<table>
<thead>
<tr>
<th>Author</th>
<th>Title</th>
<th>Date Submitted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scott Marion</td>
<td>Establishing a state consortium for assessment: A discussion of factors to consider</td>
<td>1/4/2010</td>
</tr>
<tr>
<td>Kathleen Tannian</td>
<td><em>No Title</em></td>
<td>1/6/2010</td>
</tr>
<tr>
<td>Hugh Burkhardt</td>
<td>Race to the Top Assessment Program Submission from Hugh Burkhardt for the Mathematics Assessment Resource Service</td>
<td>1/7/2010</td>
</tr>
<tr>
<td>Kimberly Hymes</td>
<td>Testimony Submitted to U.S. Department of Education</td>
<td>1/8/2010</td>
</tr>
<tr>
<td>Terry Paul</td>
<td>Thoughts on Testing Arising from the Common Core State Standards Movement</td>
<td>1/8/2010</td>
</tr>
</tbody>
</table>
Establishing a state consortium for assessment:
A discussion of factors to consider

Charles A. DePascale
National Center for the Improvement of Educational Assessment (NCIEA)

December 2009

Unparalleled demands on state assessment programs and conditions associated with the availability of federal monies, particularly the $350 million portion of the $4 billion Race to the Top funds set aside for assessment, have led to a renewed interest in the establishment of state consortia for the development and administration of assessment programs. The motivation to establish state consortia may be driven by a combination of factors such as a) the desire for national, cross-state comparisons on a common assessment produced by a consortium of states, b) the belief that a consortium of states working together is more efficient and cost effective than individual states developing their own assessments, c) the belief that a consortium of states working together will produce higher quality assessments than any individual state working on its own, and d) the belief that a consortium of states working together increases the equity in assessment resources available across states. Similarly, there are a variety of purposes for which an assessment consortium might be established including a) developing common, general summative assessments to be administered across states, b) developing common components of a general, comprehensive assessment system, c) developing specific assessments to measure the knowledge and skills of particular subgroups of students (e.g., ELL or SWD), and d) conducting research and determining best practices in the design and use of various assessment models, item types, and alternative formats to create comprehensive assessment systems.

Regardless of the motivation for establishing the consortium or the specific purpose for which the consortium is established, there are common factors that will impact the operation of any consortium that should be considered as it is being established. In this document, we provide a framework that includes four major areas to consider when establishing a consortium:

1. Role of consortium members
2. Governance of the consortium
3. Management of the consortium
4. Structural organization of the consortium.

Although these areas are interrelated, the issues encountered within each area are significant and distinct enough to warrant discussion within its own section. In part because there are no hard and fast rules regarding the organization of a state consortium, it is critical that any proposal to develop a consortium include consideration of how issues within each of these four areas will be addressed.

It is common to write and speak naively of a state or consortium of states developing an assessment program. Also, it appears that the assessment consortium idea is attractive to
many people because of the assumption that consortia reduce workloads of state assessment personnel. Before beginning the discussion in this document, it is important to establish two axioms on which the arguments presented in this document are based:

Axiom 1: States play only a partial role in the development and administration of state assessment programs.

Under current state and federal requirements, states do not have the capacity, expertise, or desire to fully design, develop, and administer a state assessment program. A variety of advisors, partners, assessment contractors, and vendors play a crucial role in state assessment programs. The level of involvement and the particular areas of involvement of states will vary based on capacity and expertise, but some external support is required in virtually all cases.

Axiom 2: A consortium is an entity distinct from the institutions or organizations that it comprises.

Whenever a group of states, institutions, organizations, or even individuals is convened the management of that group becomes a distinct task requiring the allocation of resources above and beyond those assigned to accomplish the tasks for which the group was convened.

When reading the document, it is also important to realize that a consortium is dynamic and fluid, and that the roles, governance, management, and structural needs of the consortium may change over time as the consortium moves from a focus on design and development to the operational administration of an assessment to maintaining and growing an operational assessment program. It is also important to realize that although this document is concerned primarily with states as members or leaders of a consortium, the leadership of a consortium may fall to other partner organizations or institutions involved in the consortium.

Role of Consortium Members

State leaders must be clear when joining a consortium about how intensive a role they expect to play and whether they have the capacity and/or resources to meet this level of involvement.

The function that state personnel serve within a consortium varies greatly across consortia. Factors such as the size of the consortium, the expertise and capacity of its members, and the perceived importance and impact of the purpose and products of the consortium interact to determine the appropriate organizational model for the consortium and the optimum level of involvement of the individual states. The level of state involvement in a viable consortium can range from total involvement in all operational decisions to serving as an advisory or policymaking board to minimal direct involvement in operations decisions.
In general, there is likely to be an inverse relationship between the size of the consortium and the level of control, influence, and involvement of individual states. If the goal of the consortium is to produce a common product then individual states will be more likely to be asked to compromise and give up individual practices and preferences as the size of the group increases. Also, it simply quickly becomes too unwieldy and inefficient to seek state input and approval on all operational decisions, and perhaps all policy decisions, as the size of the consortium increases.

Conversely, there is likely to be a direct relationship between a state’s level of involvement in the consortium and its capacity, interests, and level of expertise. States may expect to exert more operational influence and control in a consortium convened to develop multiple-choice, summative assessments (an area in which they have extensive experience) than in a consortium whose purpose is to develop the infrastructure and hardware for a technologically-based assessment system (an area in which they have limited expertise). Similarly, states are more likely to desire to maintain policy control on projects dealing with areas in which they have made a significant investment or have a passionate interest. However, states may be more willing to cede control to the consortium if they have limited staff and resources to devote to the project.

The perceived importance or impact of the project will also impact a state’s level of involvement in the consortium and their willingness to compromise on issues brought before the membership. A project that impacts a very small percentage of students or has minimal consequences attached to its results may be one in which states are less inclined to exert control or influence. Conversely, a project which has significant legal, financial, or political implications is one in which an individual state may expect to be heavily involved in all major policy decisions and/or maintain a high level of control and influence over the decision-making process.

**Governance of the Consortium**

*There are a range of possible governance structures for a consortium ranging from full operational partnership to a “users” group model. The governance structure will affect all other aspects of the consortium, so potential members will have to weigh various considerations when deciding on the governance structure of the proposed consortium.*

When a consortium of states is created, it will be necessary to establish a governance structure for its operation. The purpose of the governance structure is to define the relationship and roles of the individual members of the consortium as well as to establish the procedures and protocol by which the consortium will operate. The mission of the consortium, its purposes and goals, and the products it expects to produce or services it expects to deliver should also be clearly delineated and understood to ensure that they are supported by the established governance structure.

One of the first decisions to make regarding the governance of the consortium is the role of the individual members in its governance. Three common governance models for consortium are
members are operational partners in the consortium,
members serve as a board of directors for the consortium, and
members serve on an advisory committee for the consortium.

A fourth model is the “user group” model in which the members are consumers of a common product produced by an external contractor/vendor and exert only indirect influence through organized feedback and market choices. In the user group model, the development of the assessment is likely to be non-collaborative, but the consortium of states may be directly involved in the development of ancillary products and services related to the interpretation and use of assessment results.

If it determined that member states have a direct role in the governance of the consortium as partners or as a board of directors then additional decisions regarding the distribution of power among member states must be made. In a consortium where states are voting members, a basic question to be resolved is whether the states are equal partners – that is, one state/one vote. Although equal representation and voting rights may be the preferred approach in many cases, there are cases in which other approaches might also be considered such as the following:

- There is a wide discrepancy in the populations of the member states, and consequently in their level of use of the products and services.
- There is variation in the level of commitment or in the contribution that member states are making to the consortium.
- The stakes associated with the product or services of the consortium are much higher in some member states than others.

Like design of the United States government, it may be the case that the consortium does not establish a single voting policy applicable to all situations. There may be some topics in which each state has a single vote, others in which votes are proportional, and some which require unanimous agreement among the states.

Although a clear consensus among the states is preferred for most decisions, the situations for which unanimous agreement is required – or in which an individual state has veto power – should be limited in number and to issues in which unanimity is critical. Veto power, in general, is counterproductive in a project designed to produce a quality product in a timely manner. The result of allowing a single state to block decisions is most likely either a slowing down of the process or a narrowing of the product to only those elements on which all member states agree.

Management of the Consortium

Managing the consortium is a task distinct from managing the assessment program that, like governance, can range from a very intensive commitment (e.g., multiple FTE) to a somewhat reduced role (e.g., ½ FTE). This will require that the consortium budget significant resources for a high-quality person or organization to fill this role.

Management of the assessment program that the consortium was constituted to develop is an obvious responsibility of the consortium. Above and beyond the management of the tests and related assessment services, however, is the perhaps less obvious task of managing the consortium itself. Like any organization, the consortium must be properly organized and managed to effectively and efficiently carry out its tasks and meet its

C. DePascale, NCIEA, 12/23/2009
goals. The amount of resources needed to manage the consortium depends, of course, on all of the factors previously discussed including the size and structure of the consortium, the level and type of involvement of its members, as well as the number and type of products designed and implemented by the consortium. Some resources, however, will have to be devoted to the management of every consortium of states.

Management of the consortium refers to those tasks and services not directly related to the administration of the assessment program that must be performed to ensure the smooth operation of the consortium. At the most basic level, these tasks might involve support tasks such as coordinating communications among consortium members, scheduling meetings, and making travel arrangements. In consortia with active involvement of the members, consortium management is also likely to include higher level project management responsibilities to coordinate decision-making among the members. Management of the consortium might also involve serving as a liaison between the consortium members and their various contractors and vendors or actually serving as the contracting agent for the consortium. As the focus of the consortium expands beyond the development and delivery of an assessment, consortium management might also be responsible for tasks or contracts related to support materials, technical advisory roles, professional development services, and research and development.

In many cases the individual or group responsible for the management of the consortium will be a third party external to both the state members of the consortium and to the assessment contractor, but clearly serving as the states’ advocate. Management external to the states can be advantageous to the state members because no single state is asked to assume the burden of consortium management and no single state is placed in a perceived position of power within the consortium. Establishing management control external to the contractor is a logical option in cases where there are multiple assessment contractors and cases in which policy conflicts across states will need to be negotiated and resolved – both of which are likely occurrences in a consortium assembled to develop and administer a comprehensive state assessment program.

The consortium model in which the consortium is managed by the assessment contractor is most likely to be found is the user group model. As described previously, the user group model is a largely non-collaborative model in which multiple states are purchasing the same product from a single agent. In this model, the common assessment, and perhaps the consortium itself, is inextricably linked to a specific contractor. In a common scenario for this model, the agent would be a test publisher selling an off-the-shelf test. In another scenario the agent could be a state whose custom-made assessment program other states have agreed to administer. However, it is likely that a state will not have either the capacity or the desire to actively manage the sale of assessment products and services to other states. Of course, internal management of the consortium by the assessment contractor is not limited to consortia which are non-collaborative or in which the contractor has all of the control. States may serve as board members or advisory panels exerting influence on the design and direction of an assessment program managed by the assessment contractor. Also, as previously noted, none of the consortium classifications presented here are totally distinct or permanent. A consortium may begin
as a partnership with heavy state involvement during the design and development phase of an assessment program and begin to function more like a user group as the assessment program becomes operational

**Structural and Legal Organization of the Consortium**

*States’ legal and fiscal structures and rules can derail the best laid assessment plans.* Therefore, it is crucial that state assessment leaders and consortia organizers clearly understand the structural, legal, and fiscal constraints of the various state members and design the consortium structure to account for these issues.

Distinct from the governance and management issues are a variety of topics related to the structural organization of the consortium. One set of issues relates to the structure of the consortium and the formal interactions of its members:

- the legal organization of the consortium,
- rights, responsibilities, and obligations of consortium members, and
- protocol for allowing additional states to join the consortium or for states to leave the consortium.

A second set of issues relates to the funding of the consortium’s activities and the manner in which states pay for and use assessment materials produced by and for the consortium:

- funding options,
- procedures for entering into contracts with third parties, and
- procedures for the expenditure and collection of funds.

Finally, related to both the rights and responsibilities of consortium members and the financial issues listed above are issues related to the ownership and use of materials purchased or produced by the consortium,

Each of these topics is too complex to fully discuss within this document, but several key issues are highlighted in the following paragraphs. As with the areas of management, governance, and roles of consortium members discussed previously in this document, the structural needs of the consortium are likely to change over time. It is critical that the consortium anticipates those changes and builds in the flexibility to adapt to the changing circumstances.

Dependent upon the design of the consortium, its legal organization can range from a relatively informal agreement among states to the formal establishment of a separate corporation (for-profit or not-for-profit). As the size of the consortium grows or the risks and responsibilities assumed by the members grows the more likely it is that the issues the consortium faces will become more complicated and the consortium will require legal advice and formal agreements among states. As a starting point, it will be necessary to establish procedures to protect individual consortium members, partners, and contractors from and minimize the potential impact of a) decisions by individual states to no longer participate in the consortium or b) the inability of individual states to meet their obligations to the consortium.

Fairly early in the process, it will become necessary for the consortium to generate revenue to support its efforts and to expend money to pay for services.
multiple state governments, legal requirements for both generating and expending resources may be complicated and vary significantly across states. It may be the case that initial work of the consortium is funded, at least in part, through federal grants, grants from private foundations or similar organizations, support from partner organizations, or contributions from member states. The level of support needed for the initial work of the consortium will vary tremendously based on the extent to which the consortium is directly involved in the design, development, and ownership of a custom assessment as opposed to purchasing an assessment developed and owned by a commercial test publisher.

At some point, the consortium or its individual state members will also have to enter into contractual agreements with assessment contractors or related vendors for activities related to the administration of an assessment program. In most cases it is likely that this will involve individual state contracts with either the consortium (as a separate corporation) or directly with a general assessment contractor. It is less likely that the consortium will adopt a model in which state partners will be required to enter into individual contracts with a variety of vendors providing services related to the administration of the assessment (e.g., printing/production, shipping, scoring, reporting). It may be the case, however, that individual states, or groups of states, may require specialized optional services due to factors related to their population, location, or legal requirements that are distinct from those required by the other states in the consortium.

When the consortium reaches the point that it is purchasing or producing materials (e.g., test items, test booklets, ancillary test materials, or professional development tools), issues related to the ownership and appropriate use of those materials will need to be resolved. The consortium may try to avoid ownership issues by establishing procedures in which a) nothing is owned by the consortium (i.e., all materials are owned by an assessment contractor) or b) all materials are jointly owned and available for use by all consortium members. Neither of these approaches, however, will totally eliminate the need for the consortium to confront issues related to how, when, and by whom the materials can and should be used within the consortium, by states or non-commercial groups outside of the consortium, or by commercial groups not directly affiliated with the consortium. The consortium will also have to establish policies to handle damages caused by intentional or unintentional misuse of assessment materials.

**Examples**

In this document, we provided a framework for the establishment of state assessment consortia that includes consideration of four critical factors:

- Role of consortium members
- Management of the consortium
- Governance of the consortium
- Structural organization of the consortium.

To conclude this overview, we will use that framework to provide brief descriptions and comparisons of three existing state assessment consortia that differ on one or more of the dimensions discussed in this paper:

- New England Common Assessment Program (NECAP),
• World-Class Instructional Design and Assessment Consortium (WIDA), and
• Achieve ADP Assessment Consortium (Achieve)

These three programs differ in many of the factors described above as well as in the type of assessments that are offered, the use of those assessments, and the general purpose of the consortium. They provide models of consortia in which the assessment is jointly owned (NECAP, WIDA) v. privately owned (Achieve); consortia in which assessment contractors and other partners play different but critical roles; and three significantly different models of consortium governance, management, and structure.

NECAP is essentially a partnership of four states in the development of a custom state assessment program designed to meet NCLB Title I requirements with tests in English language arts and mathematics at grades 3 through 8 and 11, science at grades 4, 8, and 11, and writing at grades 5, 8, and 11. New Hampshire, Rhode Island, and Vermont administer all NECAP tests. Maine administers the English language arts, mathematics, and writing NECAP tests at grades 3 through 8.

The WIDA Consortium is a non-profit cooperative of 20 states and the District of Columbia focused on standards and assessments for English language learners. Consortium members administer the Access for ELLs tests developed by the Consortium as their state English proficiency assessment required by Title III of NCLB.

The Achieve ADP Assessment Consortium is a consortium of 15 states who are members of the larger 35-state American Diploma Project (ADP) network dedicated to improving the level of college and career readiness of high school graduates. States use the Algebra II and Algebra I end-of-course tests administered through the consortium in a wide variety of ways. Some states administer the tests to all students completing either Algebra I or Algebra II. Other states administer the tests available to districts on a voluntary basis. Some states administer only the Algebra I or the Algebra II test. In the future, some states may use one of the end-of-course exams as a state high school assessment to meet NCLB Title I requirements or as a component in student graduation decisions. One goal of the consortium is that the Algebra II test will be a valid indicator of readiness for an entry-level college mathematics course and will be used to support placement decisions at postsecondary institutions.

There is some overlap in membership across the three consortia. The four New England states are all also members of the WIDA Consortium. Seven of the WIDA Consortium states are also members of the ADP Assessment Consortium. Rhode Island is a member of all three consortia.

The table on the following page provides a very brief summary of the organization and operation of the three consortia based on the framework provided in this document. The table highlights the significant differences among the consortia in areas such as ownership of the assessment, management of the consortium, and participation of the membership.
## Overview of the Organization and Operation of Three State Assessment Consortia: NECAP, WIDA, Achieve Algebra

<table>
<thead>
<tr>
<th>Role of Members</th>
<th>Governance</th>
<th>Management</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>New England Common Assessment Program</strong></td>
<td>- Four member states actively participate in the development of a common assessment with their assessment contractor – Measured Progress</td>
<td>- Member states are equal partners in all decisions regarding the assessment program.</td>
<td>- Assessment is owned jointly by the individual states.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- States issue identical RFP and enter into individual contracts with the assessment contractor.</td>
</tr>
<tr>
<td></td>
<td>- Original member states (Wisconsin, Delaware, and Arkansas) and eight second-wave states (including District of Columbia) worked with partners at the Center for Applied Linguistics, University of Wisconsin system, and University of Illinois to develop standards and an assessment framework.</td>
<td>- The WIDA Consortium is a fully-staffed non-profit cooperative.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Current member states serve on the WIDA Board and play a variety of roles in various WIDA projects.</td>
<td>- In addition to administration and operations, WIDA staff also includes groups focused on assessment, professional development, and ongoing research.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- In addition to the assessment, the consortium is involved in work on standards, offers professional development services to members, and conducts research and validation projects.</td>
<td></td>
</tr>
<tr>
<td><strong>WIDA Consortium</strong></td>
<td>- Original nine member states actively participated in the development of content standards and design of the tests and testing program with Achieve.</td>
<td>- Achieve manages the consortium with significant input from the member states and serves as primary contact with the assessment contractor – 2+ FTE</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Current fifteen member states serve on item review committees; approve proposed contract changes and schedules with their assessment contractor - Pearson.</td>
<td></td>
<td>- Assessment is owned by Pearson</td>
</tr>
<tr>
<td></td>
<td>- States are asked to approve changes to test design and schedules. Consensus agreement is required for significant changes to contract with Pearson.</td>
<td>- Pearson assumed initial development costs prior to sale of operational tests.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- A single procurement RFP was issued by one state. Other member states may enter into agreements with Pearson to purchase tests through that contract.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Entry of new states is facilitated by Achieve through the ADP Network.</td>
<td></td>
</tr>
</tbody>
</table>

*C. DePascale, NCIEA, 12/23/2009*
My name is Kathleen Tannian. I am a member of the Pembroke Teachers Union in Pembroke, Massachusetts. As a teacher teaching in an area in which there is a documented shortage in Massachusetts, Modern World Languages, specifically Spanish, I would like to note that one factor that is missing in our education system is the feeling that the high school experience is a high-stakes challenge for students, a factor which is definitely ever-present for students in other countries. A high-stakes assessment after high school to determine which majors a person may choose in college is what other countries use. In the United States, senior year is commonly thought to be a time to relax. Often extracurricular activities and activities that do not have anything to do with academic benchmarks are most valued today by the some of the leaders in public education.

My colleagues and I are somewhat dismayed when we hear students focus more on who will yell the loudest at the pep rally rather than who will score the highest on their upcoming midterms. To make teachers responsible for an entire system in which they are treated at times by leadership as enemies of the all-important pursuit of extracurricular and school day fun, would be to discourage any thinking person of integrity form staying in a broken system. We should beware punishing teachers for poor performance by students. The business model does not work in schools. Intrinsic reward for students and for teachers is the best motivation. The promise of a better life is what drives the greatest high schools in the world. It is no coincidence that the allies in that pursuit of that better life, the teachers, are valued in those countries.

Sincerely, Kathleen Tannian
The purpose of this email is to confirm your submission of written input regarding the Race to the Top Assessment Program.

Thanks for your response. This is the follow up.

To ensure that your input is fully considered, if you have not already done so we ask that you identify clearly in the body of a follow-up email the specific question, purpose, and characteristic that each of your suggestions addresses and to arrange your submission in the order of the questions listed later in this notice.

Our paper (attached again for easy linkage, with the task exemplars) mainly relates to **General Assessment Question 1**.

**GAQ1**

Our paper sets out design principles that underlie previous implementations of high-stakes assessment around the world that meet your goals, along with some practical moves. It is thus intended to help in your review of the proposals you will receive, from testing agencies and others. Given the track record of US assessment, we think it unlikely that any of these proposals will deliver what you ask for. (Any ambitious program needs good engineering, i.e. a design and development process combining all the relevant knowledge and skill, with imagination with rigor)

We have not made a detailed proposal because, as our paper shows, there are various policy decisions that are needed to guide even strategic design -- on cost, time, linkage to teaching and professional development. If we were to make such a proposal, the high-stakes element would be for something like:

- annual tests, each taking 2-3 hours
  - ~1 hour of short items focused on the year's new content
  - ~2 hours of 5-20 minute substantial tasks, assessing mathematical thinking balanced so that "teaching to the test" leads to a rich and balanced curriculum.
- Initially norm-referenced with year to year and grade to grade linkage from piloting
- Scored by teachers (on screen?) with scoring training as powerful professional development
- Scoring samples monitored, with corrective scaling
- Tests released after a specific testing date (security is somewhat illusory)

with, ultimately, all tasks drawn from a public bank large enough for learning them all in advance to be unrealistic (and, where done, good education)

Teachers would be provided with a similar range of tasks, "practice tests" and scoring guidance for use in class throughout the school year.
There follow some comments on your other questions, as labelled:

**GAQ2:**
**Type:** criterion-referencing inevitably drives down standards through fragmentation - the easiest way to meet a criterion, which in fairness must be offered, is in a short item (see Section 2B of http://www.educationaldesigner.org/ed/volume1/issue3/article9/index.htm); adaptive testing tends to suffer from the same problem. Worthwhile tasks have a higher "total cognitive load" than their separate parts

**Frequency:** Fit for purpose.
Formatively, every day: doing tasks unaided is "the game" for which teaching is the training; the reasoning shown should guide instruction.
For accountability: once a year is plenty -- real growth is slow (though you can test new content)

**Format:** FACT: high-stakes tests determine the implemented curriculum in most classrooms - so, if you want mathematical thinking, the task-set must be balanced across your goals.
WITHOUT THIS, PROGRESS WILL NOT HAPPEN

**GAQ3 and 4**
Separate responsibility according to skill, along the lines we suggest. For LEAs
Task design: commission tasks from several agencies with track records for various task types
Test design: assign test assembly and balancing to math ed specialists in the LEA, with some outside advice
Test delivery: Assessment specialists in the LEA, with an outside agency; math ed specialists in the LEA handle scoring training of teachers as a (powerful) part of the professional development program. (Human scoring should be the main cost)

**GAQ4**
There are various well-established ways of doing this. (eg SVMI/MAC have been doing this for 10 years)

Scanning student papers and marking on-screen saves handling costs and facilitates monitoring, but there is an overhead

**GAQ5**
Competency-based testing, like criterion-referencing, forces fragmentation and 'false positives' -- skills students can't use, and see as irrelevant. This is not assessing mathematical thinking.

High expectations are stimulated by a ladder of increasingly rich, more challenging problems.

**GAQ6**
This is problematic for any test. (see the section on Accuracy in our paper). School value-added scores over a few years can be reliable evidence. More detailed analysis is unreliable, because of all the other factors involved. Going for it drives the system away from broad and balanced assessment.
Pilot substantial tasks across several grades, using the score distributions for linkage

Having the balance of task-types determined by an appropriately balanced panel (see our paper) Statistics cannot do it for you.

100%. See GAQ1: "ultimately, all tasks drawn [should be] from a public bank large enough for learning them all in advance to be unrealistic (and, if done, good education)"

Also, if you have not already done so please include in the body of a follow-up email a description of your involvement, if any, in statewide assessment practices.

We covered this in the introduction to our paper.

The Secretary is committed to ensuring absolute transparency and openness during the input gathering process. To that end, all input will be posted on the Department's website.

Fine

Thank you for your interest in this program.

We in MARS see this as a critically important enterprise in moving US education in Mathematics forward to higher real standards

This message has been checked for viruses but the contents of an attachment may still contain software viruses which could damage your computer system: you are advised to perform your own checks. Email communications with the University of Nottingham may be monitored as permitted by UK legislation.
MARS is an international collaboration with 25 years experience, in collaboration with major test providers\textsuperscript{2} in the UK and the US, in the design of assessment that matches the aims set out in your admirable document of October 20 (called RTAP below). In it you recognize, indeed emphasize, that the approach you describe is far from current practice in high-stakes testing in the US. However, working models in the UK, Australia and elsewhere in the past show that your criteria can work well and have wide public acceptability as high-stakes assessment. This leads us to believe that the summary of that experience in this paper may be of value in your current enterprise.

We begin with practical steps that make the aims set out in RTAP more achievable, lifting test design and delivery out of the groove of current practice – as one must. In section 2 we outline the strategic design principles that, if sustained, can make this possibility a reality.

Finally, long experience has shown that exemplification is essential in making meaning clear – descriptions are easily misinterpreted\textsuperscript{3}. So we complement this brief analytic account with some illustrative tasks\textsuperscript{4} that show the variety and balance that we believe need to be included, if tests are to recognize, and thus encourage, the kinds of \textit{thinking with mathematics} that are needed in the modern world.

1. Moves to improve assessment

The following departures from current US practice have all worked well in large-scale use. They are essentially independent, each contributing value to the assessment system. The quality of the detailed engineering of their realization is important – only sound principles, with imaginative design with systematic development and testing, produce robust high-quality solutions.

\textbf{Separate test design from test delivery}

These two are very different activities, demanding quite different capabilities. When test providers control the design of the tests, various pressures lead them to make life easy for themselves and cheap for their school system customers. This introduces constraints on the design (eg machine-scored multiple-choice items only, or mainly) that preclude assessing important kinds of performance (substantial chains of reasoning, or even procedural skills!). The focus moves to meeting the customers’ primary need for low-cost reports, without worrying much about what is \textit{assessed} – this distorts the implemented curriculum in most classrooms.

More balanced tests still offer business opportunities for test providers – indeed because such tests are more expensive, the potential is greater. The challenge of

\textsuperscript{1} Contact Hugh.Burkhardt@nottingham.ac.uk

\textsuperscript{2} In the UK they include the Joint Matriculation Board (now AQA) and the Midlands Examining Group (now OCR); in the US, CTB McGraw Hill and various state and school district systems, including the Silicon Valley Mathematics Initiative;

\textsuperscript{3} \textit{problem solving} and \textit{validity} are but two terms that are regularly used with widely different meanings.

\textsuperscript{4} Tasks that demand substantial chains of reasoning are many items, in the original statistical sense of score point.
getting school systems to choose high-quality tests is addressed below; a prerequisite is that such tests are widely available, and that their added value in encouraging higher standards is recognized.

Separate task design from test design

The design and development of assessment tasks that test whether our children can think with mathematics is a highly skilled and specialized activity\(^5\) – one in which a few US groups have good track records. In contrast, the assembly of tests from a pool of high-quality tasks essentially involves societal decisions on the balance of different kinds of task – this should reflect a national view of the balance of performance types that is important at each grade or stage. (In traditional “standards”, which are essentially a list of topics for each grade, this balance is not specified – an unintended but inevitable consequence of this criterion-based approach is the dominance in tests of short items – a travesty of performance in mathematics.)

The balancing process of test assembly needs to carry authority and credibility\(^6\). This could be provided, for example, by a knowledgeable panel from (say) the National Academy’s scientists, engineers\(^7\) and pure mathematicians, the mathematics education professions, business, and state or national government. Such a group would instruct a design team to assemble a number of draft tests with specific proportions of the various task types, drawn from the pool of well-engineered tasks. The panel would then critique the drafts, choosing a suitably revised version. Such a process would command respect, avoiding the “race to the bottom” that competition to deliver routine, predictable and cheap tests produces. Complementing this, the design and development of tasks for the pool\(^8\) would be commissioned from a number of design teams in the usual way, encouraging constructive competition for higher quality. Together they would reflect the full range and variety of performance types. The selection of the teams could well involve the panel suggested in the previous paragraph.

Involve teachers

There are enormous benefits in involving teachers in the high-stakes assessment process while, of course, monitoring their work to ensure reliability. The advantages include:

- **linking assessment and teaching** This link is inevitable, because high-stakes tests largely determine the implemented curriculum in most classrooms; the essence of RTAP, reflected in this paper, is to move to tests worth teaching

---

\(^5\) MARS has found that it costs about $3,000 to develop a substantial task that gives every student who has been well taught the opportunity to show what they know and can do in thinking with mathematics. This contrasts with ~$1 per item paid to “item writers” by testing companies, who then spend heavily on determining the statistical properties of the test.

\(^6\) Recent history has shown that standards are surprisingly “flexible”. NCLB required state tests to be aligned with the state’s standards, many of which were of high-quality. Perhaps because of cost pressures, standards were often interpreted in a way that led to minimal changes in most state tests.

\(^7\) There are many more of these users of mathematics than there are pure mathematicians, who often have a more inward-looking view of the subject than most people require.

\(^8\) The aim should be a task pool that is large enough so that it can be public – ie teaching and learning how to solve all the problems would not be a feasible approach. The sample selected for each test could, of course, remain secure as long as necessary – typically until a given date.
Further, the same range of rich tasks will become an integral part of the pattern of classroom activities – “the game” for which teaching is “the training”.

- **professional development** Sessions focused on a rich assessment task, with scoring guidance and samples of student work at various levels, are among the most powerful and motivating kinds of professional development, generating focused discussions on mathematics, and the teaching needed to develop students’ performance in it.

- **containing costs** Apart from the direct benefits above, teacher scoring with rigorous monitoring of random samples can cut costs substantially. There are various well-established ways of doing this, some contributing to the further professional development of the teachers,

It remains to be seen if the ingrained distrust of teacher (or even human) involvement in scoring can be overcome.

**Explore marketing opportunities**

Meeting the challenge of gaining wide public acceptability will need leadership, incentives, and creative public relations.

If the cost of high-quality assessment, roughly $15-30 per student test, precludes its large-scale take up in school systems, other possibilities for funding its introduction should be explored.

For example⁹, the test could be marketed to parents (as with current AP tests). Many parents spend much larger sums on tutoring and other means to help their children achieve higher-standards (or just higher scores). If successful, this “trojan horse” approach could well lead to schools embracing the test – then, on equity grounds, to public support for those of limited means. Relying on current narrow tests would then be a sign of limited educational ambition by a school or a district.

**2. Design principles for high-quality assessment**

We have outlined above some practical moves that would contribute to a substantial rise in quality of assessment in mathematics. We now sketch a few key principles for assessment design. Much of what follows is implicit in RTAP; nonetheless we feel it is worth emphasizing.

**Observe the principles of good strategic design**¹⁰ It is the essence of strategic design of an innovation to take very seriously the properties of the system you are trying to improve – as it is, not as it is intended to be. This includes the likely responses of all the key groups and how they can be moved in the directions required. Otherwise unintended consequences will follow.

This does not mean that profound change is impossible, but it usually requires well-aligned pressure and/or encouragement and effective support for all concerned. The design and development process must ensure that the people who must change can change with the support you are proposing to give – through materials and network support for their professional development.

---

⁹ We offer this suggestion tentatively – marketing is not among our areas of expertise.

Tests should fit their purpose – one size does not fit all
It should be unnecessary to restate that the purposes of an assessment should dictate its design – were it not so often neglected. A diagnostic test that is focused on a single topic, a formative assessment lesson that guides teacher and students to develop their reasoning on a problem, a summative test that gives an overview of each student’s current level of performance for accountability purposes – they all have their different design priorities11. (Below we mainly consider the last, reflecting its dominance.)

WYTIWYG and the “measurement only” fallacy
Most teachers know that “what you test is what you get” – that the teaching and learning in most classrooms will be dominated by the types of task in high-stakes examination. Thus tests are not only measurement, they are a major driver of what is taught and learned– the implemented curriculum will not be better than the tested curriculum. Thus if you want high standards you need tests that sample, in a balanced way, all the types of performance that mathematical competence implies. Then teaching to the test can raise standards.

Assess mathematics, not individual “standards”
In complex performances, the whole is more than the sum of the parts. Basketball is more than dribbling, shooting and the other elements. Playing the piano is more than scales and arpeggios. Writing is more than spelling, grammar and syntax. So it is with mathematics. “Standards” or other forms of domain description are weak models of performance, which is more vividly described by a rich collection of task exemplars. The two complement each other, the tasks show the performance goals while the domain description illuminates what the tasks assess and guides the balancing of tests.

Accuracy
Performance in mathematics, even in K-12, is a complex business. Assessing a student’s level of performance is not a precise measurement. Tests that claim that it is make it so by only by narrowing the tests to assess fragments of performance. The impression of accuracy is enhanced by averaging – over many items, many students, many classes. The real significance of such averages is unclear – while the statistical error is reduced, the systematic errors in what is assessed remain unchanged. Statistics make a rough measurement into an accurate one – of something else.

11 In considering the design of assessment, it is worth remembering the overall purpose: helping children, not tattooing them.
“Target Performance” Tasks
Some exemplar tasks from the
Mathematics Assessment Resource Service
(MARS)

We distinguish:

School math tasks: The kinds of task that assess the students performance in concepts, skills and problem solving, with the structured scaffolding that helps students show their level of performance, however weak.

Target performances: The kinds of task that arise in everyday life and work, at a level that is accessible to students in appropriate curriculum programs (see other pdf)

These examples are for “college and career ready” students; similar principles of variety and balance apply at every grade.
You have the job of organising a table tennis league.

- 7 players will take part
- All matches are singles.
- Every player has to play each of the other players once.
- There are four tables at the club.
- Games will take up to half an hour.
- The first match will start at 1.00pm.

Plan how to organise the league, so that the tournament will take the shortest possible time. Put all the information on a poster so that the players can easily understand what to do.
SECURITY CAMERA

A shop owner wants to prevent shoplifting.
He decides to install a security camera on the ceiling of his shop.
The camera can turn right round through 360°.
The shop owner places the camera at point P, in the corner of the shop.

Plan view of the shop

1. The plan shows ten people are standing in the shop.
   These are labelled A, B, C, D, E, F, G, H, J, K.
   Which people cannot be seen by the camera at P?

2. The shopkeeper says that "15% of the shop is hidden from the camera"
   Show clearly that he is right.

3. (a) Show the best place for the camera, so that it can see as much of the shop as possible.
   (b) Explain how you know that this is the best place
MAGIC CROSSES

There are two rules for making Magic Crosses:
• They must contain all the numbers from 1 to 9
• The vertical total must equal the horizontal total.
This is called the Magic Total.

In the Magic Cross shown here, for example,
the Magic Total is 27 because:
the vertical total is 5+4+9+3+6 = 27;
the horizontal total is 8+1+9+2+7 = 27.

1. Which Magic Totals are possible?
Which are impossible?
Prove that you are right.

2. Make up a new problem like this and solve it.

For example: Arrange all the numbers from 1 to 9 in the boxes so that
the total along each line is equal. Call this the Magic Total.
What different values can this Magic Total take?
A grocer wants to sell strawberries in small boxes. He wants to make the boxes from square card 30 cm long and 30 cm wide as shown below:

The shaded areas are cut away and the rest is folded along the dashed lines. The sides are folded up and stuck together using flaps. The lid has two flaps which are not glued.

1. Calculate the volume of this box.

2. Suppose he starts with the same square of card, but changes the 7 cm to a different measurement. What is the largest volume he can make the box?
PROOFS OF THE PYTHAGOREAN THEOREM?

Here are three attempts to prove the Pythagorean theorem. Look carefully at each attempt. Which is the best 'proof'? Explain your reasoning as fully as possible.

Attempt 1:
Suppose a right-angled triangle has sides of length $a$, $b$ and $c$.
Draw squares on the three sides as shown.
Divide these squares into smaller squares.
You can see that the number of squares on the two shorter sides add up to make the number of squares on the longest side.
So: $a^2 + b^2 = c^2$

Attempt 2
Suppose that you start with four right angled triangles with sides of length $a$, $b$ and $c$ and a square tray with sides of length $a+b$.

You can arrange the triangles into the tray in two different ways as shown here.
In the first way, you leave two square holes. These have a combined area of $a^2 + b^2$.
In the second way you leave one large square hole. This has an area of $c^2$.
Since these areas are equal $a^2 + b^2 = c^2$. 

Copyright © 2009 by Mathematics Assessment Resource Service. All rights reserved.
**Attempt 3:**
The proof of the Pythagorean theorem is clear from this diagram.
The squares on the two shorter sides of the black triangle are each made from two congruent triangles. These fit together to make the square on the longest side—the hypotenuse.

The best proof is attempt number ........
This is because ........

My criticisms of the others are........
FEARLESS FRAMES

Fearless Frames PLC make metal frames for containers.

1. A customer asks Fearless Frames to make a large container which is a rectangular cuboid with a square cross section. The company has only 60 metres of suitable metal tubing in stock. Find the dimensions of the container which holds the maximum volume the company can make using 60 metres of tubing.

2. The customer changes his mind! He asks for a container which is a prism with a cross section which is an equilateral triangle. Investigate the maximum volume of the container that can be made using 60 metres of tubing for the frame.

3. What advice do you think Fearless Frames should offer to this customer?
Max has received this email. It describes a scheme for making money.

From: A Crook  
Date: Thursday 15th January 2009  
To: B Careful  
Subject: Get rich quick!

Dear friend,

Do you want to get rich quick? Just follow the instructions carefully below and you may never need to work again:

1. At the bottom of this email there are 8 names and addresses. Send $5 to the name at the top of this list.  
2. Delete that name and add your own name and address at the bottom of the list.  
3. Send this email to 5 new friends.

1. If that process goes as planned, how much money would be sent to Max?  
2. What could possibly go wrong? Explain your answer clearly.  
3. Why do they make Ponzi schemes like this illegal?
“School Math” Tasks
Some exemplar tasks from the Mathematics Assessment Resource Service (MARS)

We distinguish:

School math tasks: The kinds of task that assess the students performance in concepts, skills and problem solving, with the structured scaffolding that helps students show their level of performance, however weak.

Target performances: The kinds of task that arise in everyday life and work, at a level that is accessible to students in appropriate curriculum programs (see other pdf

These examples are for “college and career ready” students; similar principles of variety and balance apply at every grade.
Short tasks

PAINT

A painter pays $26.85 for a can of paint. The paint costs $8.95 per pint.

How many pints does he get?

MAPS

An aircraft coming in to land lets down its wheels at a distance of 12 kilometers from the runway,

What distance is this on the navigator’s map, which has a scale of 0.35 cm to 1 kilometer?
TRUE OR FALSE

One of the numbers below has the same value as $3.5 \times 10^{-3}$.

Write true under the correct number.

$35 \times 10^{-4}$  $3.5 \times 10^{3}$  $0.00035$  $3500$

_______  ________  ________  _______

COFFEE

Hugh uses 13 grams of coffee powder to each liter of water.

He has a cup which holds 0.3 liters of water.

How much coffee powder should he use?

EGGS

Using only a five-minute and a three-minute egg timer, how could you boil an egg for four minutes? Explain.
<table>
<thead>
<tr>
<th>Distance</th>
<th>Time</th>
<th>Distance</th>
<th>Time</th>
<th>Distance</th>
<th>Time</th>
<th>Distance</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td>B</td>
<td></td>
<td>C</td>
<td></td>
<td>D</td>
<td></td>
</tr>
</tbody>
</table>

X
Longer Tasks

Patchwork

This problem gives you the chance to:

• identify and extend a problem
• construct a rule or formula

A sheet of square dot paper is provided for use with this item.

Kate makes patchwork cushions.

She uses right triangles and squares.

She uses triangles along the edges of each cushion. The rest is made from squares. The backs of the cushions are made of plain material, not patchwork.

Here are the first five sizes of patchwork cushions.
Kate makes cushions in many other different sizes.

She begins to figure out how many triangles and squares she needs for each size.

For size 1, she needs 4 triangles and 0 squares.

For size 2, she needs 8 triangles and 4 squares.
1. Complete this table to show how many triangles and squares she needs for each of these five sizes?

<table>
<thead>
<tr>
<th>Size $(n)$</th>
<th>Number of triangles $(t)$</th>
<th>Number of squares $(s)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. Find a rule, or a formula, that will help Kate figure out the number of triangles that she needs for cushions of different sizes. Explain how you figured it out.

3. Use the number patterns in the table to find a rule, or a formula, that will help Kate figure out the number of squares she needs for cushions of different sizes. Explain why your rule works.

4. Kate has a cushion made with 180 squares. How many triangles are in this cushion? Show how you found the number of triangles.
Funsize Cans

This problem gives you the chance to:
• Design cylindrical cans to hold a given amount

The volume of a cylinder is
\[ V = \pi r^2 h \]

The surface area of a cylinder is
\[ S = 2\pi r^2 + 2\pi rh \]

The Fresha Drink Company is marketing a new soft drink. The drink will be sold in a `Fun Size' cylindrical can which holds 200 cm³. Here are two suggestions for the radius of the cylindrical can.

I'm designing a can with radius 2 cm. My can has a radius of 5 cm.

1. Each of these cans holds 200 cm³. Find the heights of these two cans.

Are the dimensions of the cans suitable? Explain your answer.
2. Find the surface area of the two cans. Show your work

3. In order to keep costs low, the Fresha Drink Company wants to sell the drink in cylindrical cans that use the smallest amount of aluminum.

Find the radius and height of a can which holds 200 cm³ and uses the smallest amount of aluminum. Show clearly how you figured out the size of the can. Make your dimensions correct to the nearest 0.5 centimeter. *(You may find it helpful to use graph paper.)*
Circle Pattern

This problem gives you the chance to:

• explore fractions in context

Here is a developing circle pattern.

Here is one black circle.

Two white circles of half the radius have been added to the diagram.

1. Show that the fraction of the diagram that is now black is one half.

____________________________________
____________________________________
____________________________________
____________________________________

Four black circles have now been added.

2. What fraction of the diagram is now black?

____________________________________
____________________________________
____________________________________
____________________________________
3. Fill in the table to show what happens as the pattern continues.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Black fraction</th>
<th>White fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>One black circle</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Two white circles</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
</tr>
<tr>
<td>Four black circles</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eight white circles</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sixteen black circles</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. Write a description of what is happening to the black and white fractions as the pattern continues.

________________________________________________________________
________________________________________________________________
________________________________________________________________
Hopewell Geometry
This problem gives you the chance to:
• work with the Pythagorean Rule, angles and similarity in given triangles

The Hopewell people were Native Americans whose culture flourished in the central Ohio Valley about 2000 years ago.

The Hopewell people constructed earthworks using right triangles, including those below.

1. What is the length of the hypotenuse of Triangle H?
   Give your answer correct to one decimal place.

   _______________

   Show your calculation.

2. What is the size of the smallest angle in Triangle A?
   Give your answer correct to one decimal place.

   _______________

   Show your calculation.
The diagram on the next page shows the layout of some Hopewell earthworks. The centers of the Newark Octagon, the Newark Square and the Great Circle were at the corners of the shaded triangle.
The three right triangles surrounding the shaded triangle form a rectangle measuring 12 units by 14 units.

Each of these three right triangles is similar to one of the Hopewell triangles on the previous page.

For example, Triangle 3 above is similar to Hopewell Triangle C.

3. Which Hopewell triangle is similar to Triangle 1? _______________
   Explain how you decided.
   __________________________________________________________________
   __________________________________________________________________
   __________________________________________________________________

4. Is the shaded triangle a right triangle? _______________
   Explain how you decided, showing all your work.
   __________________________________________________________________
   __________________________________________________________________
Sidewalk Patterns

This problem gives you the chance to:
• work with patterns
• work out the \( n^{th} \) term of a sequence

In Prague some sidewalks are made of small square blocks of stone.

The blocks are in different shades to make patterns that are in various sizes.

![Pattern #1](image1)
![Pattern #2](image2)
![Pattern #3](image3)

1. Draw the next pattern in this series.
2. Complete the table below

<table>
<thead>
<tr>
<th>Pattern number, n</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of white blocks</td>
<td>12</td>
<td>40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of gray blocks</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of blocks</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. What do you notice about the number of white blocks and the number of gray blocks?
________________________________________________________________________

4. The total number of blocks can be found by squaring the number of blocks along one side of the pattern.
   a. Fill in the blank spaces in this list.

   \[ 25 = 5^2 \quad 81 = ____ \quad 169 = ____ \quad 289 = 17^2 \]

   b. How many blocks will pattern #5 need? _____________________

   c. How many blocks will pattern \(n\) need? _____________________

5. a. If you know the total number of blocks in a pattern you can work out the number of white blocks in it. Explain how you can do this.
________________________________________________________________________
________________________________________________________________________

   b. Pattern #6 has a total of 625 blocks.
   How many white blocks are needed for pattern #6? _____________________
   Show how you figured this out.
Hi there - Attached, please find a copy of my testimony which was presented during the 11/18 hearing in Atlanta. It was requested that I forward this hard copy to this email address.

Please let me know if I can provide any additional information. Thanks! Kim Hymes

---

Kim Hymes  
Director, Policy & Advocacy  
Council for Exceptional Children  
1110 North Glebe Road, Ste. 300  
Arlington, VA 22201  
P 703.264.9441  
F 703.243.0410

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

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

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

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

**Creating Assessments that are Accessible to Diverse Learners**

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

Specifically, CEC urges the Department to fund grants that create assessments which:
- Use multiple measures that are norm referenced on students with disabilities
- Are formative and summative in nature in an effort to provide educators with useful feedback
- Take into account accommodations and modifications
- Utilize the principles of Universal Design for Learning
- Include computer adaptive testing
Creating better Alternate Assessments based on Alternate Achievement Standards (AA-AAS) and Alternate Assessments based on Modified Achievement Standards (AA-MAS)

As you know, current federal policy allows states to use an alternate assessment based on alternate achievement standards (AA-AAS) and an alternate assessment based on modified achievement standards (AA-MAS) for certain students with disabilities. While this policy has been in place for some time, the consistency and availability of these assessments varies widely between states.

A recent study by the National Center for Special Education Research\(^1\) within the Institute Of Education Sciences, found that many states approach the AA-AAS differently. Some states use a portfolio or body of evidence to constitute the entire assessment. Others use techniques such as a rating scale/checklist, performance task/events, or multiple choice/constructed response assessments. The inconsistent approach to these assessments across states creates varying standards and expectations and fails to provide the information we need to accurately determine a student’s achievement.

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

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

Creating Assessments that Provide Meaningful Feedback to Educators & Families

As the Department considers its grant proposal, CEC encourages the Department to place a strong emphasis on the importance of creating assessments that yield meaningful information for educators and families. Assessments should be tools that help inform instruction, identify areas of strength and weakness, and help inform decision making. However, assessments can only be effective if they are presented in a way that enables a student to accurately demonstrate their knowledge and skill. Educators need meaningful professional development to help them understand how to

use assessment data to inform and drive instruction. Parents need to understand what complex scores show about how their child is learning, and educators must be able to describe results and help parents interpret this complex data meaningfully.

To this end, CEC encourages the Department to fund grants that included professional development and training. Considering how assessments can provide meaningful feedback to educators and parents from the first stage of assessment creation, will help ensure their success.

**Conclusion**

In conclusion, CEC appreciates this opportunity to provide feedback as the Department moves forward in funding grants through the RTTT Assessment Competition and we will submit written comments that expand on these issues and discuss students with gifts and talents. All students will benefit from assessments that allow them to effectively demonstrate their knowledge and skill. Our ability to have a true understanding of how our students are performing depends on having accurate assessments from which to evaluate them by.
Thoughts on Testing Arising from the Common Core State Standards Movement

Terrance Paul, CEO, Renaissance Learning, Inc.,

December 20, 2009
Introduction

The state-led common core standards movement provides an important opportunity to improve student achievement and our international competitiveness. In addition, the common core state standards movement puts summative testing on the front burner, creating an opportunity to also reduce the cost of summative testing and provide the means to invest in a balanced assessment system that incorporates both summative and formative elements. While summative assessments are clearly essential to satisfy the purpose of state oversight of district and school performance or “accountability,” it is universally recognized that the primary drivers of improved student achievement are interim and formative assessments, designed to inform instruction, differentiate learning, and support the important school improvement framework called Response to Intervention (RTI). Finally, there is a movement toward computer-based testing (CBT), particularly computer-adaptive testing (CAT), to reduce assessment costs, provide more information, and save teachers time while making it practical to increase the frequency of interim testing. Legislatures, however, will be reluctant to approve monies to expand interim assessments and computer-based testing unless the cost of summative testing can be reduced.

Renaissance Learning is in a unique position to provide input to deliberations on these issues. Renaissance is the leading provider of computer-based tests for U.S. schools. We administer more than one million web-based computerized tests to K-12 students per day. Our STAR computer-adaptive tests are the most widely used computerized reading and math tests, in addition to being among the most highly rated by the federally funded National Center on Response to Intervention (NCRTI) and by teachers themselves. Teachers like our STAR reading, early literacy, and math tests because they are easy to use, short (typically less than 15 minutes), self-administered by students, inexpensive, and provide immediate feedback.

Our perspective on testing is perhaps unique because it is based on our experience working with hundreds of thousands of teachers in thousands of schools. We build tests with usability in mind in addition to the traditional criteria of reliability and validity—tests that teachers like and use in their classroom daily to inform instruction. We know that if teachers do not see the immediate practical use for the information collected and if the student spends too long at the computer, the test won’t be used. It won’t be used especially in classrooms where it’s most needed. Access, time at the computer, and the number of computers that work, is the single biggest impediment for the expansion of computer testing.

To make tests usable and inexpensive, we focus on maximizing the fundamental ratios I/t and I/C, the amount of useful information (I) collected versus the time (t) it takes to collect it and the amount of useful information (I) collected versus total cost (C). Generally the most expensive component of testing cost is teacher time rather than the cost of the test itself. Test length, the time the student is at the computer, is directly correlated to cost. Keeping computer tests short is key to reducing their cost, while increasing teacher satisfaction and usefulness.

We believe there is now an unprecedented window of opportunity to get student testing right because of the state-led initiative for common core standards and the need for improved lower cost tests. Unfortunately, there is an almost equal opportunity to get it wrong. We offer the following five observations regarding building a balanced and coherent assessment system to accelerate student achievement.
1. **Minimize deployment time.**

Perhaps the most important factor to consider is deployment time. We cannot wait five or more years to roll out new tests to improve student achievement. No one can wait for better, less expensive tests—not students, teachers, parents, or politicians.

The new common core state standards will be available this spring. We should be thinking of what can be done for the next school year. Certainly two years would be too long to wait. There are many who say that now is the time to not only move to full CBT for summative testing but also to add more and new types of open-ended (also called constructed response) and performance items, which will take many years to research and develop. Compared to multiple-choice items, open-ended items, while useful in formative assessments, are problematic for summative tests because they make tests longer and more expensive and can also cause the gap between groups to be overstated. This is because linguistic, writing, and keyboarding skills are confounded with the target academic construct whether math, reading, or science. While it’s bad to understate the gap, it’s worse to overstate it.

The fact is we can’t wait for the development of more complex and unproven tests. Instead, states need to roll out improved testing systems aligned to the new standards in stages. First, deploy the kinds of assessments that are already proven, which can be developed inexpensively and rolled out quickly. After that it may be appropriate to deploy more complex tests, but only after the research, development, and field testing are complete and the costs known.

By prioritizing carefully and using primarily multiple-choice items, which yield the most information with the least investment of time and money, it is very reasonable to expect that improved state assessment systems and reduced-cost summative tests can be deployed in less than two years from state adoption of common core and state specific standards.

2. **Interim benchmark, progress monitoring, screening, and mastery assessments are more important to roll out quickly than once a year summative tests.**

No one questions that interim assessments are more important to improve student learning than the once-a-year summative tests, and for good reason: their purposes are different. The once-a-year summative test purpose is accountability, to measure the effectiveness of schools and districts. The purpose of interim assessments is to inform instructional decisions to improve student learning. The summative test purpose is long-term; the interim test purpose is immediate and key to improving student achievement.

Two categories of interim tests need to be deployed quickly: 1) periodic benchmark assessments aligned to state specific and common core standards and 2) assessments which support the research-based elements of effective instruction, including instructional match and differentiated learning. These latter types of interim assessments have already been defined by the federally funded NCRTI, namely progress monitoring, screening, diagnostic, and mastery tests. RTI has become the primary research-based framework for school improvement nationwide, supported by both state and federal funding and, as such, needs to be supported by the assessment system.
3. **Computer-based tests should be short; access is the big issue.**

Though we are the leading provider of computerized tests for schools, we would be the first to admit that paper is hard to beat, especially for high-stakes summative tests. High-stakes tests must be secure, measure growth, and provide comparable information across schools and districts. Now with common core state standards, states will also want comparability across states. In the case of interim and other formative assessments, frequency and instant feedback are more important issues to consider. (NCRTI requires progress-monitoring tests to have a sufficient number of items to administer equivalent forms at least monthly.)

The single biggest problem with going to full CBT for summative tests in U.S. schools today is computer access. Our experience in classrooms across the U.S. has shown us that there are huge disparities in access, across states, across districts within states, between grades, and even disparities between schools within the same district. There are disparities in the number of computers, the number of computers that work, and the amount of time the students are allowed to become familiar with computer assessments. Published surveys on the number of computers per student, in our experience, considerably under-estimate the disparity.

While it will take a long time to reduce or eliminate the disparity in computer access in U.S. schools, the impact of the disparity in computer availability can be greatly reduced by making computer tests short. This has been our approach with our widely used STAR tests; they are very short, less than 15 minutes, yet provide similar reliability to a 50-minute paper assessment. STAR tests are short because of computer-adaptive technology (CAT) and careful attention to what we test and how we test it using very efficient multiple-choice item types.

In addition to making computer-based tests shorter, the computer access problem can be further mitigated by having computer-based interim assessments that mirror the look of the summative test so all students have an opportunity to practice both with the computer and item types.

We believe high-stakes tests could be much shorter and thus support computer testing sooner rather than later. The next two sections explain why.

4. **New developments from cognitive science suggest multiple-choice items measure more than just knowledge.**

New developments in cognitive science regarding how our brains work shows knowledge is more important to higher-order thinking than once believed; particularly the interplay between limited short-term working memory and long-term memory and the importance of what is called “chunking.” Wide knowledge, including vocabulary and even “mere facts,” greatly facilitates higher-order thinking (Willingham, 2009). Knowledge, it turns out, is not only the fuel for thinking but is itself constructed through higher-order thinking (understanding, application, analysis, synthesis, and evaluation) and the doing, producing, and performing of tasks. In other words, when you test for knowledge, you also test the capacity to think and do.

These findings from cognitive science explain why the criticism that multiple-choice items cannot test higher-order thinking is wrong. This actually is old news to the psychometric community. Study after study since the 1970s has shown multiple-choice to be the equal of open-ended items in testing higher-order thinking skills (Bennett, Rock, & Wang, 1991; Bridgeman,
Undoubtedly, this is why college entrance exams such as the ACT and SAT, the most high-stakes tests in the land, use mostly multiple-choice questions. ACT uses 100% multiple-choice and, while the SAT includes a writing test which is optional in the ACT, college admission offices typically either do not consider the writing score or give it little weight, which further suggests the importance of multiple-choice items.

Not only can multiple-choice items test higher-order thinking skills as well as open-ended items, but they do so at far less cost with much higher I/t. Also, as previously mentioned, open-ended items introduce bias because of their higher linguistic load. For example, a study of the international TIMSS test (which uses both multiple-choice and open-ended items) showed open-ended math and science items favored girls versus boys since girls, on average, exhibit higher language abilities, though the overall country rankings were not significantly changed (Hastedt & Sibberns, 2005). If the country rankings don’t change, gender bias is introduced, and open-ended items make the test longer, one is left with the question as to why TIMSS and other summative tests continue to use them.

The answer to that question is that the problem with multiple-choice items is not that they can’t measure higher-order thinking skills, but rather they don’t look like they do. Multiple-choice items lack what is called “face validity.” Face validity of course can lead to mistaken beliefs, for example, the mistaken belief that the sun revolves around the earth. For hundreds of years institutional and political barriers prevented this mistaken belief based on face validity from being disproved by science. Many institutional and political barriers also stand in the way of the use of multiple-choice items, not the least of which are the revenues generated for research, development, and scoring for ever more complex and expensive tests. The federal Government Accounting Office (2003) estimated that multiple-choice summative tests under NCLB would cost states $1.9 billion, but mixed-item formats (multiple-choice plus open-ended items) would cost $5.3 billion.

In addition to face validity issues, mistaken beliefs about multiple-choice tests are also partly caused by confusing the purposes of summative versus interim assessments. Summative tests are not intended to inform instruction like interim and formative assessments. In the case of formative assessments, open-ended items can be useful to help teachers identify incorrect thinking processes. The admonition by math teachers to “show your work” rings in many adult ears decades after they were in the classroom. Clearly, performance is essential to assess subjects such as writing or the playing of a musical instrument. However for summative accountability tests in reading, math, science and social studies, the case for open-ended items is at best weak and at worst harmful, especially if the excessive cost of the summative test slows the deployment of interim and formative assessments which directly help teachers improve student achievement.

5. Perhaps it is time to reduce the summative test footprint.

It is understandable that state legislatures will be reluctant to increase spending on interim and formative assessments at the school and district level unless there are cost savings on the state summative test. Clearly, using multiple-choice items in summative tests can significantly reduce costs. The question is whether it might be possible to reduce the length of the summative test even more so they can be computer delivered utilizing computer-adaptive testing. Computer-
adaptive testing has many benefits such as reducing test time and eliminating the out-of-range problems in the testing of low-achieving students on one end and gifted students on the other.

The answer to this question of whether summative tests can be made even shorter is “yes” if the purpose of the summative test is defined narrowly for accountability, but the answer may change if there are other broader purposes.

If the purpose of the summative test is solely accountability, that is to satisfy the state need to measure the effectiveness of districts and schools and growth of subgroups, it would seem that most state tests are already too long. School and district performance can be measured with shorter tests than those currently used because it is not necessary to know how well each student is performing, but rather how well groups of students are performing.

The enemy of short and inexpensive is not accountability but rather other purposes such as reporting on individual student achievement and using that information to measure teacher effectiveness. In both cases, it seems that this should be the responsibility of the local educational authority (LEA) versus the state educational authority (SEA).

In the case of reporting on individual student performance, districts and schools already have more and better information than states. The combination of information from interim low-stakes assessments administered throughout the year together with student portfolios provides LEAs more complete and nuanced information on what students know and can do than a single high-stakes test no matter how large.

With regard to using student test results and value-added scores to measure teacher effectiveness, clearly this should also be a job for the LEA. First, the summative test is so limited in its coverage that perhaps at most only 30% of teachers in a K–12 system would have directly relatable test results. K–2 teachers, music, health, social studies, special ed, reading specialists, and many other middle and high school teachers will have no relatable test results. Second, even for the 30% of teachers for which there are state summative test results, many researchers have concluded because of reliability issues it would not be fair to put much weight on one or two testing events for something as complex as evaluating teacher performance (Braun, 2005). The LEA has more testing information and other quantitative and qualitative information on all staff to more fairly evaluate teachers than the SEA. Finally, of course, it’s the LEA and not the SEA, who contracts with teachers or teacher unions.

Just as it is important to not define the purpose of the state summative test too broadly, it is important to not define it too narrowly. There is now a call for more summative end-of-course testing of high school students in grades 9–12. With more end-of-course testing, some suggest that testing students in reading and basic math should not be done beyond the eighth grade. This would be a mistake.

While it is sometimes said that the gateway skill for college is algebra, the only true gateway skill is reading. Reading is the fundamental skill underlying all academic subjects. The next most important gateway skill for college after reading is basic math—the math and computational skills leading up to algebra. The adoption of high standards do not change the fact that
employers, tech schools, and colleges should be able to trust that a high school diploma means
that, at a minimum, students can read well and can do basic math and that, therefore, remedial
reading and math courses are not needed. Many studies have shown that students do very little
reading in high school and do even less math practice. A skill not practiced is forgotten. Testing
students in the eighth grade in reading and math does not tell you if students can read well and
do basic math at graduation. Therefore, testing reading and basic math through 12th grade is
essential for summative testing.
We cannot wait five or ten years to roll out new tests to improve student achievement. No one can wait for better tests—not students, teachers, parents, or politicians. It is urgent that we deploy both the new common core state standards and better more balanced systems of assessment. There is particular urgency to deploy interim assessments to directly inform instruction and support RTI.

Clearly for a variety of reasons, including cost, security, comparability, fast deployment, computer testing, and the need for expanded interim testing, the summative test must be dramatically reduced in length. With the state-driven development of common core state standards, careful definition of the purpose of state summative testing and use of multiple-choice items, the goal of reducing the cost and footprint of the summative test is very achievable and with it the goal of providing more balanced and coherent testing systems to accelerate student achievement.
References


