standards. Any assessment tool or item type should be used if it better measures those individual standards. Sometimes simple item formats will work best; for other standards, more complexity is required. If we focus on what the standard or objective requires instead of the format of the item or test we can build better assessments. I've had no or surprisingly few problems over 30 years of practice integrating several types of item formats in the same exam,

combining their scores for a total score. This is not new. I like the approach of the older intelligence tests: create items that measure the construct, integrate them all into a common exam, and administer them by hook or by crook (they usually administered them individually by trained examiners). We need that attitude and approach today; technology will be a great help.

- New online test development and test administration models are adding new test designs and new item formats all the time. It is not unusual for an item developer to choose between one of two dozen different formats, as well as multiple item scoring options (e.g., partial scoring). When finished the items can be combined in an assessment and published for immediate administration. It is relatively easy to customize these tests for individual students, different courses, parts of courses, different circumstances, etc.
- A note about item development. Most performance assessment projects run into trouble at the beginning, during item development, because it takes so much effort to create these supposedly complex items. For such an effort to be successful, test items, regardless of complexity, need to be created quickly and inexpensively by subject matter experts, not programmers or other specially trained individuals. There should be only minor differences in time and effort to create a text-only-based item and a simulation-based item. We did this successfully at Novell in the early 90's, so I know it can be done if the system is designed and built properly.
- Contrary to concerns of some input-providers, performance assessments hold the promise of better measurement, and therefore solve comparability issues, for students with disabilities and English Language learners.

Security

Summative assessment, and maybe even many formative assessments, will have important consequences for students and others. It is CRITICAL that test security be carefully considered as consortia respond to USED requirements. It will be a travesty to create wonderful and new assessment systems only to have them easily subverted by cheaters and test content pirates. It was good to hear Dr. Shepherd mention security today. Here are some specific points to consider:

1. Traditional models of handling security are under attack, successfully, by thieves and cheaters. Content is easily stolen and shared on the Internet. From what I read in the input documents and from USED, there isn't a lot of concern expressed about security. The assumption that traditional security models (teacher as proctor in a classroom testing circumstance) should and will be used is a bad assumption.
2. Online and on-demand (long or unlimited windows) assessment models increase the pressure on security systems by encouraging new security risks.
3. New and mostly available security technology will be needed to combat the current and new threats. These need to be considered as consortia respond to USED requirements. USED requirements should be included that cover how the consortia will prevent and detect and deal with security breaches.

4. The USED is requiring Distance Education programs to provide better security by authenticating the students who take courses, complete assignments and take tests. This is a good first step for distance education programs, but it must be required throughout the US educational system. And it should not just be the authentication of students, but include steps to prevent cheating and test theft.
5. Besides better systems, technologies and policies, teachers, administrators, parents and students will need to be trained in proper security of educational assessments.
6. There are new online-based item types and test designs that are more protective of test content and that discourage cheating. The use of these should be encouraged.

Again, it will do no good to create a great assessment system for education if it leaves it unprotected from cheaters and test thieves. Once compromised, the system will not be able to produce test scores that are reliable, valid and fair.

Specific Recommendations for by the 5 USED Questions Discussed in January 20, 2010, Meeting

<u>Q1: Requiring Through Course Summative Assessments</u>

1. Some online test development and test administration systems have much of this capability today. They are getting better as innovations are added each day through SaaS (Software as a Service) models. They allow rapid test development and deployment, a critical requirement for Through Course assessments. They can also provide immediate reports and data sets for the important decisions and for research needed. I would suggest that the USED look into current capability, not with the purpose of requiring specific products, but with the goal of helping to create technology-based requirements for the consortia.
2. Through Course assessments will need to be delivered on-demand through small or long testing windows. It is critical that security be enhanced for this testing. New and affordable online security models are available that make sure that:
    a. The proper student is taking the test
    b. The student is not cheating or getting help from other students or the teacher
    c. The student is not able to capture the test content and share it with others.
3. Performance assessments as defined as "using all means and methods to measure the desired construct as defined in the core standards," can be easily created and administered as evidenced today by such assessment occurring for non-educational testing purposes.

<u>Q2: Creating Common EOC Summative Exams for High School Subjects</u>

1. The issues of technology, performance assessment, and test security for EOC assessments are similar to those for Through Course Summative Assessments as listed above.

<u>Q3: Computerized Testing Challenges</u>

1. Innovative item types and test designs will actually improve the measurement of skills and knowledge of students with disabilities and English language learners. Audio, video, animations and simulations are occurring more commonly in non-educational exams (certification exams, for example). It is ONLY when such accommodations are applied that scores are meaningful and comparability across student populations can occur. Accommodations—and those can be greatly enhanced by technology—are the ONLY way to make assessment fair for these groups of students.
2. Comparability of paper-and-pencil with technology-based assessments should be a low priority. Making it a high priority will continue to "dumb-down" the assessment systems and force compliance with much older models. Randy Bennett's expert comments on this topic make a lot of sense.
3. Availability of technology, including Internet bandwidth. Technology resources are growing as we all know. Sources of funding to make such technology available for assessment purposes may need to be encouraged. Underused resources can be recruited to help out in the effort. I'm thinking of community computers and locations, community businesses can offer their facilities in the evenings, home computers and bandwidth can be used for high-stakes assessments as we are seeing with Distance Education end of course tests today. We should be thinking of these solutions instead of continuing to focus on what schools have or don't have. Ideally, all educational assessment can occur outside of the school environment and schedule, taking no time from teachers for instruction. While this is a radical thought and may go farther than desired, it is occurring today in distance education programs. A version of that can be applied to K-12 and typical higher education assessment situations.
4. Security issues. CBT and online testing have been criticized as creating more security problems than solving them. That is simply untrue. Some new security risks are introduced with technology-based tests; that is true, but others are resolved at the same time. But the largest benefit comes from the fact that technology-based tests allow many NEW security methods, such as online monitoring through webcams, real-time analysis of test performance looking for anomalies, un-intrusive authentication methods, and on and on.
5. A mild challenge is to incorporate principles of Universal Design. This is a good direction whether students are from special educational populations or not. Tests are more comparable if problems of items presentation and response input are eliminated. And they are eliminated by providing many ways for students to view or receive test item content and many ways for students to answer. It is only a mild challenge because many of these "innovations" are in place in existing systems, they just have not been made "universal", complete and available to all.

<u>Q4: Innovation and Improvement</u>

1. Using technology in testing will encourage and allow the development and implementation of hundreds of new innovations each year. This is a very good thing.
2. Research may be needed for many if not most of these innovations. Thorough, complete research isn't always required before implementing an obviously valuable and necessary

innovation. As mentioned above, even a short-term research program may take longer than the useful life of the innovation (as replacement innovations become available).

Q5: Two Specific Issues

1. I don't have any useful comments on the first topic of VAM. It's not an area where I have experience.
2. I want to repeat my comments on performance assessment. If we view assessment as either performance assessment (complex question types, essays, etc.) or something else (multiple choice assessment; traditional assessment, etc.) we will always struggle with generalizability and comparability. Instead, if we create integrated (i.e., using many different item types in the same assessment) assessments comprised of item types produced to measure specific standards, we should be able to measure any student, even those with disabilities or ELL students, equivalently. The focus has to be on the content standards, not the test format. The test and item formats will follow from the content standards.

Thank you for accepting my input.

Race to the Top Assessment Meeting: Technical Assessment

Dr. John C. Foster, President/CEO NOCTI (formerly known as the National Occupational Competency Testing Institute)

NOCTI is a non-profit assessment company serving the secondary, community college, and university levels of the Career-Technical Education (CTE) field. We very much appreciate the focus on technical assessment under RTTF, which recognizes that it is different from other types of assessment topics discussed at RTTF forums.

Technical assessment differs for three major reasons:

1) Technical assessments are of an immediate and critical nature due to their impact on assuring technical competence of the incumbent as well as the incoming workforce. Assured competence is desperately needed as employers scramble for talent coming out of the current recession and to maintain this country's global competitiveness.

2) Technical assessments/certifications are dynamic, needing to reflect continual upgrades in technical skills, processes, materials and equipment. Yet these assessments must meet the same psychometric standards as their more academic counterparts with static content.

3) Work with technical skill subject matter experts presents challenges. They can't commit days of time to design a test and they sometimes represent regional differences, which must be compensated for, if one is proposing a nationally standardized assessment.

All these issues add up to additional costs. The Department of Labor's funding initiative that embraces a competency modeling format (think performance testing) and the recently announced grant to establish a certifying exam for those entering new fields like medical information systems are exceptional steps; however, they are not enough, and what's more, these initiatives are not familiar enough to the education community.

NOCTI has seen many positive assessment decisions being made across the nation and has begun to see some cross-state collaboration. We have also seen some statewide technical articulation of secondary to post-secondary programs; so we believe that some states are picking up the Race to the Top challenge.

We have also, however, witnessed our share of concerns. Most of those concerns have to do with adequate and equitable resources. As a brief illustration, we were contracted by one state to develop dynamic technical skill attainment tests in 50 occupational areas. Our contract was approximately 1.2 million dollars over a three-year period. During that same period, another third party developer received 23 million dollars to develop 3 static academic tests to be given at 3 different levels. Quick division shows resources roughly 108 times higher on development of each academic test when compared with each technical test. When policy makers indicate that assessment of technical competence is supposed to assure the maintenance of this

country's infrastructure and the sustainability of our economic prosperity, it is difficult to reconcile inequities of this magnitude!

NOCTI has demonstrated a willingness to help with the ongoing assessment efforts. We have demonstrated that willingness by participating as a partner in the National Research Center for Career and Technical Education and by working with OVAE and the Data Quality Institutes. We have also sent a letter to Secretary Arne Duncan, basically underscoring that we would be happy to contribute our national perspective and expertise. Our business requires in-depth knowledge of the process of building a test. We understand the need for item analysis, documented validity and reliability, and we know what it takes to develop a nationwide test with the proper weightings and ratings. We hope to be of service to the education community, particularly the CTE community. We would strongly encourage equitable funding through Race to the Top or any other funding streams available to advance this critical work.

Thank You!

STATE OF MAINE
DEPARTMENT OF EDUCATION
23 STATE HOUSE STATION
AUGUSTA, MAINE
04333-0023

JOHN ELIAS BALDACCI
GOVERNOR

SUSAN A. GENDRON
COMMISSIONER

January 20, 2010

TO: Office of Elementary and Secondary Education
Race to the Top Assessment Program
U.S. Department of Education
400 Maryland Avenue, SW. room 3W339
Washington, D.C. 20202

FROM: Susan A. Gendron
Commissioner of Education
Maine Department of Education
33 State House Station
Augusta, Maine 04333-0333
207 624-6620

RE: Project and Consortium Management

On behalf of the state of Maine and 36 other states as listed in the attachment I am submitting for your consideration a Memorandum of Understanding signed by 37 states that are committed to establishing a framework of collaboration for states in supporting assessment of the common core standards. The agreement articulates tasks in support of a Multi-State Consortium in its implementation of an approved Standards and Assessment Section of a Race to the Top grant. The MOU outlines a set of working principles, the roles of states and local districts within the consortium and a set of tasks that the Consortium would undertake.

Invitations have been extended to leading assessment experts across the country to assist the State Consortium Developing Balanced Assessments of the Common Core in creating a design team to assist the states in the development of this comprehensive integrated summative and formative assessment system.

It is hoped that this consortium can become the umbrella organization for states to achieve a rigorous internationally benchmarked common reference assessment which would include selected response, constructed response and performance components; utilize new technologies to provide adaptive tools and access to information, computer based scoring and reporting; develop rubrics that embody the standards, create examples of good work; implement high quality professional learning focused on examination of student work, curriculum, and assessment development and moderated scoring; and above all utilize all of this information to improve teaching and learning.

Attached is a copy of the Memorandum of Understanding for the State Consortium Developing Balanced Assessments of the Common Core Standards and the list of partner states as of this date.

MOU for a State Consortium Developing Balanced Assessments of the Common Core Standards

This Non-Binding Memorandum of Understanding ("MOU") is entered into by and between the Balanced Assessment Consortium and the State of Maine. The purpose of this agreement is to establish a framework of collaboration for states in supporting assessment of the common core standards. The agreement also articulates tasks in support of a Multi-State Consortium in its implementation of an approved Standards and Assessment Section of a Race to the Top grant. The MOU outlines a set of working principles, the roles of states and local districts within the consortium, and a set of tasks that the Consortium would undertake.

Working Principles

A consortium of states developing a balanced assessment system for evaluating the common core standards would start with working principles derived from an examination of successful state systems in the U.S. and high-achieving systems internationally. For example:

1) **Assessments are grounded in a thoughtful, standards-based curriculum and are managed as part of a tightly integrated system of standards, curriculum, assessment, instruction, and teacher development**.

- Curriculum guidance is lean, clear, and focused on what students should know and be able to *do* as a result of their learning experiences. Assessment expectations are described in the curriculum frameworks or course syllabi and are exemplified by samples of student work.
- Curriculum and assessments are organized around a well-defined set of learning progressions within subject areas. These guide teaching decisions, classroom-based assessment, and external assessment.
- Teachers and other curriculum experts are involved in developing curriculum and assessments which guide professional learning and teaching. Thus, everything that comes to schools is well-aligned and pulling in the same direction.

2) **Assessments elicit evidence of actual student performance on challenging tasks that prepare students for the demands of college and career in the 21st century.** Curriculum and assessments seek to teach and evaluate a broad array of skills and competencies that generalize to higher education and work settings. They emphasize deep knowledge of core concepts within and across the disciplines, including problem solving, analysis, synthesis, and critical thinking, and include essays and open-ended tasks and problems, as well as selected response items.

3) **Teachers are involved in the development of curriculum and the development and scoring of assessments**. Scoring processes are moderated to ensure consistency and to enable teachers to deeply understand the standards and to develop stronger curriculum and instruction leading to greater student proficiency. The moderated scoring process is a strong professional learning experience that helps drive the instructional improvements that enable student learning, as teachers become more skilled at their own assessment practices and their development of curriculum to teach the standards. The assessment systems are designed to increase the capacity of teachers to prepare students for the contemporary demands of college and career.

4) **Assessments are structured to <u>continuously improve teaching and learning</u>**. Assessment *as*, *of*, and *for* learning is enabled by several features of assessment systems:

- The use of school-based, curriculum-embedded assessments provides teachers with models of good curriculum and assessment practice, enhances curriculum equity within and across schools, and allows teachers to see and evaluate student learning in ways that can feed back into instructional and curriculum decisions.

- Close examination of student work and moderated teacher scoring of both school-based components and externally developed open-ended examinations are sources of ongoing professional development that improve teaching.

- Developing both school-based and external assessments around learning progressions allows teachers to see where students are on multiple dimensions of learning and to strategically support their progress.

5) Assessment and accountability systems are designed to <u>improve the quality of learning and schooling</u>. Assessments aim to encourage and support the learning of ambitious intellectual skills in the way they are designed and used for informing teaching, learning, and schooling. Accountability systems publicly report outcomes and take these into account, along with other indicators of school performance, in a well-designed system focused on continual improvement for schools.

6) Assessment and accountability systems <u>use multiple measures</u> to evaluate students and schools.

Multiple measures of learning and performance are used to evaluate skills and knowledge. Students engage in a variety of tasks and tests that are both curriculum-embedded and on-demand, providing many ways to demonstrate and evaluate their learning. These are combined in reporting systems at the school and beyond the school level. School reporting and accountability are also based on multiple measures. Assessment data are combined with other information about schools' resources, capacities, practices, and outcomes to design intensive professional development supports and interventions that improve school performance.

7) <u>New technologies enable greater assessment quality</u> and information systems that support accountability.

New technologies enhance and transform the way the assessment process is developed, delivered, and used, providing adaptive tools and access to information resources for students to demonstrate their learning, and providing appropriate feedback by supporting both teacher scoring and computer-based scoring (now possible for both selected response and some forms of constructed-response items). By using technology to reduce costs for delivery of more open-ended assessment formats, scoring, and reporting, resources can be redirected to improvements in assessment quality.

Technology also organizes data about student learning, enhancing system accountability for instruction and reporting by providing more efficient, accurate, and timely information to teachers, parents, administrators, and policymakers. Technology helps to integrate information at as part of longitudinal data systems, contributing to a rich profile of accomplishment for every student.

<u>State and Local Roles within a Consortium</u>

States working within the Consortium would:

- Adopt and augment the Common Core standards as appropriate to their context.

- Create and deploy curriculum frameworks that address the standards—drawing on exemplars and tested curriculum models.
- Build and manage an assessment system that includes both on-demand and curriculum-embedded assessments that evaluate the full range of standards and allow evaluation of student progress. The Consortium may develop both joint assessments (commonly implemented by states) as well as other assessment tasks and items linked to the standards (and grounded in curriculum units) that can be incorporated into states' individual assessment plans for formative or summative purposes.
- Develop rubrics that embody the standards, and clear examples of good work, benchmarked to performance standards.
- Create oversight / moderation / audit systems for ensuring the comparability of locally managed and scored assessment components.
- Ensure that teacher and leader education and development infuse knowledge of learning, curriculum, and assessment.
- Implement high-quality professional learning focused on examination of student work, curriculum and assessment development, and moderated scoring.

Districts and schools would:

- Examine the standards and evaluate current curriculum, assessment, and instructional practice in light of the standards.
- Evaluate state curriculum guidance, and further develop and adapt curriculum to support local student learning, select and augment curriculum materials, and continually evaluate and revise curriculum in light of student learning outcomes.
- Incorporate formative assessments into the curriculum, organized around the standards, curriculum, and learning sequences to inform teaching and student learning.
- Participate in administering and scoring relevant portions of the on-demand and curriculum-embedded components of the assessment system, and examining student work and outcomes.
- Help design and engage in professional development around learning, teaching, curriculum, & assessment.
- Engage in review and moderation processes to examine assessments and student work, within and beyond the school.

Tasks the Consortium Would Undertake

The consortium of states would build on successful efforts already launched in a number of states, seeking to integrate the best knowledge and exemplars from existing efforts, so as to use resources efficiently, take advantage of well-tested approaches, and avoid reinventing the wheel. It would bring together leading curriculum and assessment experts to advise and support efforts to create a system for evaluating the Common Core, building on the most credible and well-vetted knowledge available in the field. With these supports, the Consortium could:

1. Support the Development of Curriculum Frameworks: When the Common Core standards have been released, vetted, and adopted, consortia of states would work with curriculum and assessment experts to develop (or adapt from previously successful work) curriculum frameworks, syllabi, and other materials mapped to the standards. There has been enormous investment in the United States in high-quality curriculum, for example through NSF and other organizations at the national level, and in many states and districts. Other English-speaking nations have also developed high quality curriculum materials linked to standards and learning

progressions that could be evaluated in this process. This effort would inventory and cull from efforts with a strong evidence base of success to support states in building out curriculum frameworks around which they can organize deeper curriculum development at the local level, state and local assessment development, instructional supports, and professional development.

2. Create a Digital Curriculum and Assessment Library: The results of this effort should ultimately be made available on-line in a digital platform that offers materials for curriculum building and, eventually, model syllabi for specific courses linked to the standards, formative and summative assessment tasks and instruments linked to the curriculum materials, and materials for training teachers and school leaders in both strategies for teaching specific curriculum concepts / units and assessment development and scoring. In addition, as described below, an electronic scoring platform supporting training, calibrating, benchmarking, and reporting would be developed and made available across the states.

3. Develop State and Local Assessments: The state consortium would work to create a **common reference examination, which includes selected-response, constructed response and performance components** aimed at higher-order skills, linked to the Common Core standards for grades 3-8, like the NECAP assessment recently developed by a set of New England states. This assessment would be designed to incorporate more rigorous and analytic multiple-choice and open-ended items than many tests currently include and would include strategically selected curriculum-embedded performance assessments at the classroom level that can be part of the summative evaluation, while also providing formative information.

These curriculum-embedded components would be developed around core concepts or major skills that are particularly salient in evaluating students' progress in English language arts and mathematics. (Eventually, work on science could be included.) Exemplars to evaluate and build upon are already available in many states and in nations like England that have developed a set of "tests and tasks" for use in classrooms that help teachers evaluate students' learning in relation to well-described learning progressions in reading, writing, mathematics, and other subjects.

Curriculum-embedded components would link to the skills evaluated in the "on-demand" test, allowing for more ambitious tasks that take more time and require more student effort than can be allocated in a 2 or 3-hour test on a single day; these components would evaluate skills in ways that expect more student-initiated planning, management of information and ideas, interaction with other materials and people, and production of more extended responses that reveal additional abilities of students (oral presentations, exhibitions, and product development, as well as written responses) that are associated with college and career success.

In the context of summative assessments, curriculum-embedded tasks would be standardized, scored in moderated fashion, and scores would be aggregated up to count as part of the external assessment. Curriculum-embedded assessments would also include marker tasks that are designed to be used formatively to check for essential understandings and to give teachers useful information and feedback as part of ongoing instruction. Thoughtful curriculum guidance would outline the scaffolding and formative assessment needed to prepare students to succeed on the summative assessments.

All components of the system would incorporate **principles of universal design** that seek to remove construct-irrelevant aspects of tasks that could increase barriers for non-native English speakers and students with other specific learning needs. In addition, designers who are skilled

at developing linguistically supportive assessments and tests for students with learning disabilities would be engaged from the beginning in considering how to develop the assessments for maximum access, as well as how to design appropriate accommodations and modifications to enable as many students as possible to be validly assessed within the system.

The emphasis on evaluating **student growth over time** and on tying standards to a conception of learning progressions should encourage a growth oriented frame for both the "on-demand" examination and the more extended classroom assessments. The Consortium may consider the viability of incorporating computer-based adaptive testing that creates vertically scaled assessments based on the full range of learning progressions in ELA and math. This would allow students to be evaluated in ways that give greater information about their abilities and their growth over time. This approach would not preclude the evaluation of grade-level standards, which could be part of any students' assessment, nor would it preclude a significant number of constructed response, open-ended items, as the technology for machine-scoring structured open-ended items is now fairly well-developed. Strategic use of partial teacher scoring for these items would also be a desirable element of the system to support teachers' understanding of the standards and assessments, and their planning for instruction.

The emphasis on evaluating student growth should also inform the development of the curriculum-embedded elements of the system, which should be selected or developed to strategically evaluate students' progress along the learning continuum. Centrally developed tasks administered and scored by teachers with moderation (see below), using common rubrics, would be part of the set of reported scores. In states with experience and capacity, it may be possible to begin to incorporate information about student learning that teachers develop from their own classroom evidence, linked to the standards and learning progressions and guided by the curriculum frameworks. This could be an optional aspect of the Consortium's work for states and communities with interest and capacity.

At the **high school level**, the Consortium might explore one or both of two options for assessment:

- **Course- or syllabus-based systems** like those in England, Australia, Singapore, Hong Kong, Alberta (Canada), as well as the International Baccalaureate. Generally conceptualized as end-of-course-exams in this country, this approach should become a more comprehensive course assessment approach like that pursued in these other countries. Such an approach would include within-course performance assessments that count toward the examination score, as well as high-quality assessment end-of-course components that feature constructed response as well as selected response items. Within-course performance assessments would tap central modes of inquiry in the disciplines, ensuring that students have the opportunity to engage in scientific investigations, literary analyses and other genres of writing, speaking and listening; mathematical modeling and applications; social scientific research. Such an approach might require an ELA and math assessment at a key juncture that evaluates an appropriate benchmark level for high school standards, and then, as in high-achieving nations, allow for pursuit of other courses/ assessments that are selected by students according to their interests and expertise. These could serve as additional information on the diploma for colleges and employers.

- **Standards-driven systems** that might include a more comprehensive benchmark assessment in ELA and mathematics complemented by collections of evidence that demonstrate

students' abilities to meet certain standards within and across the disciplines. This set of assessments would allow more curriculum flexibility in how to meet the standards. Systems like these are used in some provinces in Canada and Australia, in states like Rhode Island, Wyoming, Nebraska, and New Hampshire, and in systems of schools like the New York Performance Standards Consortium, the Asia Society, and Envision Schools. Sometimes these sets of evidence are organized into structured portfolios, such as the Technology portfolio in New Hampshire and the broader Graduation portfolios in these sets of schools that require specific tasks in each content area, scored with common rubrics and moderation.

- **A mixed model** could combine elements of both course- and standards-driven models, allowing some demonstrations of proficiency to occur in any one of a range of courses (rather than a single, predetermined course) or even outside the bounds of a course, like the efforts by some states to allow students to pass courses via demonstrations of competence rather than seat time (e.g. NH, OH). Such a system could also include specific components intended to develop and display research and inquiry skills that might also be interdisciplinary, such as the Project Work requirements in England, Singapore, and the International Baccalaureate, and the Senior Project requirements in Pennsylvania and Ohio.

4. Develop Moderation and Auditing Systems for Teacher-Scored Work: The consortium would develop protocols for managing moderation and auditing systems and training scorers so as to enable comparable, consistent scoring of performance assessments. In other nations' and states' systems that include these features routinely, procedures have been developed to ensure both widespread teacher involvement – often as part of professional development time – and to create common standards and high levels of reliability in evaluating student work. A range of models are possible, and the consortium would serve as a resource to individual states in developing and implementing strong, efficient approaches.

5. Develop Technology to Support the Assessment System: Technology should be used to enhance these assessments in a number of ways: by delivering the assessments; in on-line tasks of higher-order abilities, allowing students to search for information or manipulate variables and tracking information about the students' problem-solving processes; in some cases, scoring the results or delivering the responses to trained scorers / teachers to assess from an electronic platform. Such a platform may also support training and calibration of scorers and moderation of scores, as well as efficient aggregation of results in ways that support reporting and research about the responses. This use of technology is already being used in the International Baccalaureate assessment system, which includes both on-demand and classroom-based components.

In order to gain the efficiency and cost benefits of machine scoring and the teaching and learning benefits of teachers' moderated scoring, a mixed system could be developed where computer-based scoring is incorporated on constructed response tasks where useful – though teachers would score some of these tasks for anchoring and learning purposes – while other tasks that require human scoring engage most teachers in scoring to support improvements in instruction.

RESPONSIBILITIES OF ALL SEAs PARTICIPATING IN THE CONSORTIUM

1) Each participating SEA in the Consortium will appoint a key contact person.

2) These key contacts from each State will maintain frequent communication with the parties administering the Balanced Assessment Consortium to facilitate cooperation under this MOU.

3) Participating SEA grant personnel will work together to determine appropriate timelines for project updates and status reports throughout the whole grant period.

This Non-binding Memorandum of Understanding shall be effective beginning with the date of the last signature hereon:

SEA Superintendent/- Participating State Chief/Commissioner (or equivalent authorized signatory)

Susan A. Gendron

	1-7-10
Signature	Date
Susan A. Gendron	Commissioner of Education Maine Department of Education
Print Name	Title

Signatory States

Consortium Developing Balanced Assessments of the Common Core Standards

Alabama
Arizona
Arkansas
California
Connecticut
Delaware
Illinois
Indiana
Georgia
Iowa
Kansas
Kentucky
Maine
Maryland
Massachusetts
Michigan
Mississippi
Missouri
Montana
Nebraska
New Hampshire
New Jersey
New York
North Carolina
North Dakota
Ohio
Oklahoma
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Utah
Washington DC
West Virginia
Wisconsin
Wyoming

3

National Association for Gifted Children
1707 L Street, NW, Suite 550
Washington, DC 20036
(202) 785-4268
www.nagc.org

STATEMENT OF NAGC BEFORE THE U.S. DEPARTMENT OF EDUCATION ASSESSMENT MEETING

JANUARY 20, 2010
WASHINGTON, DC

Hello, my name is Nancy Green; I am the Executive Director of the National Association for Gifted Children, the nation's largest organization dedicated to supporting gifted learners.

Our nation as a whole has long neglected the needs of our highest-potential students. A number of issues are to blame including the lack of a comprehensive national gifted education policy, a piecemeal system of identification and services that varies sharply between states and within states, and a focus too heavily weighted toward achieving proficiency alone. For many reasons, the status quo is not serving our nation and our more than 3 million academically gifted children.

On behalf of these students, I offer the following recommendations regarding the Race to the Top Assessment Fund.

NAGC vigorously supports the intent of the Race to the Top Assessment program. To help further improve the program, we recommend a special focus on

- the precision of assessments for students at the upper end of the achievement spectrum, and
- the use of computer adaptive assessments for this special population of students

Empirical and anecdotal evidence demonstrate that the assessment and accountability provisions in the *No Child Left Behind Act* have had negative consequences for our nation's advanced learners. The law's assessment and accountability system created a much-needed focus for educators to address the needs of students who do not read or perform mathematics at grade level. Unfortunately, this incentive structure has driven educational systems to neglect the unique learning needs of students performing above the proficient level in reading and mathematics. As a result, the academic progress of higher-performing students has languished during this same time period.

The current system has also helped widen dramatically the racial achievement gap between students on the higher end of the achievement spectrum. As districts focus on their struggling learners, new research shows that the achievement gaps in reading and math skills between Kindergarten and fifth grade grow roughly twice as fast for students who begin school with scores one standard deviation above the mean as for those who begin one standard deviation below the mean.[1] This trend is especially stark between high-achieving African-American and White students. Other recent research has found similar results.[2] Half of lower-income students who were classified as high-achievers in math early in their schooling lost this status later in elementary school.[3] Clearly, we need to do better.

---

[1] Reardon, F. (2008, March). *Differential Growth in the Black-White Achievement Gap During Elementary School among Initially High- and Low-Scoring Students (Working Paper # 2008-07).* Palo Alto, CA. Stanford University.

[2] Hanushek, E.A. & Rivkin, S.G. (2006, October). *School Quality and the Black-White Achievement Gap*. Cambridge, MA: National Bureau of Economic Research; Gribben, M. A., Campbell, H. L., & Mathew, J. (2008). *Are Advanced Students Advancing? Examining Achievement Trends Beyond Proficiency*. Alexandria, VA: Human Resources Research Organization.

[3] Wyner, J., Bridgeland, J.M., & Diulio, J.J. (2008). *The Achievement Trap: How America is Failing Millions of High-Achieving Students from Lower Income Families.* Lansdowne, VA: Jack Kent Cooke Foundation; Hanushek, E. A. & Rivkin, S. G. (2006). *School Quality and the Black-White Achievement Gap*. Cambridge, MA: National Bureau of Economic Research

NAGC understands that high-ability learners present a special challenge. By definition, gifted and talented students are capable of performing at a level beyond their age peers and require more academic rigor than is typically offered in the regular classroom. For the fortunate few, their schools adjust the curriculum and instruction they receive and provide appropriate services. Unfortunately this is not the reality for most. We believe that access to accurate student performance measures would be the first step to recognizing the unique learning needs of gifted and talented learners and in adapting instruction to student ability.

Focusing grants on the design and quality of assessment systems to support improved teaching and learning, rather than just on accountability policies, is a critical improvement. However, to achieve the Race to the Top Assessment program goals, funded projects must encourage the development of assessments that accurately measure learning gains for students across the entire achievement spectrum.

The issue of the precision of assessments across the full performance continuum is a critical one for students with unique learning needs. Most of the discussion among policy makers focuses on the reliability of assessments for students with disabilities, which has sparked innovation in the development of new assessments and accommodations for these students.

By contrast, this same innovation has not developed for gifted students. This special population also needs assessments that are able to accurately measure their achievement, which frequently exceeds grade-level standards. NAGC strongly urges the Department to include a priority for valid and reliable measures of student achievement and growth at the upper end of the performance spectrum. We believe that this stated priority will bring much needed attention to the unique learning needs of gifted and talented students.

Computer adaptive assessments, which are growing in popularity, are another issue of special interest. The promise of these assessments is to be able to pinpoint students' ability levels with test questions based on previous test performance. With this data, educators will have the important information they need to differentiate and adapt instruction to the meet student needs. Many local school districts currently use computer adaptive assessments to determine gifted student mastery of a subject. This allows teachers to implement pedagogical strategies, such as academic acceleration and flexible ability grouping, to ensure that gifted and talented students meet not only a grade-level "standard," but also their individual learning potential.

We understand that the new common core standards are organized in "grade bands." In other words, the standards cover multiple grades. We encourage the corresponding grade-level assessments to also cover multiple grades. This strategy may be able to mitigate the "ceiling effect," which so many gifted and talented students experience when taking standardized tests.

NAGC is concerned that an emphasis on measuring student performance against only grade-level standards will continue to harm some gifted and talented students by preventing them from reaching their full academic potential. Educators will continue to be driven to teach to the middle rather than to maximize student talents. Together, we must champion a new movement that strives for excellence in education rather than the mediocrity that results from merely asking students to reach proficiency. We must challenge educators and the public to abolish preconceived notions of what a child at a certain age ought to know and, instead, provide students with the academic support to exceed their own, and our, expectations.

NAGC applauds Secretary Duncan's efforts to improve the quality of assessments in our public schools. We urge the Department to remain strong in its efforts to ensure that assessment information is used to inform and tailor instruction for high-ability students. Thank you for your attention.

Attached: December 2, 2009, letter to Secretary Duncan on RTT-Assessments

NATIONAL ASSOCIATION FOR
Gifted Children

National Association for Gifted Children
1707 L Street, NW, Suite 550
Washington, DC 20036
(202) 785-4268
www.nagc.org

December 2, 2009

Honorable Arne Duncan
Secretary
U.S. Department of Education
400 Maryland Avenue, SW
Washington, DC 20202

RE: Race to the Top Assessment Program

Dear Secretary Duncan:

In recent months, the U.S. Department of Education's Race to the Top program and proposed reforms have generated overwhelming interest and action in determining the best path forward for improving student achievement in this country. The National Association for Gifted Children (NAGC) applauds the administration for sparking this dialogue, and we vigorously support the intent of the Race to the Top Assessment program. Focusing the grants on the design and quality of assessment systems to support improved teaching and learning rather than on accountability policies is a critical development. However, to achieve the goals of the Race to the Top Assessment program, projects funded must encourage the development of assessments that accurately measure learning for *all* students, not just those around the mean. To best reflect this aim, we have several recommendations for the proposed framework outlined in the October 20, 2009, Race to the Top Assessment Program Executive Summary.

(1) In the *Design of Assessment Systems – General Requirements* section

Recommendation 1: Rather than assessments focusing solely on the extent to which each individual student is on track, at each grade level tested, toward college or career readiness, we recommend that the framework acknowledge and incorporate the reality that **some students' achievement surpasses grade-level proficiency. Enhanced assessments must be able to accurately measure these students' knowledge, skills, and abilities in order to be able to show student learning gains.**

Currently, summative tests only track and assess student proficiency and mastery of an age- and grade-specific standard. Improved assessments will capture the extent of a student's ability so that educators can accelerate his/her instruction appropriately.

Current assessments are also limited in their ability to validly and reliably measure student achievement the further away a student moves from the mean. In other words, the assessment becomes less reliable at the upper and lower spectrums of achievement. For our most highly able students, access to tests or test items that are above grade level would provide educators with more precise understandings of how to teach and challenge these students.

Recommendation 2: NAGC recommends that the expectations for the information gathered from the assessments should be expanded to include "**Determinations of a student's current academic performance and future academic potential.**"

We applaud the efforts to boost college and career readiness, but we believe that the next generation of assessments should provide more precise understandings of each student's absolute achievement rather than achievement confined to his or her age and grade-level placement. Assessments that are able to accurately measure a broad spectrum of achievement, especially above the mean performance, will allow for more appropriate instruction and planning for students who may be on a different learning and graduation trajectory than their age peers.

(2) In the *Design of Assessment Systems – Required Characteristics* section

Recommendation 1: NAGC suggests that the framework make clear that the new generation of assessments must be valid, reliable, and fair for all students, including those already at or above grade level.

Because the current federal focus in NCLB, as well as recent federal assessment pilot projects, is and has been on students performing below or near proficiency levels, it may not be clear to grant applicants that the next generation of assessments must be able to measure the performance of all students, including advanced students who typically score at the highest percentiles of their state tests. Clearly, the goal of using assessment data to plan instruction for students cannot be met for top students if the assessments cannot accurately measure their knowledge and skill levels.

(3) In the *Design of Assessment Systems – Desired Characteristics* section

Recommendation 1: NAGC recommends including **teachers who teach gifted and talented students into item # 1**. NAGC strongly supports the inclusion of teachers in scoring of constructed responses and performance tests in order to effectively measure student mastery of higher-order content and skills and to better understand performance expectations. We believe that it is critical that teachers with expertise in instructing students who are gifted and talented would ensure that assessments and instruction improve for these students.

Recommendation 2: NAGC supports item #4 on building the technology infrastructure, but recommends that the item notes that **the assessments should be calibrated to the performance levels of the individual students, including those performing at the top end of the achievement spectrum**.

(4) In the *Design of Assessment Systems – LEA-Level Activities* section

Recommendation 1: NAGC recommends **adding "gifted and talented students" to the list** of student populations that should be included in the pilot test of the new assessments to ensure that new assessments are able to measure learning growth for those students at the top end of the performance scale.

Recommendation 2: NAGC recommends adding the following to the list of activities for participating LEAs: "**Designing systems that support students who surpass proficiency levels with flexible modifications, informed by the new assessment data, to course and grade level placements and other strategies to meet their learning needs.**"

(5) After the *Question on the Assessment of Students with Disabilities*

Recommendation 1: NAGC recommends adding a new section entitled "**Questions on the Assessment of Gifted and Talented Students.**" NAGC strongly believes that gifted and talented students have unique learning needs that are not being met in many schools under current conditions. Directly incorporating gifted and talented students into this assessment reform package would be an important step toward meeting the needs of these students. This view of gifted and talented students is consistent with the newly adopted definition of "teaching skills" in HEOA, which recognizes that this population has unique learning needs.

Thank you for the opportunity to comment on these important reforms. We appreciate your efforts to improve this critical area for our country's students.

Sincerely,

Nancy Green

Nancy Green
Executive Director

January 20, 2010

A Submission to the Assessment Session

From: Myrtice J. Irish, Ed.D.

Representative, Community Advisory Commity

Los Angeles Unified School District

Assistant Professor, Special Education

National University

Contact Information: e mail: mirish@nu.edu

Telephone: (909) 919 – 7634, (909) 806 – 3346

Cell Phone: (909) 240 – 4474

Question 2: The need for a system of end of course or grade summative exams that accurately monitor student progress in the curriculum and encourage better teaching

Introduction:

I started teaching in 1963. In the intervening years, I have been a general and special education classroom teacher, a department chair, a teacher union vice president and economic negotiator, a title 1 coordinator, an assistant principal, and finally a central office administrator. I am currently retired from public education but remain involved as a university professor involved in the preparation of students to be special education teachers and as a community representative on the Community Advisory Council of the Los Angeles Unified School District. I offer this abbreviated history of my varied educational career because I want it understand that I have a very long view of public education trends and their results. This history also allows me to tell you categorically that without exception learning is the result of the relationship between a teacher and his or her students as they together explore the concepts and develop the skills required by the state mandated curriculum.

Bean Counting your Way to Improvement – a description of current practice!

The assessment system adopted to meet the requirements of NCLB is a norm referenced exam administered in mid year that in the perception of the students counts for nothing. But, in fact, it counts for a great deal as the results of this assessment are being misused in a very major way. The assessment theoretically assesses the student's knowledge of curriculum content and it's results are also used to develop the API and AYP scores that determine whether schools are regarded as successful. Individual and group scores are communicated to schools and parents along with a verbal descriptor which describes which category of the normal curve the student's score places him. The terms in current use are Advanced, Proficient, Basic, Below Basic and Far Below Basic. Schools who do not meet the state target on the mandated test must find a way to meet the target and this is where the misuse begins. As I retired there were actually workshops being held to train administrators and teachers to identify which students could be most easily moved to the proficient level, thus increasing the school's likelihood of meeting the state mandated performance target. The practice was referred to as looking for the students who could deliver the most AYP bang for the buck. The student's who achieved in the Basic Category appear to meet the definition of what is being looked for. The focus of remedial services has shifted to this group to the detriment of other groups of students particularly the gifted and the students with disabilities who fall in the mild to moderate category. Funds to lower achieving categories are regarded as wasteful in these circumstances as the movement required is much larger and therefore there is much less likelihood of success. This sharpening of focus on one group produces a drain of resources which once applied to other students because they are in the category judged most easily improved.

Funds diverted are largely being used for the purchase of researched based programs which promise quick improvements and are delivered through scripted lessons which standardize instructional approach and delivery in a demand for conformity to the script. Behind the overt research based results guaranteed sales pitch is the belief that these programs will succeed because they are not subject to or dependent on the idiosyncratic organization and planning skills of individual teachers. Training for working with this group thus becomes learning how to follow the script and track results. This ultimately leads to a decline of teacher abilities in curriculum knowledge, planning for instruction and practice of a variety of instructional strategies.

A Modest Proposal: (to be expanded later)

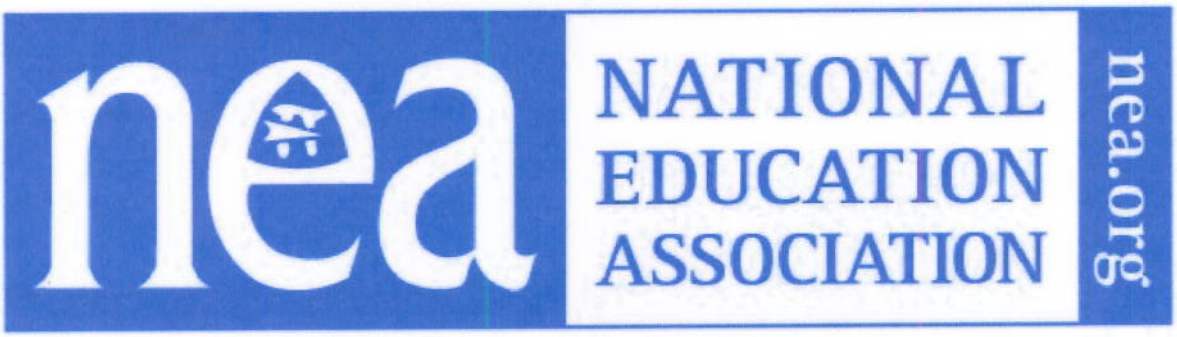
What is required is a different form of assessment designed to assess student progress in the concepts and skills specified in the curriculum mandated for each grade level. The assessment would come in the form of subject and or grade end tests that would count toward the student's final grade and promotion.

I propose a representative group of school districts working in partnership with a university faculty of education in order to development a curricular assessment program. The first task for such a partnership would be building a test bank of curricular test items which could be validated and have reliability established by norming items in the school of the district partners. A sufficient bank of items could be developed and ready for implementation state wide within 3-5 years.

Conclusion:

Aside from ensuring that students are actually learning the concepts and skills mandated by the curriculum, this type of system automatically reinforces the importance of curriculum and instructional strategies in the minds and practice of teachers. Administrators could once again functional as educational leaders rather than being participants in a bean counting enterprise focused on how to beat the system in the most expeditious way. The original title of principal was **principal teacher** and it is time we got back to that and refocused our attention and resources on the advancement of all students not just the those that count the most toward an increased API score.

9



Great Public Schools for Every Student

Testimony of

Barbara Kapinus

Senior Policy Analyst

Education Policy and Practice Department

National Education Association

January 20, 2010

For Public Hearing on Race to the Top Funding for Assessment Systems

Because I am speaking on behalf of the National Education Association, I will be emphasizing the importance of considering the role of teachers in the development, implementation and improvement of assessment systems. Because formative assessment is a teacher tool for classroom assessment and planning, I will be emphasizing its importance as part of a complete assessment system that leverages education improvement.

NEA wants every child who attends a public school to receive a complete, worthwhile education. The key to creating and sustaining such schools is the development of accomplished teachers who continually strive to improve their professional practice. Simply administering assessment after assessment by itself will not will not accomplish the goal of schools providing complete educations. This has been stated by researchers, educators, and testing experts (e.g., Ed Haertel, Linda Darling Hammond, Scott Marion, and Lorri Shepard) as testimony on the Race to the Top assessment funding. However, effective assessment systems, when accompanied by other resources for the well-being of students can be a tremendous factor in the improvement of the quality of education. In order to construct such assessment systems, it is necessary to reexamine the goals of educational assessment and put aside much of what we have been doing in education assessment.

Current assessment practices are linked to the goal of attaining specific scores levels on highly constrained tests in order to prod states, districts and schools into doing a better

job. For accountability related to NCLB many states use standardized tests that are relatively easy to administer and score but have little connection to what a student learns as part of a high quality education.

The process of assessing student learning should be directly linked to *improving* student learning. Current large-scale assessments are only indirectly linked to student learning with assessment tasks mainly proxies for broader education goals. Much money has been poured into those assessments but there is no research evidence that they improve student learning. Improved scores on those tests may reflect improved test taking skills or the results of focused drills rather than increased learning of complex skills and concepts. Large-scale assessments give general information on student achievement and progress, but they do not offer reliable or valid information on the specific instructional needs of individual students.

To promote learning in the classroom, **we need better assessment tools that directly impact student learning –what are known as formative assessments**. By formative assessments we mean assessments that provide detailed information about student learning directly and clearly to both students and teachers. Commercial and interim assessments do not accomplish this. Formative assessments allow teachers to make adjustments to instruction and learning activities to ensure student achievement of learning goals. They also provide opportunities for teachers to develop essential assessment skills and deep knowledge of both standards and their students.

There is an urgent need to provide resources for professional development related to formative assessment use that, according to Randy Bennett of ETS and other proponents of assessment *of, for* and *as* learning, is an essential yet neglected component of assessment systems. According to Linda Darling Hammond, involving teachers in the process of producing and scoring formative assessments develop capacities for analyzing student work and performance and making instructional decisions based on complex student learning data. In addition, Darling Hammond points out that the provision of professional development through hands on involvement of teachers in the development, choosing, and scoring of formative assessments is a key common factor across countries whose students do well on international assessments. Finally, Darling Hammond provides compelling evidence for the incorporation of formative assessments into the US system with the fact that those countries excelling on international assessments such as PISA have complete assessment systems with substantive formative assessment components. The current US system is incomplete, lacking this component.

Formative assessments offer one of the most effective ways of building the capacity of teachers to analyze student work, plan and adjust instruction that focuses on the progression of learning and student needs, and to understand the nature of education goals. Again, the indications from high performing countries are that the quality of the teachers is supported by their participation in assessment development and scoring form the classroom to the state levels.

A complete assessment system should address the range of factors in learning, not simply enabling skills and basic content. It should provide opportunities for students to demonstrate the ability to apply skills and knowledge to real life situations and to demonstrate the essential dispositions for continued learning and participation in a democratic society. Formative assessments will not be employed, if the large scale assessments do not address the same level of skills and content as the formative assessments. Current multiple choice and short answer tasks that make up most large scale assessments make such alignment impossible. Such tasks do not tap students abilities to apply knowledge and skills to complex situations. A complete assessment *system* must components and tasks that produce 1) information to help teachers and students discern progress and areas of needed attention, 2) data that indicates whether schools and districts and states that support education are providing needed resources and policy support, and 3) a continuing effort to build the capacity of teachers to teach, students to learn, and parents and the public to support education.

The nation and the states are faced with limited resources for education and must us them wisely. It makes little sense to invest only in developing more large-scale summative assessments that have already consumed huge resources and continue to ignore the critical role of formative assessments. These tools are a key to connecting large-scale summative assessments to classroom practice. In terms of value for investment, formative assessment not only address a missing system component but they also provide opportunities for teachers to deepen their understanding of educational goals, students needs, and their own pedagogy. They serve the dual purpose of assessing

progress and building teacher capacity. They also provide a third benefit: students with better capacities for self assessment. I can't think of any employer who would not want their employees to be good at determining for themselves their success on the job.

Experts such as Ed Haertel and Eva Baker have already given compelling testimony that things need to be done differently in this country if our education system is to improve. NEA suggests that, for a change, efforts should start with accomplished teachers and the building of increased teacher capacity to educate through the development, use, and researching of a wide range of assessments and assessment tasks. Certainly we have the knowledge among our 3.2 million members to provide substantive, useful input into this critical effort to improve education through improving assessment systems. We propose that substantive attention to thee roles and development of teacher capacity be included in the requirements for this funding.

Center for Educational Improvement

PO Box 1355
Vienna, VA 22183
571-213-3192
Email: cmason@edimprovement.org
www.edimprovement.org

January 19, 2009

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

To Whom it May Concern:

Thank you for the opportunity to comment on the proposed Race to the Top Assessment Program. We applaud the Department for taking the initiative to establish an ambitious and competitive process to revise the standardized assessments. This is long overdue. For far too long children and teachers have suffered as states have systematically implemented a wealth of procedures that have dramatically revised the school calendar as schools, districts, and states have scrambled to show that students are scoring higher scores on outdated and antiquated tests. We are looking forward to an era with a renewed emphasis on TEACHING and INSTRUCTION.

In preparing this response, we have reviewed testimony by others and attended three of the public RTT Assessment meetings. We have also reflected on research regarding learning and achievement. As the Department proceeds with synthesizing all of the input it has received, we urge you to consider these factors:

1. Technology is available to present exciting options for student demonstration of problem solving, teaming, and application of skills. We hope the Department will encourage applicants to push the envelope in the use of these innovative technologies so that higher order thinking skills can be demonstrated in creative ways that reflect our everyday use of technology. Tests can be interactive and include diagrams, videoclips, and interesting visuals.

2. Research on human learning clearly demonstrates that when individuals are able to set their own goals for learning on topics of interest to the individual that more learning occurs. Knowing this, is it essential that each and every child be held to the same standards across a multitude of subjects? Or rather would better overall results be obtained if students were to identify perhaps 2-3 areas/subjects for intensive investigation and be tested at a higher level in these areas, with a more general test administered for other subjects? Let self-determination and student-driven instruction be the lynch-pins for this new program of assessments.

   We believe, and considerable research on student-centered instruction supports, that when students are more engaged that they learn more, more quickly. Recent research in the area of universal design for learning (UDL) and related brain research clearly shows that motivation and

affective centers of the brain are "lit up" when individuals are pursuing subjects of interest to them. CAST and others have described this in relation to Vygotsky's "zone of proximal development" –the optimal zone for learning where individuals are duly challenged but also encouraged by their rate of learning.

3. UDL also provides a way to facilitate test accommodations through building accommodations into tests, rather than adding accommodations after the test has been designed. Such an approach to test design would allow for the greatest consideration of the diversity of learners and their needs and strengths and would facilitate testing for individuals with disabilities as well as students with limited English proficiency (or English as a Second Language).

4. Those on the front line using the assessments and instructing students need to be involved in designing the assessments and reviewing templates for data summary and feedback to teachers. We recommend that input be obtained both from panels of teachers representing diverse perspectives and also from teachers who have received awards for their teaching excellence.

5. International exam and accreditation systems incorporate teamwork into their measurement of student skills --- teamwork is critical to our workforce and should be part of what is measured.

6. Formative assessment and authentic, curriculum-based assessments are key for measuring student learning. Rather than designing one test, flexibility is needed so that teachers and schools could choose to target the topics/subjects they are emphasizing. This would mean that in addition to a common platform of topics, that schools that specialize in the arts, in technology, in math, etc. would have options for showcasing their strengths.

7. The assessment design process with the consortia of states, state requirements for procurement, and the other Department of Education requirements should not impede our ability to create the very best tests. Sometimes the best designs come not from a group process with all the compromises that are often made, but rather from the genius of a single person or a small group of people. However, consensus will be needed at some point. Knowing this, we urge the Department not to burden the creative process or dilute the final product through rigid adherence to lock step efforts of a cross-state consortia. Instead, here are some alternatives:

    a. The Department might be able to implement a 2 or 3 step process, perhaps by supporting a selected group of researchers, test designers, and educational experts to come together for 2-3 day meetings (or perhaps a series of meetings) to consider options and barriers and come to conclusions regarding the creative aspects PRIOR to the state consortia process. Perhaps then the results of these meetings could serve as a springboard for the state consortia.
    b. A vehicle like the SBIR (Small Business Innovations Research) process might be used to fund small scale pilots prior to widescale design and implementation.
    c. Could other ways be found to phase-in the implementation so that adequate feedback and consensus is found prior to final development?
    d. States should not be required to stick with the assessment vendors they team with. That would be the equivalent of telling consumers to buy a Ford simply because it was made in the US rather than allowing performance to be the yardstick for marketplace endorsements.

e. Some states may want to sign on to 2-3 consortia as part of different approaches to being involved and designing tests. While initially they may be able to sign on to more than one team, the number of sign-ons should be limited – this would require states to use some discretion in their endorsement process. Perhaps the competition could include a pre-ap phase where up to 3 sign-ons are allowed for individual states. At the point of the invitation for the final application, states should be allowed to only sign on to only one plan.
f. At least three national projects should be funded.

In closing, thanks again for the opportunity to overhaul our archaic assessment systems. We are looking forward to this next era of educational assessment and educational innovations. Please contact me if I can be of assistance in considering these important next steps.

Sincerely,

Christine Y. Mason, Ph.D.
Executive Director

8

assessment
solutions group

Statement for Race to the Top Assessment Program

The Cost of New Higher Quality Assessments: A Comprehensive Analysis of the Potential Costs for Future State Assessments

John F. Olson, Ph.D.
Ed Roeber, Ph.D.
Barry Topol, CPA
Assessment Solutions Group
January 20, 2010

In our opinion, the RTTT and Common Core Assessments are two key initiatives for improving education and states will have an opportunity to receive some much needed resources and assistance to help them make important improvements to their assessments and create a higher-quality assessment system. State assessments need to be improved so that they do a better job of measuring the critical skills students will need in the 21st century, are integrated into the curriculum, help students learn, and provide teachers opportunities to develop new instructional strategies. Higher-quality summative assessments should be well-designed, include more constructed response and performance items, and be affordable to states. It is critical that states be able to afford the long term, ongoing administration costs associated with any new assessment system. Should states design a new assessment system without a thorough review of costs and affordability prior to project implementation, the sustainability of any improvements to the current system of assessments could be in jeopardy.

The Assessment Solutions Group currently is conducting a study of the costs for various assessment designs, including different models and options for higher-quality state assessments. With the financial support of a grant from the Nellie Mae Foundation, we are collaborating with Dr. Linda Darling-Hammond and her colleagues at Stanford University and are in the process of completing a research paper which focuses on the design and costs of a higher-quality

Assessment Solutions Group - "*your* assessment experts"
428 Parker St.
Newton, MA 02459
210-859-5593
jolson@assessmentgroup.org
www.assessmentgroup.org

assessment system that will meet the federal characteristics of a next generation assessment system, yet be at an affordable price to states.

The research we are conducting for this study and the preliminary results reported in our draft paper demonstrate that under the right conditions, states can dramatically improve their assessments and develop a higher-quality assessment system at an affordable cost. However, states must be careful to design an efficient assessment system and understand its ongoing administration costs, as well as future state budget allocations prior to committing to an innovative, high-quality assessment and implementing it in their state. Professional help in all these areas is highly recommended.

Based on our detailed cost analysis, the overall finding is that a higher-quality assessment, one that meets the criteria spelled out in the USED Design of Assessment Systems, can be done for about the same money as is spent for a traditional state assessment today. Our draft paper provides more details of the findings from the study that we are conducting, which follow:

- Our preliminary results indicate that such a higher-quality assessment system can be developed that has increased numbers of constructed-response items, performance-based items, teacher scoring of these items, and other features, and will cost roughly the same, or in some cases less, than what a typical state pays for an assessment today.

- It was found that the development cost of a new high-quality assessment is relatively inexpensive relative to the total cost of the assessment. The ongoing administration costs are by far the largest component of any assessment system. A key factor in the sustainability of new and improved assessments, and whether or not states can adopt and use them, will be the ongoing administration costs that need to be managed. It also is

critical that the next generation of assessments be designed so they are affordable on an ongoing basis, after federal grant money is exhausted.

- Of the cost reduction strategies that were examined, teacher scoring (with and without stipend) and formation of consortia were the two most important means of controlling overall assessment cost. While there are issues to be tackled regarding the implementation of teacher scoring and the formation of state consortia, both strategies should be a part of the plans for any future assessment system.

- In order to share costs across states, it will be important to have states participate in assessment consortia to share the overhead associated with development and management of assessments. Larger consortia are more cost effective, although the majority of cost savings relative to a single state case can be seen at the 10 state consortia size. Cost savings of over 20% (of total assessment costs) can be attained even at the 10 state consortia level vs. a state going it alone. Participating in a consortium allows for the spreading of the fixed costs of development and overhead functions over larger numbers of states and students.

- To help control costs, teachers will need to score the performance items. This practice is used in many high performing countries and helps build teacher expertise and understanding of student performance expectations.

- The use of computer-based technology is recommended but not required to achieve the savings needed to implement a new higher-quality assessment at current prices. Online assessment systems can save significant costs vs. paper-and-pencil based test administration. Additionally, distributed scoring of CR items can help save on costs.

Finally, the use of computerized artificial intelligence scoring holds significant promise in making high quality assessments more affordable.

- A clear and efficient assessment design must be created and specified in an equally clear RFP. Such a design can minimize costs, particularly development, printing, and logistics costs. One of the causes of expensive assessments today is that RFPs are generated without clear specifications. This leads to confusion among vendors when they respond to the RFP, as well as when the program plan for implementation of the assessment is developed, which ultimately leads to higher costs. In addition, RFPs for state consortia services should allow for multiple bidders and not be organized as sole source contracts.

In the paper, an average cost of $20/student was calculated for a typical assessment and $55/student for a high-quality assessment. Costs for the high-quality assessment were reduced to about $40/student with state consortia, and to less than $20 with teacher scoring of the items. Including all cost reduction strategies that were analyzed could bring the cost down to only $10/student.

In summary, we believe that states need to work together to create the next generation of higher-quality assessments. More work is needed to make efficient new designs that are affordable and meet federal requirements. This is not an easy task and will require dedicated resources, use of outside experts, and willing and cooperative states. States need to know what they will be spending for an assessment prior to engaging in the actual development and implementation of a plan. State consortia proposals should include a detailed test design, as well as an independent, detailed cost analysis to ensure that future assessments are affordable and sustainable. We hope that the findings from our ongoing study of assessment costs will help shed more light on what states can do to create higher-quality assessments.

Assessment Solutions Group - "*your* assessment experts"
428 Parker St.
Newton, MA 02459
210-859-5593
jolson@assessmentgroup.org
www.assessmentgroup.org

CTB McGraw-Hill

Remarks on General and Technical Assessment Topics
Race to the Top Public Meeting

Richard J. Patz, Ph.D.
Vice President, Research & Product Development
CTB/McGraw-Hill

Washington, DC
January 20, 2010

Founded in 1926, CTB/McGraw-Hill has a long track record of innovation and distinction in serving the needs of students, educators, and policy makers with high quality, valid and reliable assessments. Right now, for example, we are delivering online writing assessments with artificial intelligence scoring for both formative and summative state tests, technology-enriched interim assessments in some of the nation's largest school districts and states, vertical scales that support growth models in a number of state NCLB programs, and language assessments that meet the Title III needs of a consortium of states. No organization has greater expertise and experience in building high quality tests of achievement on nationally appropriate, K-12 standards, and conducting the rigorous large scale research studies required to validate such instruments. We care deeply about the education and testing of our nation's students, and we appreciate opportunities now and in the future to help shape meaningful education reform.

To address the questions on General and Technical Assessment issued in advance of the meeting, I offer these observations for the Department's consideration:

1. Summative assessments—those that have real consequences for students, teachers, and schools—must be standardized and validated. They should be built in compliance with the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), reviewed and approved by appropriately credentialed technical advisory committees, and administered with appropriately high levels of test security. Assessments meeting these requirements may measure new Common Core Standards, may include diverse item types eliciting complex student performances (see Yen & Ferrara, 1997, for example), and may be designed in innovative ways (see Patz, 2006, for example).
2. If the "through-course" assessments under Department consideration could be built and used as "interim assessments" and not used for summative purposes requiring standardization, then important goals might be achieved sooner. Such interim assessments can be thoroughly researched, scaled, and linked to summative tests through appropriate data analyses. Freed of the burden of standardization and validation required for summative use, these tests could be used much more flexibly and in a variety of ways (e.g., administered at different times). Test security would be much simpler, teacher professional development

could be enhanced, and the pedagogical value of a comprehensive summative assessment (e.g., at end of course) would not be lost.

3. It is more important that tools for formative and interim assessments integrate seamlessly with instructional practice than it is for these assessments to integrate seamlessly with summative tests. The tools that enable a teacher to utilize student performance data in real time or near real time and adapt instruction accordingly are distinctly different from the tools required to securely administer a high stakes large scale assessment. Although it is possible to create one system incorporating all these tools, separate systems may be more feasible and optimal.
4. Opportunities to improve the current K-12 state testing landscape abound. In this endeavor much of the finest talent available to help guide this improvement resides in the testing organizations. Finding ways to engage this talent, learn from past successes and failures, and leverage existing capabilities, will be critical to designing future assessment programs that are truly innovative, technically defensible, and feasible.

CTB/McGraw-Hill supports the goals of the Common Core Standards initiative and the Race to the Top grant program. We look forward to opportunities to support the U.S. Department of Education, the states and their consortia, and the agencies and organizations that are advancing these important goals.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing.* Washington, D.C.: American Educational Research Association. 1999.

Patz, R. J. (2006). Building NCLB science assessments: Psychometric and practical considerations. *Measurement: Interdisciplinary Research and Perspectives, 4*, 199-239.

Yen, W., and Ferrara, S. (1997). Maryland State Performance Assessment Program: Performance assessment with psychometric quality suitable for high stakes usage. *Educational and Psychological Measurement, Vol. 57*, 60-84.

Race to the Top Assessment Program
In Response to Request for Public and Expert Input
Topic 3.C. General and Technical Assessment

Lauren B. Resnick

Distinguished University Professor of
Psychology and Cognitive Science
and
Learning Sciences and Education Policy
University of Pittsburgh

Director
Institute for Learning
Learning Research and Development Center
University of Pittsburgh

The Problem: The "Testing Bind" That is Lowering the Quality of Instruction in America

Over the past two decades, our country has been trying to build a standards-based accountability system as a foundation for a more equitable and higher achieving education system. In practice, however, we have created a *test-based* accountability system that does not reflect the standards we aimed for at the beginning of the 1990s, much less today's emerging Twenty-First Century Skills. The state tests, on which the No Child Left Behind Act (NCLB) rests, are not well aligned with state standards. These tests, with only a few exceptions, systematically overrepresent the most basic skills and knowledge and omit the complex knowledge and reasoning skills we are seeking in the Twenty-First Century.[1] The misrepresentation of standards by most current accountability tests has had negative effects on teaching and instruction, especially for poor and minority students. The tests carry consequences, and many educators serving poor students aim to raise test scores in the most direct—in some cases, the only—way they know: They provide practice on exercises that substantially match the format and content of their state's end-of-year accountability tests. Principals and district administrators encourage this practice. They provide practice materials, and they introduce *interim*

assessments that largely mirror the end-of-year tests rather than modeling the kinds of performance intended by the standards. Though no one intended to do so, we have created a "testing bind" that, as it tightens, drives attention away from the intended standards. The effects are greatest in the poorest schools.

The Solution: An Educative Assessment System

It is possible to rectify the testing bind by creating an *Educative Assessment System* that, taken as a whole, provides reliable accountability for student, school, and educator performance; models the kinds of instruction that are valued; and provides detailed guidance to teachers as they work to ensure that each individual child meets the standards and expectations. There is little likelihood that any single testing instrument can meet all three of these goals. But an Educative Assessment System can meet all three and can move overall education performance upward.

An educative assessment system would meet the four main goals of assessment—one, modeling the kinds of student performances we aim for; two, providing diagnostic tools for instruction that meet individual student needs; three, holding the education system accountable for its results; and four, certifying student competence—through a *system* of assessments that are linked but do not directly mirror one another. Consider a system made up of the following assessments:

- *End-of-year summative tests* (*end-of-course exams in high school*). These would likely be mostly machine scorable for cost and technical reasons,[2] thus limiting the extent to which the tests can align well to standards that call for high levels of cognitive processing. Experiments with new technologies could change this limitation over time,[3] but it will likely be several years before these new approaches are ready for widespread use. Meanwhile, intelligent use of current forms of summative testing can be used as *part of* an educative assessment system. But it will be important not to mimic them in interim assessments and workbook materials.

- *End-of-unit examinations*. The current interim assessments that mimic end-of-year accountability tests would be discontinued in an educative assessment system. In their place, states can implement end-of-unit performance examinations that model the kind of high cognitive demand instruction intended by the state's standards. To build such end-of-unit exams, states will need to adopt curriculum and instruction based on the state's standards that spell out what should be taught, and perhaps provide detailed guidance to schools wishing to actively prepare their students for the exams. This is a logical step in ending the testing bind: Instead of supporting the use of practice materials that mimic end-of-year tests, states can provide instructional tools that help teachers prepare students for the end-of-unit examinations.[4] The end-of-unit examinations can and should include extended written work and others kinds of performances. Student responses can be graded locally (teacher participation in grading is a good form of professional learning). Teacher grades can be

validated using one of a number of methods that have been used in other countries (e.g., cross-school or cross-state grading exercises; re-grading of a sample of student papers at the state level). Validated interim examination grades can constitute part of a school's accountability score.[5]

- *Formative assessments.* These are locally administered assessments given in the course of teaching. They are intended to guide the teacher (and perhaps students themselves) in next steps for individual students. They may consist of tools for recording observations of students at work (e.g., listening to students read; evaluating the quality of small group discussion) as well as brief exercises built into the teaching process. There is no need for formative assessment data to be processed centrally; what matters is the swift and intelligent flow of information back to teachers, perhaps accompanied by detailed suggestions for instruction. Traditional criteria of psychometric reliability and generalizability do not apply because no major, long-term decisions will be made on the basis of formative assessments in educative assessment systems. Formative assessments should carry no "stakes"—for student, teacher or school.

The Opportunity: An Open Window for an Educative Assessment System

The current moment in American education reform presents a unique opportunity for creating one or more models of educative assessment for this century. The opportunity is created by the emerging common core standards. Mathematics and literacy standards will soon be released for state consideration. There is a high likelihood that many states will adopt these standards and will then be seeking to build consortia for implementing them. If states proceed as in the past—directly from standards statements to end-of-year tests—it is entirely possible that the new tests will not be substantially different from current ones, and that the testing bind will continue unabated.

But the adoption of common core standards also contains the opportunity to move in a new direction. The standards will outline a rough set of expected *learning progressions*—perhaps four to six expected topics to be taught, student competencies to be achieved—for each school year. These progressions will represent the best wisdom of practice that we now have. The standards-based progressions can be used to get to work on defining curriculum/instruction guidelines that can form the basis for new end-of-unit examinations. Research can be *built into* the plans for creating and using these exams, both to guide instruction and as part of accountability; we do not have to wait until the research is complete in order to begin. Meanwhile, new formative assessment systems are coming into use, some making maximal use of new technologies that will likely transform how we use data and information in education. As a nation, we are ready to assemble new teams and make real progress toward an educative assessment system.

The toughest part of getting started may be in the formation of consortia of states, assessment and examination developers, and organizations that can assist in implementation of the new systems. Progress is underway in each of the parts of the educative assessment system (summative tests, end-of-unit examinations, formative assessments)—but there is no one organization that has a track record in all three of them. Existing organizations will have to join in new associations. States may wish to work with more than one of these as they develop their educative assessment systems. The request for a proposed Race to the Top Assessment Program should therefore allow for maximum flexibility in assembling the components of an educative assessment system, while still asking for evidence that the new consortia will likely be able to design and implement a real *system*¸ not just a collection of parts.

---

[1] The problem cannot be fixed by changing cut scores, so that states no longer deem as being *proficient* test performances that barely meet NAEP standards for *basic* levels of achievement. The tests are fundamentally misaligned with Twenty-First Century expectations. For an analysis, see Resnick, L. B., Stein, M. K., & Coon, S. (2008). Standards-Based Reform: A Powerful Idea Unmoored. In R.D. Kahlenberg (Ed.). *Improving On No Child Left Behind: Getting Education Reform Back on Track*. New York: The Century Foundation Press.

[2] In current testing technology, a large number of independent items are needed to ensure generalizability of test scores.

[3] See National Academy of Education. (2009). Education Policy White Paper: Standards, Assessment, and Accountability. Washington, DC: Author. (Available at www.naeducation.org).

[4] For a description of approaches to providing this kind of instructional guidance in forms that do not suppress teacher ingenuity and judgment, see Resnick, L.B. (in press). Nested learning systems for the thinking curriculum. *Educational Researcher*.

[5] For example, 60% of the score might be based on traditional-format summative tests; 40% on cumulated interim examinations.

Comments on General Assessments – A Parent Perspective

LAUSD Parent Collaborative
1360 W. Temple St., Rm. 104
Los Angeles, Ca. 90026
Phone 213 481 3350
Fax 213 481 3392

Bill Ring
1st Vice Chairperson
& Director, TransParent®
lausdparents@aol.com
Phone 310 600 2015

I would like to spend my time today addressing the last question posed by the Department : Are there other issues that need additional focused research?

Initiatives come and go. Who thinks about asking students how they like to learn?

Discussions about standards and assessments reminded me of something I contributed to an online dialogue about the California Master Plan in 2002. At the time, I had just helped an 8th grade student in his research for a paper he was doing on education. His last question to me was this: " how do people decide what kids need to learn? I added two questions to that: WHO are those people and WHY do they have such authority?

I told him that there are philosophical differences among people around the world as to what children need to learn, when they need to learn it and who should teach them. I even asked him how much of what he learns does he think he will recall? How much of what he learns will he apply? How much of what he needs to learn does he already know?

Who, exactly, is setting the standards? Regardless of the answer, I don't think that raising standards in the existing system is going to change it into anything other than what it is.

Don't we want to teach our children to think critically for themselves, to make better decisions than we've made? Don't we want our kids going to schools that encourage and foster the different ways to make learning happen? How will such a focus on testing and standards ensure such an outcome? I hope we're learning to teach our children to look for meaning in their lives and not just disconnected facts."

I said that eight years ago – and here we are today and for me, the same questions apply, the same concern over narrowing of the curriculum, the same concern expressed by others in the question, " Are we measuring the wrong end of the student? (seat time).

Someone far greater than me once said that education is not about the filling of a bucket but the lighting of a fire.

You ask, "Are there other issues that need additional focused research?"

Some parents with students at our high schools are pleased because the teachers like their children yet some of these same students are receiving 3.5 GPAs and higher yet they cannot pass the state's high school exit exam. This should be exhibit one in the argument that parents don't know what questions to ask or what to look for and worse, that the school, the district and the state have failed to build the requisite capacity in the school community to fully comprehend standards and assessments and what

the students need to and are expected to learn. It also exposes grade inflation and begs a broader set of assessment questions for me.

Parents are not generally welcomed as partners in policy making and decision making in LAUSD. As perhaps a provocative example, when it comes time to talk about wages and benefits and working conditions for the teachers who are teaching our children and administrators who are running our schools, union representatives sit down with the district to talk about such things but the parents are not included in the dialogue. We and our children are the clients and we need to be included in these discussions which are directly connected to student achievement and our budget crisis. All the talk about teacher quality, employee evaluation, seniority, etc. raises some questions for parents about how districts deal with collective bargaining and what is truly in the best interests of children.

Schools also don't always educate parents to become equal partners in shared school-site decision-making councils; they often just don't want us involved.

Last May, I was a part of a group of parents presenting to the California State Board of Education about the inadequate training of parents which has led to a lack of capacity and frankly, a lack of trust. Parents routinely see little monitoring and evaluation of decisions to track results and as we all know you cannot manage what you don't measure. Schools often fear having candid conversations around data and when they do wade in, there is difficulty in having a common conversation with all members of the school community.

Practice is a reflection of what we value.

I am here to make the case for the need to include parents – and secondary students - in the assessment and evaluation of teachers, principals and other administrators and yes, even superintendents. We parents are consumers – like everyone – but in this case, we're consumers of educational services provided to our children. In the commercial marketplace, when products and services are introduced, it is typical that consumers are engaged in such development. Their opinions are sought, surveys are taken, user groups are assembled and polled and certainly in the standards and assessment business, providers must recognize that users and developers work together to seek solutions.

Where in RttT is the commitment and specific acknowledgement and recognition of the need to build capacity in our parents and in our broader school communities?

I would like to see pilots of a school inspection system, an "inspectorate", if you will, as suggested by California's Stephen Blake. Blake speaks of a state-specific inspection framework but perhaps this idea could be introduced as a part of or in conjunction with Randy Bennett's school-based model for assessment innovation… even while recognizing Tom VanderArk's conclusion in 2005 that "giving failing schools autonomy is a bad idea". We must change the culture to embrace, train and trust parents and families as partners in this work.

Here are a couple of additional candidates for research by the Department and consortia:

The draft of California's RttT plan dated January 5, 2010 included a bullet point to "Increase choice and empowerment of parents" and stated that "California is encouraging…parent-school partnerships to improve student achievement…" It goes on to say that parents in the lowest-achieving schools as defined by RttT will have the option to apply to transfer to a higher performing school in another district…"

Is opting-out an example of partnership? Instead of abandoning the system, we should be driving it.

“Furthermore”, the draft states, “parents in schools that are in multiple years of program improvement can trigger one of the turnaround options under RttT”. If one were to rank the lowest performing schools in California now, one would find some charter schools near the bottom of the list.

Scott Marion asks, “Do other forms of school reform work better?”

Well, parents with students in failing charter schools have **already** pulled the trigger yet the schools are still failing their children - NOW WHAT?

The Chancellor of the California State University system publicly stated at the launch of CSU’s Center to Close the Achievement Gap last summer that less than 50% of the algebra teachers in California had ever taken a course in it or were taught to teach it. Would you consider this important information when considering student assessments?

Has there been any thought given to the time, resources, the learning curve in successfully adopting any of the suggested courses of action with respect to assessments of any kind? A new district bulletin will not change district culture.

If we parents do not take charge of our children's education, no amount of "reform" is ever going to fix the schools. Unless parents get organized and get informed, no amount of money is ever going to be enough. By the way, I would deeply appreciate someone showing me the studies proving a connection between increased spending and increased learning?

PEARSON

Jon S. Twing, PhD
Executive Vice President
Assessment & Information group of Pearson

Pearson
2510 North Dodge Street
Iowa City IA 52245 USA
Telephone: 319 339 6407
Fax: 319 358 4224
jon.s.twing@pearson.com

January 20, 2010

Office of Elementary and Secondary Education
Attention: Race to the Top Assessment Program—Public Input Meetings
U.S. Department of Education
400 Maryland Avenue SW
Room 3W339
Washington, DC 20202

To Whom It May Concern:

On December 24, 2009 the U.S. Department of Education issued notice of additional public and expert input meetings and request for input on the Race to the Top Assessment Program. The notice includes questions under the topics of Project and Consortium Management, Procurement, and General and Technical Assessment. We thank the Department for the opportunity to comment. Below are recommendations from Pearson on each topic.

Project and Consortium Management
While there are a number of consortium models that states may consider when forming an assessment consortium, given recommendations provided by expert witnesses at the January 13-14 public meetings, and our own experiences, it is worth replicating and building on the American Diploma Project (ADP) Algebra I & II consortium model. Key lessons learned from the ADP Assessment Consortium are described in the June 2009 panel presentation from the Council of Chief State School Officers (CCSSO) National Student Assessment Conference attached to this document.

The ADP consortium model includes the following desired characteristics.

- Organized by a lead state
- Operated by a consortium board (one member from each state's department of education)
- Managed by a project manager (third party convener to the states and manages the contractor)
- Ongoing participation of K-12 and post-secondary educators
- A focus on post-secondary validity research

Ohio serves as the lead state and contracts directly with Achieve for external management of the consortium and with Pearson for assessment development and delivery. The participating states executed a multistate

procurement participation agreement amongst themselves, a memorandum of understanding with Achieve, and a task order with Pearson to purchase the assessment. Nine states constituted the original consortium; which has expanded to include six additional states since its inception. A similar procurement model would work well for a RTTT common assessment consortium which may also grow over time as states transition from their current assessment programs.

For any new consortium formed, ideally the states will have agreed on the purposes and goals of the assessment prior to entering the consortium and will therefore be participating as equal partners (e.g., one vote per state). In consortia where there is variation in the level of usage by states (e.g., one state makes the assessment a district option and another state requires the assessment) or the stakes attached to the assessment (e.g., no stakes versus used for accountability), it may be necessary to establish tiers of participation levels whereby some states are allowed more weight in decision-making than other states. Detailed governance structures need to be in place early to define how decisions will be made.

Due to the size and complexity of consortia that are likely to form from the RTTT common assessments it may be necessary for consortia to:

- Be organized by more than one lead state (e.g., one lead could focus on K-8 assessments and the other lead could focus on high school assessments).
- Utilize multiple assessment contractors (preferably a prime contractor with one or more subcontractors).
- Create steering committees among state members to ensure adequate progress in key areas (e.g., technology and innovation, integration of curriculum and assessment, professional development, and use of data for decision making).
- Fund assessment development through the consortium (versus funded through the contractor) to establish a fixed price model.

The more committed the consortium members are to working as a unit, and share common purposes and intended uses for the assessment, the more efficient and successful the consortium will be. Significant amounts of customization by individual states will take-away from the benefits of working as a consortium and the ability to effectively compare results. Financial gains will also be reduced as the variability increases amongst the consortium.

Procurement

We agree with the competitive procurement process recommended at the January 13-14, 2010 public input meetings:

- States establish consortia and apply for RTTT assessment grants.
- Each consortium *develops* an RFP to procure a contractor or contractors for its common assessment program.
    - The RFP would include all of the standard terms and conditions which are mandated for the participating states, so the participation by each state's procurement and Attorney General's office is vital at this stage in the process.
    - The RFP would be led by the lead state.
- The Department awards funds to qualified consortium assessment applicants.
- Successful consortia *release* the RFP for contractors to bid on.

  - Sufficient time must be allowed to contractors for the Q&A process, as multiple states are involved extra time should be allotted.
  - The deadline for submission of proposals should take into consideration the challenges that contractors will encounter in responding to a proposal of this magnitude.
- Contractors submit proposals to the lead state.
- Consortia award vendor contracts.

Based on our experience, at least two months should be allowed from release of the RFP to submission of proposals by contractors, to allow sufficient time for Q&A and for contractors to include any answers received during the Q&A into the design of their response and bid. Following the submission of responses, the consortium may need one to two months to select a contractor or contractors, and then an additional two months to harmonize the terms and conditions and finalize the contract. Given that the Department likely will award RTTT assessment grants in September 2010, we anticipate most vendor contracts will not be in place until late in the first quarter of 2011.

When developing RFPs to select a contractor, states should include specific requirements for standardization, comparability and quality. For example, the management system supporting the constructed response scoring system should have industry recognized certification and follow processes to promote rigor and consistency, such as International Organization for Standardization (ISO) certification. When procuring technology solutions, states should require systems to be open in order to avoid becoming locked in to any single vendor's products and services. The most effective way to ensure vendor independence is to pursue interoperability. Interoperability, when used to describe computing systems, implies that they were designed to work with other systems which adhere to a common standard for communicating. There are two common approaches to achieving vendor independence, open architecture and open source software. The Department should encourage states to require open architecture not open source software.

Open source software is a software licensing model whereby the source code for the software product is made available in the public domain for anyone to modify as they see fit. The attraction of open source software is that the software can be procured without paying licensing fees to a vendor. Additionally, if the consumer of the software wants to enhance or extend the software product, they have the ability to apply their own software engineers to create the next version. The trade off for these benefits is that the consumer assumes the burden of ensuring the system is compatible with the rest of their computing environment. An example of open source software is the Linux operating system. There are many versions of Linux available as open source products. If you select a version which is in some way not compatible with your computer, you must either work directly with the vendor/distributor to correct the problem, or modify the product yourself.

Open architecture is a software design principle in which software products are built around industry standards so that components from multiple vendors can be interchanged without modification. With the open architecture approach, vendor independence is implicit in the model as each vendor's products are designed to interoperate with products from other vendors based upon industry standard protocols for data storage and communication. Web browsers and web servers are good examples of open architecture software products. Since browsers and servers are both built upon a set of standard protocols (HTTP), users do not need to know which vendor's web server is being used to host content. Similarly, the creator of the content does not need to worry about which

browser is being used to retrieve the content as the underlying architecture has been designed to ensure compatibility. Within the education domain, open architecture standards exist such as the Schools Interoperability Framework (SIF) and the Question and Test Interoperability Framework (QTI). There are many solutions available in the market that have been designed to interoperate based upon these standards, some of them are distributed under open source agreements, some are licensed through software vendors.

When establishing guidelines for procuring software for states or consortia, it is our recommendation that the guidelines pursue vendor independence through the use of open architectures versus open source software. With interoperability as the primary means to achieve independence from any single product or vendor, we believe mandating open source products will limit the number of available solutions unnecessarily without specifically addressing the goal of vendor independence.

General and Technical Assessment

Both the federal government and the States have invested a significant amount of time and money in existing No Child Left Behind (NCLB) assessments and accountability systems. Some amount of preparation and transition will be required before states are able to support and implement the new generation of assessment systems. In our December 2 response, we proposed a ***bridging strategy*** that will allow states to move from today's primarily paper-based, multiple-choice summative assessments to significantly enhanced summative assessments. Ultimately states would move to an integrated student-centered learning system with on-demand assessments throughout the school year that are linked with instruction and provide information for continuous teacher capacity building and professional development. This bridging strategy recognizes two core functions of educational accountability that must be preserved in the transition:

1. Provide individual student achievement data.
2. Support meaningful comparisons across schools, states, and internationally.

Our motivation for proposing a bridging strategy is to first prioritize transition to an online platform with significantly enhanced summative assessments as the primary assessment of record. This assessment would most likely start as a fixed-form, online criterion-referenced test administered at the end of the year, but would become more flexible and adaptive as we move into the future. This bridge assessment would initially be comprised of English language arts (ELA) and mathematics for grades 3–8 and a series of end-of-course assessments for middle/high school. The content domain for the bridge assessment would be defined by the common standards, and operationally defined through specific item and test specifications. The assessment, however, would have a greater focus on performance-based assessment and problem-solving, with items and tasks that—thanks to ongoing research and innovation in both item design and technology—produce information regarding students' academic knowledge and skills that until now have been difficult to assess with traditional multiple-choice tests.

As we push the envelope with new technologies in computer-based assessment to better capture and measure student performance, we do not want to be limited or constrained to strict comparability of traditional paper multiple-choice tests or formats; however we do want to ensure students are assessed in fair and equitable ways in both modes. It is not so much that we should seek paper-based assessments and online assessments to be comparable in a psychometric sense so much as we should make them compatible together in parts of a larger system.

Similarly, as we develop and implement new systems of assessment and learning including new measures, we need to routinely collect and use sound, research-based evidence to ensure the system and its components are valid, reliable, and fair so that we can continuously refine and improve the system without being mired by the past or abandoning rigor and quality.

If you have any questions or would like to discuss our response further, please contact me at 319-339-6407 or by email at jon.s.twing@pearson.com or my colleague Shilpi Niyogi, Vice President, National Services, at 202-434-0975 or by email at shilpi.niyogi@pearson.com

Sincerely,

Jon S. Twing, PhD
Executive Vice President
Assessment & Information Group of Pearson

Lessons Learned and the Road Ahead for the American Diploma Project (ADP) Assessment Consortium

CCSSO Student Assessment Conference

June 23, 2009

PEARSON

Achieve

Welcome and Introductions

- Laura Slover—Vice President, Content and Policy Research, Achieve
- Stan Heffner—Associate Superintendent for Curriculum and Assessment, Ohio Department of Education
- Gayle Potter—Associate Director for Curriculum, Assessment and Research, Arkansas Department of Education
- Bernie Sandruck—Mathematics Professor and Division Chair, Howard Community College
- Shilpi Niyogi—Vice President, National Services, Pearson

Agenda

- **Lessons Learned**
  - Forces shaping the American Diploma Project & the Assessment Consortium
  - State Leadership Perspectives from Arkansas and Ohio
  - The Assessment Vendor's Role as a Partner in Reform
  - Higher Education's Role
- **The Road Ahead**
  - Where do we go from here?

Forces Shaping the American Diploma Project & the Assessment Consortium

PEARSON

Achieve

Achieve and the American Diploma Project

- Achieve was created by the nation's governors and business leaders in 1996 following the first National Education Summit.

- Achieve is a bipartisan, non-profit organization that helps states raise academic standards, improve assessments, and strengthen accountability to prepare all young people for postsecondary education, work, and citizenship.

- Achieve currently is working with 35 states through the American Diploma Project (ADP) Network to design and implement policies that aim to close the "expectations gap."

The American Diploma Project

- The American Diploma Project (ADP) was created to ensure all graduates leave high school ready for college and careers.
- Early research by ADP sought to identify “must-have” knowledge and skills graduates will need to be successful in college and the workplace.
- Found a convergence between the skills that high school graduates need to be successful in college and those they need to be successful in a job that supports a family and offers career advancement.
- Developed ADP benchmarks that include the core content and skills in mathematics and English all students should have when they graduate high school.

ADP Key Findings

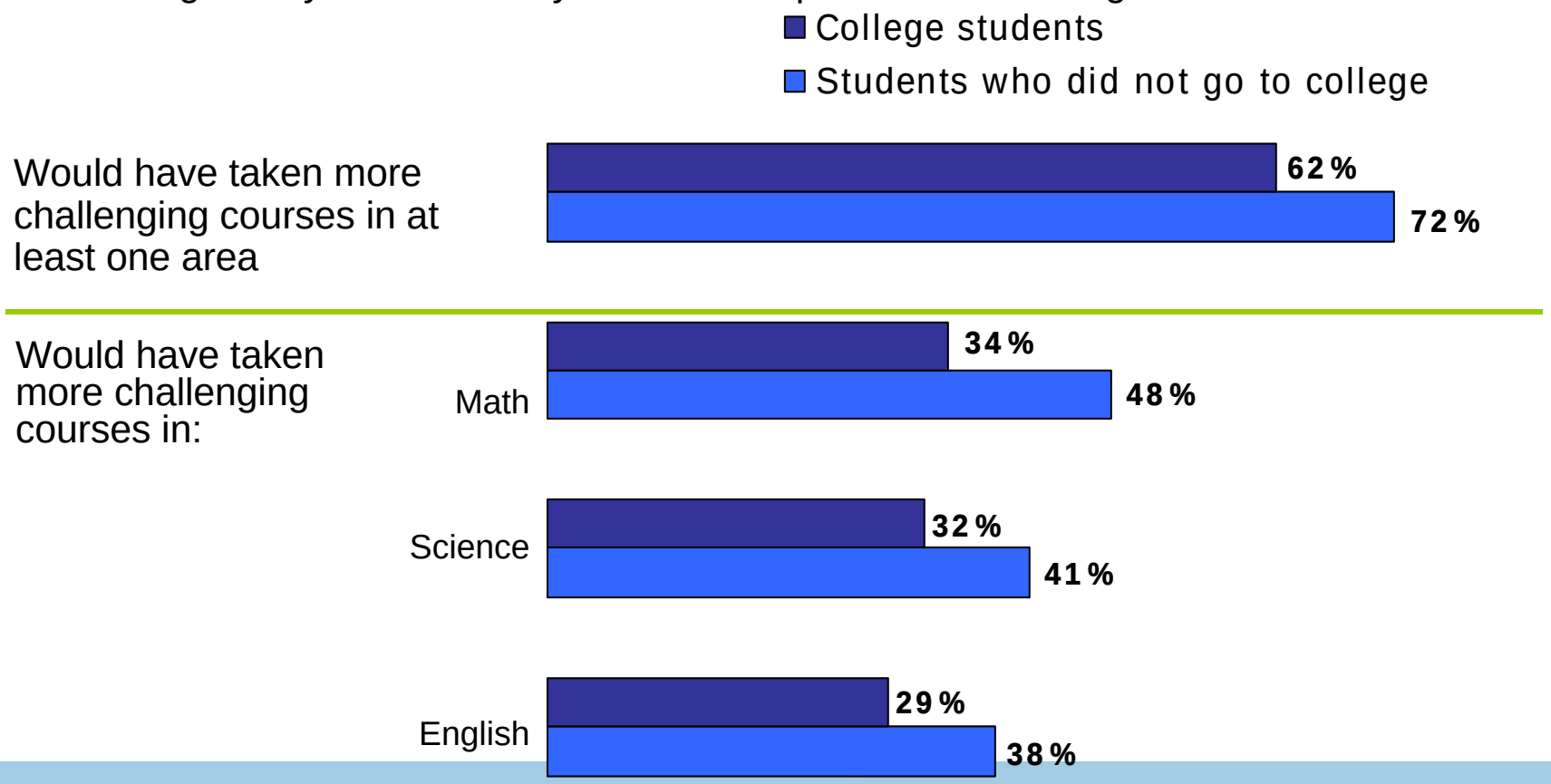
- In mathematics, graduates need strong computation skills, ability to solve challenging problems, reasoning skills, geometry, data analysis, statistics, and advanced algebra.

- Essentially, they need the content and skills *typically taught* in courses such as Algebra I, Algebra II and Geometry, as well as data analysis and statistics.

- In English, graduates need strong reading, writing and oral communication skills equal to four years of grade-level coursework, as well as research and logical reasoning skills.

The Majority of Graduates Would Have Taken Harder Courses, Particularly in Mathematics

Knowing what you know today about the expectations of college/work …

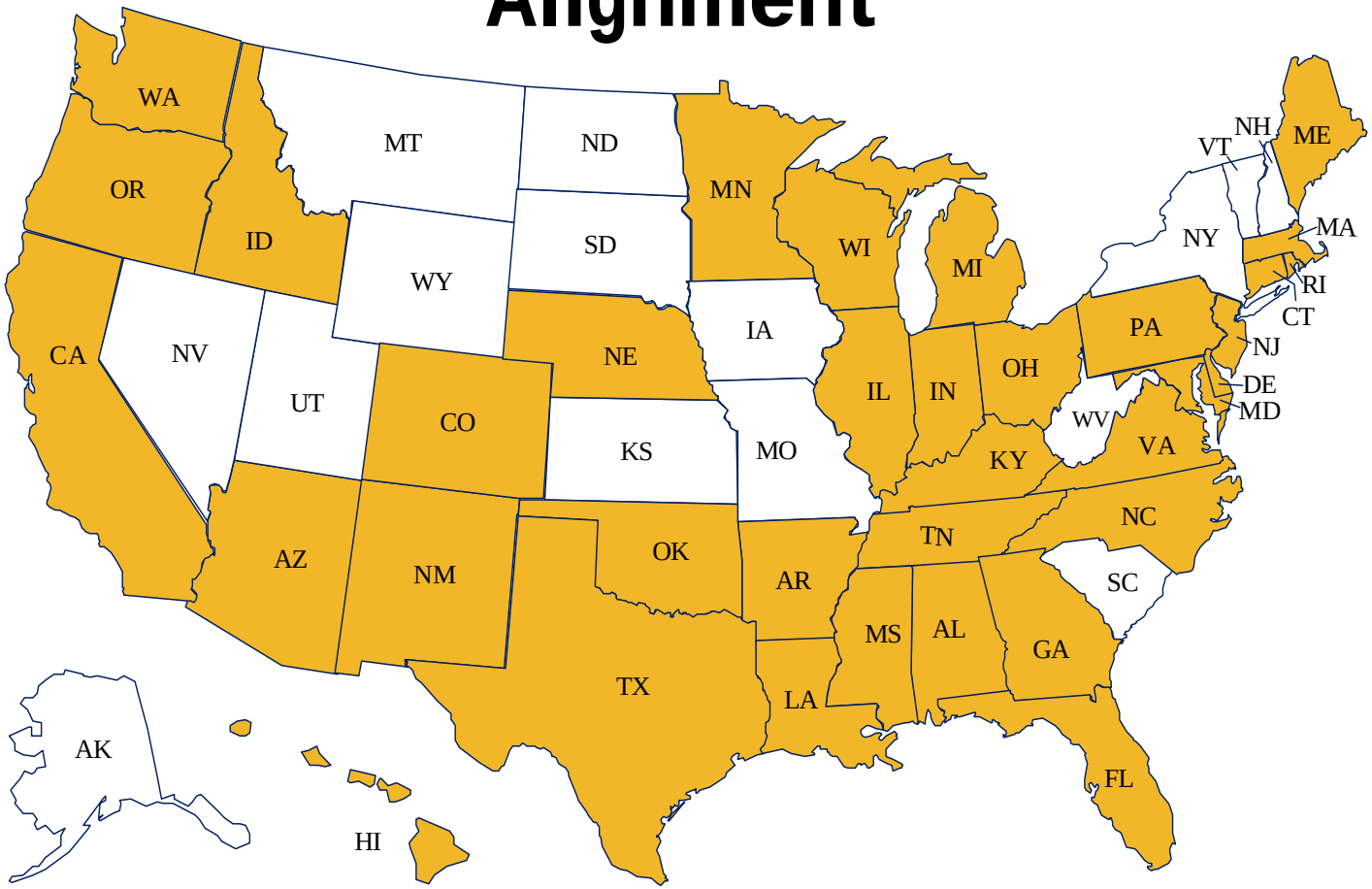


Source: Peter D. Hart Research Associates/Public Opinion Strategies. (2005) *Rising to the Challenge: Are High School Graduates Prepared for College and Work?* Washington, DC: Achieve.

ADP Policy Agenda: Close the Expectations Gap

- Align high school standards with college and career expectations.
- Require all students to take curriculum aligned with standards.
- **Include "college-ready" tests, aligned with state standards, in high school assessment system.**
- Hold high schools accountable for graduating students college- and career-ready, and hold postsecondary institutions accountable for student success.

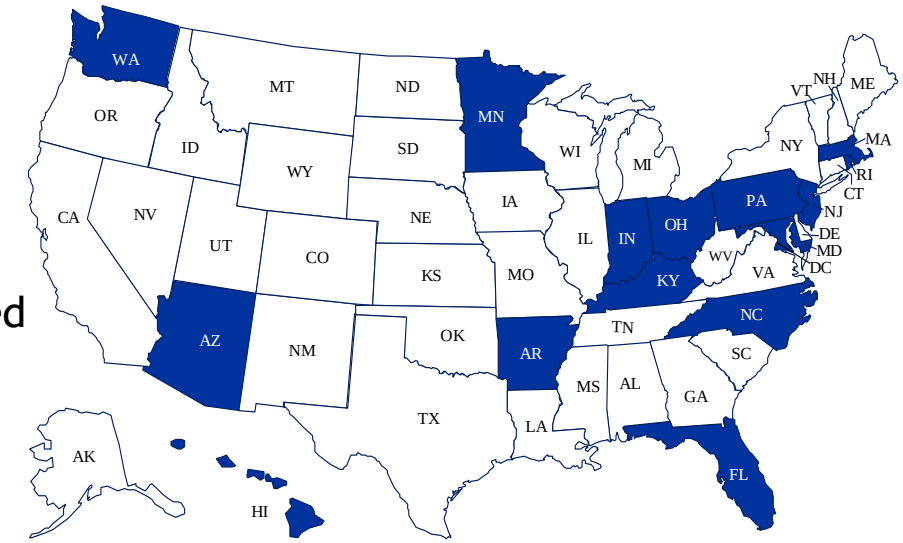
35 ADP States Working Toward Alignment



ADP Assessment Consortium

In summer 2006, nine states issued RFP for the development of an Algebra II EOC Exam:

- Arkansas, Indiana, Kentucky, Maryland, Massachusetts, New Jersey, Ohio, Pennsylvania, Rhode Island
- Ohio acted as "lead state" in unprecedented multi-state procurement arrangement
- Pearson awarded contract in 2007 with addition of Algebra I in 2008
- Since the consortium began, six additional states have joined: Arizona, Florida, Hawaii, Minnesota, North Carolina, and Washington.



More states are expressing interest in the exam and joining the consortium…

Purposes of the Common Exam

The ADP Algebra I and II Exams are designed to serve 3 goals:

- To improve high school curriculum and instruction and to help ensure rigor and consistency for the sake of equity
- To provide a common measure of student performance within and across states over time
- To serve as an indicator of readiness for more advanced course work, specifically:
  - The Algebra II EOC exam is designed to assesses readiness for first-year college credit-bearing courses, but not intended as a replacement for existing college placement tests

Exam Content Standards and Design

- States worked together to agree on test content and design specifications:
  - Involved state mathematics leads, high school teachers and two-and four-year higher education mathematics faculty
- The resulting ADP Algebra II End-of-Course standards are generally aligned with ADP mathematics benchmarks, which represent knowledge students should have to be prepared for postsecondary success
  - Performance levels will be common across states
  - Consortium cross-state report published annually

State Leadership in Building and Sustaining an Effective Consortium

PEARSON

Achieve

Consortium Structure

ADP Assessment Consortium States

Arizona, Arkansas, Florida, Hawaii, Indiana, Kentucky, Maryland, Massachusetts, Minnesota, New Jersey, North Carolina, Ohio, Pennsylvania, Rhode Island and Washington

OhioDAS

Achieve

PEARSON

The Coordination and Direction Team

- Includes assessment directors or other high-ranking policy-making officials from each member state in the consortium
- Oversees production and implementation of the Algebra I & II program
- Ensures that legal and policy needs of each state are addressed during team deliberations and decision making

The CDT and Decision Making

- The CDT states have:
  - Multiple representatives
  - One vote
- Process of consensus, collaboration, and compromise:
  - Relating to the policies of each state
  - Repeated as necessary
  - Less discussion time needed as process matures
- For legal matters, decisions must be unanimous

New State Members

- Membership is extended to any state wanting to join

- New member states must:
  - Agree to terms stated in the contract (but may add language)
  - Participate in all CDT meetings, item development and review meetings
  - Make firm, long range commitments
    - Contractual
    - Personnel
    - Financial

Ohio: Role of the Lead State

- Appoint a procurement officer
- Conduct procurement
- Provide guidance to the CDT
- Administer the contract
  - Manage contract amendments
  - Incorporate new states into the contract

Procurement Award

NASPO
National Association of State Procurement Officials

2008 Cronin Award for Procurement Excellence

Ohio Multi-State Procurement of Unique Services

Implementation Date: March 23, 2007

Originator: The multi-state Algebra II End-of-Course (EOC) Exam Contract is administered by the Ohio Department of Administrative Services (DAS) for the Ohio Department of Education (ODE). The purpose of the Project is to procure a common EOC Exam and related services as defined in jointly developed Project specifications, consistent with the needs of consortium States.

Submitted by the

OFFICE OF PROCUREMENT SERVICES
GENERAL SERVICES DIVISION

OhioDAS
Department of Administrative Services

Contact Name: Mark D. Hutchison, Sr., CPPO, CPPB, C.P.M.
DAS/GSD/Procurement Services
4200 Surface Rd.
Columbus, OH 43228
Phone: (614) 466-2375
e-mail: Mark.Hutchison@das.state.oh.us

Team Members: James F. Hunley, CPPO
Dana L. King, CPPB

Challenges of a Consortium

- Future funding
  - Riches to rags
- Algebra II (?) in high schools
- Achieving acceptance with:
  - Students
  - Colleges: Use as a placement exam?
- CCSSO “common core” curriculum
  - “Race-to-the-Top”
- Mode of assessments: Change in the air

Challenges of a Consortium

- States have different requirements for
  - Accommodations
  - Calculator usage
  - Data collection and reporting
  - Test Security guidelines and procedures

- Variances in curriculum standards exists among the member states

Ohio: Lessons Learned

- Start the process early
  - The legal process can be lengthy
  - Have a lawyer shepherd the process
- Can your state enter into a joint purchasing agreement?
- Plan for multiple agencies and levels of review and approval:
  - RFP
  - Proposals
  - Joint purchasing/contracts

Benefits of a Consortium

- Interstate and Intrastate Comparison
  - Valuable information within state and throughout nation

- Financial Efficiency
  - Item development
  - Processing
  - Better use of staff and resources

Benefits of a Consortium

- Information sharing
  - Joint problem solving to accommodate member states
  - Best practices and alternative approaches

- Proactive Policymaking
  - Prepares states for any future move to common standards
  - Advantage in “Race-to-the-Top” funding

Arkansas: Lessons Learned

- Important to have high-level state leaders involved from the beginning
  - Commissioner James served on the RFP selection committee, met with higher education officials in the state to discuss use of the exam, and promoted the consortium as a positive endeavor among his constituencies
- Alignment between standards and curriculum
- Importance of commitment in terms of participation and state funding

Arkansas: State Leadership Commitment

- Arkansas has aligned its Algebra II curriculum to ADP standards as one of the original ADP states
- This effort is a good fit with legislation and policy in Arkansas, especially in regards to the Smart Core Curriculum
- Dr. James was committed to working through the procurement issues among the states, determining the appropriate language in contract, and devoting resources from Arkansas to ensure smooth implementation
- Participation in this consortium has been a springboard for Arkansas' involvement in the state-led common standards movement now underway

The Assessment Vendor's Role as a Partner in Reform

PEARSON

Achieve

The Role of the Assessment Vendor as Partner in Reform

In close collaboration with the States and Achieve:

- Balance critical state-level needs and constraints with broader Consortium objectives in a pragmatic way
- Design and implement validity research to support standard-setting and longer-term program goals
- Facilitate communications with key stakeholders including teachers, students, parents, policymakers and the public at large
- Learn to be more flexible, nimble in response to evolving state and national education policies and practices

The Role of the Assessment Vendor

Pearson's scope of work on the Algebra I & II exams includes:

- Item and test development (retain ownership of items)
- Paper and online test administration
- Scoring and reporting
- Psychometric and research services

Strong Commitment to Research

- Achieve, the Assessment Consortium, and Pearson share this commitment
- Guided by the Research Alliance comprised of leading national experts from higher education, mathematics, psychometrics, and education policy
  - Particularly influenced studies that will inform Algebra II standard setting, given the college-ready component

Approach to Standard Setting

- Given the multiple purposes of the exam, in November 2007 the Research Alliance, recommended various validity studies prior to standard setting

- Standard setting was delayed until after the spring 2009 administration

- Benefits of this approach
  - Time for additional studies
  - Time for additional instruction
  - Time for a full implementation of the ADP Algebra II Exam

Algebra II Validity Studies Overview

Based on recommendations from the Research Alliance, Achieve and Pearson designed a number of studies to inform ADP Algebra II standard setting.

- Concurrent validity studies
  - State-level high school exams
  - National-level "college-ready" placement exams
- Cross-sectional validity studies
  - Predictive
  - Criterion
- Judgment studies

Other Research to Inform Standard Setting

Standard setting will also be informed by other research conducted by Achieve:

- Analysis of Algebra II standards across numerous states
- Aligned expectations
- International perspective
- Analysis of college syllabi from judgment study participants

Algebra II Validity Studies: Timeline

Task	Dates
Decision made to defer standard setting until after spring 2009 administration	Dec 2007
Validity research plan drafted	July 2008
Validity studies conducted	Aug 2008-May 2009
Spring 2009 administration	May 1-June 12, 2009
Standard setting	July 22-24, 2009
Spring 2009 scores available online	August 24, 2009

Standard Setting

- 2 ½ day meeting in Chicago: July 22-24, 2009
- Participants include one representative from DOE each of the 15 ADP Assessment Consortium states and 14 Achieve selected participants
- Briefing book model
- Each performance level will describe student proficiency in Algebra II and preparedness for the first credit-bearing college mathematics course
    - Well Prepared, Prepared, and Needs Preparation
- Panel will make a recommendation to Achieve
- Achieve will decide the final cut scores

Higher Education's Role

PEARSON

Achieve

Maryland: A history of collaboration
Between K-12 and Higher Ed

- Before ADP
  - Core Learning Goals/High School Assessments
  - Bridge Goals

- Higher Ed & the ADP programs
  - Began with -
    - Two professors of mathematics on the state ADP benchmark committee

Maryland: Higher Ed Participation in ADP projects

Roles of Higher Ed representatives:

- Convened Benchmarks discussion meetings
- Disseminated information
- Recruited participation in item development
- Recruited participation in research studies

Maryland: Higher Ed Participation

Maryland's higher education participation in the development of the ADP Algebra II Exam:

- Scope of content
- Analysis of potential test items
- Development of scoring rubrics
- Cross-sectional studies
- Judgment studies
- Exam standards setting

Maryland: Cross-Sectional Studies

- Sept 2008 predictive study (628 students)
  - University of Maryland College Park (Precalculus)
  - University of Maryland Baltimore County (Precalculus)
  - Howard Community College (College Algebra)

- Nov/Dec 2008 criterion study (331 students) (Intermediate Alg.)
  - Howard Community College
  - Cecil College
  - Anne Arundel Community College

Maryland: Cross-Sectional Studies

- Matched a student's ADP score to his/her final mathematics course grade to determine:
  - For predictive studies, how well a student's performance on the test predicts his/her performance in the math course
  - For criterion studies, how well the test assesses what is taught in the math course

- Final results from the study will be shared with the standard setting panel in July

Maryland: Judgment Studies

- One or more College Algebra or Precalculus professors participated in the ADP Algebra II Judgment Studies in March and April 2009:
  - Anne Arundel Community College
  - Baltimore City Community College
  - Montgomery College-Rockville
  - Montgomery College-Germantown
  - Morgan State University
  - Towson University
  - University of Maryland College Park

Challenges for Higher Ed

- Varying definitions of College Algebra
- Ripple effects to other courses
- Policy Implications for placement

Next Steps

- Implementation of algebra II test results in college placement
- Increasing statewide participation in high school
- Focus of P-20 on college readiness
- Routine, on-going collaboration

Results from 2008 and the Road Ahead

PEARSON

Achieve

Spring 2008 Administration: Who Took the Exam?

State	Number of Students Tested	Approximate Number of Students Taking Algebra II (or Equivalent)	Percent of Students Enrolled in Algebra II (or Equivalent) Taking Algebra II Exam	First Year Algebra II is Required for Graduation
Total	88,344			
AZ	1,091	No State Data Available	No State Data Available	2013
AR	22,101	23,949	92%	2010
HI	5,157	No State Data Available	No State Data Available	Not Required
IN	3,027	55,337	5%	2011
KY	2,019	80,400	3%	2012
MN	205	No State Data Available	No State Data Available	2015
NJ	9,813	43,000	23%	Not Required
NC	922	33,013	3%	2013
OH	33,611	128,576	26%	2014
PA	8,371	105,000	8%	Not Required
RI	1,853	7,000	26%	Not Required
WA	174	No State Data Available	No State Data Available	2013

*MD and MA did not administer the exam in Spring 2008.

State Usage Varies

- Some states require all students who take the course(s) to take the exam(s)
- Most states allow districts to make decisions about whether to require exam(s)
- Some states make the exam(s) part of certain state initiatives

2008 Results:
How Did Students Perform?

- Interpretation challenges in 1st year
  - the number of test takers varied significantly across the states
  - performance standards or "cut scores" have not yet been established

- Although scores cannot yet be used to compare one state's performance to another, the results of the first administration provided some interesting insight when taken in aggregate.

2008 Results

State*	All Items (Total possible points = 76)		Multiple-Choice Items (Total possible points = 46)		Constructed Response Items (Total possible points = 30)	
	Average Number of Points Correct	Average Percent of Points Correct**	Average Number of Items Correct	Average Percent of Items Correct	Average Number of Points Correct	Average Percent of Points Correct
Total	20	26.9 %	17	37.8%	3	10.2%
AZ	19	25.0%	17	37.2%	2	6.4%
AR	20	26.6%	17	36.9%	3	10.8%
HI	19	25.5%	17	36.8%	2	8.2%
IN	20	25.8%	17	37.7%	2	7.6%
KY	16	20.8%	14	30.9%	2	5.4%
MN	16	21.7%	15	33.1%	1	4.2%
NJ	21	27.5%	18	38.5%	3	10.7%
NC	27	35.3%	22	48.3%	5	15.5%
OH	21	27.9%	18	39.0%	3	10.9%
PA	19	25.4%	17	36.3%	3	8.7%
RI	18	23.4%	16	34.0%	2	7.1%
WA	24	31.4%	19	42.0%	5	15.2%

Achieve Annual Cross-State Report

- **Findings**
  - Student performance was low across all states and in all content strands
  - Constructed response items were a particular challenge for students
  - Students who took Algebra II in earlier grades tended to perform better

www.achieve.org/2008Algebra2report

Spring 2009 Algebra II Administrations

Name of State	Spring 2009 Administration (# of students registered)
Arizona	4,156
Arkansas	26,887
Florida (joined 2009)	did not administer
Hawaii	7,301
Indiana	55,908
Kentucky	1,617
Maryland	2,304
Massachusetts	5,000
Minnesota	1,713
New Jersey	11,854
North Carolina	3,138
Ohio	3,035
Pennsylvania	8,645
Rhode Island	435
Washington	did not administer
Total # of students:	**131,993**

- Beginning with spring 2009 the exam is also available online.
- Over 20,000 students were expected to take the exam online this spring.
- Spring 2009 counts are based on number of students registered. Final test counts may vary and will be available in July 2009.

Spring 2009 Algebra I Administration

Name of State	Spring 2009 Administration (# of students registered)
Kentucky	610
Minnesota	9
New Jersey	34,714
Ohio	2,544
Rhode Island	2,905
Total # of students:	**40,782**

- This is the first operational administration of the ADP Algebra I exam.
- A subset of the Assessment Consortium is participating in the spring 2009 administration.
- More states have been involved in developing the Algebra I standards and item review process than have participated this year.
- Beginning with spring 2010 the exam will be available online.
- Counts are based on number of students registered. Final test counts may vary and will be available in July 2009.

The Road Ahead

Near Future

- Standard setting to establish performance levels and cut scores
- Rapid reporting to provide teachers more instructional time and the ability to use scores as final course grade
  - New Jersey submitted Enhanced Assessment Grant application to advance automated scoring methodologies for math constructed response items

Longer Term

- Next steps for policymakers
- Additional research validity studies
  - Content validity
  - International benchmarking
  - NAEP benchmarking
  - Algebra I
  - Longitudinal

How to contact us

Laura Slover: lslover@achieve.org
Stan Heffner: Stan.Heffner@ode.state.oh.us
Gayle Potter: Gayle.Potter@arkansas.gov
Bernie Sandruck: bsandruck@howardcc.edu
Shilpi Niyogi: shilpi.niyogi@pearson.com

ROSA VILLEGAS

7

January 20, 2010
Race to the Top (RttT) Hearing
Washington, DC
Re: General and Technical Assessment

Thank you for allowing me to contribute to today's hearing. My name is Rosa Villegas. I'm the parent of Leo an 11 year old 6th grader in full inclusion at Millikan Middle School who is diagnosed with Down syndrome. He loves to swim and plays on a Soccer Team. I'm also Los Angeles Unified School District's (LAUSD) Special Education Multilingual Advisory Committee Chair and LAUSD's Special Education Community Advisory Committee (CAC) Public Relations chair.

My concerns with the Race to the Top program is that no matter what people say in these hearings, our students with disabilities will not be properly represented or served. The Innovation Fund, Section 14007 caused our state to change its law and lift the cap on granting Charters to be eligible for this funding. Unfortunately, Charter Schools in LAUSD do not take our moderate to severe children. I'm also submitting data one of our CAC parents had requested from LAUSD regarding this. Charter organizations will be applying for these funds based on claims that they have indeed met eligibility requirements of Section 14007(b) (1,2,3 & 4) when in fact they have done so without having a true cross-section of typical neighborhood enrollment if compared to a regular neighborhood school.

Charters do not take students with moderate to severe disabilities so cannot truly claim to:

"(1) have significantly closed the achievement gaps between groups of students described in section 1111(b)(2) of the ESEA (20 U.S.C. 6311(b)(2))" - they do not take a reasonable enough percentage of Students with Disabilities (SWD) to make the claim, and the data I've submitted supports this.

"(2) have exceeded the State's annual measurable objectives consistent with such section 1111(b)(2) for 2 or more consecutive years or have demonstrated success in significantly increasing student academic achievement for all groups of students described in such section through another measure, such as measures described in section 1111(c)(2) of the ESEA" - While they may have "significantly increased student academic achievement" - it is not "for all groups of students", especially the SWD group and data supports this.

"(3) have made significant improvement in other areas, such as graduation rates or increased recruitment and placement of high-quality teachers and school leaders, as demonstrated with meaningful data;" – the meaningful data I am submitting shows that "improvements in other areas, such as graduation rates" is due to the lack of enrolling our SWD population and data supports this.

(4) demonstrate that they have established partnerships with the private sector, which may include philanthropic organizations, and that the private sector will provide matching funds in order to help bring results to scale." – the results will be that we have created schools that do not take students with moderate to severe disabilities and data supports this.

I would prefer that IDEA be fully funded up to the full 40% as originally intended instead of providing grant opportunities for entities who will utilize funding creatively to keep our students with disabilities from participating in their schools. Of course test scores will be higher in Charters – there are no English Language Learners or Students with moderate to severe disabilities to bring down the scores.

You cannot even start to have a discussion on assessment until you have the discussion on equity. All Charters **do not** "best serve" all children. RttT can easily become a mechanism that Charters will use to grow at the expense of our disabled and ELL kids in regular public schools. The Charter School "model for best practice" in Los Angeles is not to enroll and to discriminate against English Language Learners and our Students with Disabilites. Not one Charter School in LAUSD has been able to identify and document best practices that can be shared, and taken to scale based on demonstrated success" with The District because their success is in discriminating against and refusing to enroll English Language Learners and our Students with Disabilites.

6759 Tyrone Avenue, Van Nuys, CA 91405-4724
Phone: 818-988-8594 email: villegasr22@yahoo.com