STATE INNOVATION PRIORITIES

for State Testing Programs

Arizona, Colorado, Florida,
Georgia, Maryland, Massachusetts,
Michigan, Ohio, Pennsylvania,
Tennessee and Texas

Developed by

AccountabilityWorks

Contributors:
John H. Oswald and Theodor Rebarber. Review and input by Richard W. Cross and S. E. Phillips. Editorial assistance by Thomson W. McFarland.

ELC Education Leaders Council

Executive Summary

Over the last decade, testing has become recognized as a key lever in states' efforts to establish accountability for results in public education. The role of accountability in reform will only be enhanced by the new testing and accountability requirements of the *No Child Left Behind Act of 2001* (Public Law 107-110). To address the demand for "better, cheaper, and more" testing, states have chosen to collaborate on priorities for innovation, with the benefit of testing industry input and participation. As the product of this collaboration, the Education Leaders Council and AccountabilityWorks have assisted participating states in developing these *State Innovation Priorities for State Testing Programs* in order to define clearly state expectations for innovative testing practices that should achieve widespread adoption and implementation. This effort included an intensive Testing Summit held on February 20-21, 2002 in Austin, Texas, attended by state education and testing industry leaders, where participants provided input and suggestions regarding this document. (A list of Test Summit attendees is on pp. 19-24.) The ELC Summit was made possible by a grant from the U.S. Department of Education to the State Education Policy Network (SEPN).

In a market economy, it is expected that industry will respond to clearly articulated needs with innovative products and services. Some needs, however, may best be addressed through innovative cooperation between interested states. This document presents 25 innovations of both kinds in a table divided into two time periods, with these further divided according to whether the innovations represent a "high" priority. The table includes concise descriptions for each innovation of potential benefits as well as major challenges. Following this table, there is an extended discussion organized according to state needs: Test Content, Technical, Logistics, Interpretation, Application, and Business. It is worth noting that a number of the innovations described herein are already being implemented in some form by diverse state agencies, testing companies, and research organizations, though none are widespread and implemented to their full potential. Development work on many of the short- to mid-term innovations—such as the Common Item Bank—needs to be intensive now if they are to be ready for implementation in the short- to mid-term. The fact that there is a degree of knowledge and experience with many of these innovations encourages us to view them as promising, and therefore worthy of inclusion.

The list of innovations, grouped by timeframe and relative importance, is as follows:

TIME PERIOD 1 INNOVATIONS (IMMEDIATE):

High Priorities

1 Computerized or Web-based Reporting

2 Coordinated Diagnostic Tests

3 Collaborative Resource on State Assessment
4 Reliable System for Linking Test Records to Students
5 Computerized Scoring of Open-ended Responses
6 Integrated Test Readiness Programs

Other Priorities
7 Customized Off-the-shelf Tests
8 Randomized Forms
9 Local Scoring Option
10 Integrate Important Benefits of Norm-referenced Tests into State Standards-based Tests
11 Enhanced Quality Control Procedures
12 Distributed Scoring
13 Valid Program Evaluation Models
14 Standard Procedures for Bias Studies
15 Reports of Teacher Effectiveness

TIME PERIOD 2 INNOVATIONS (SHORT-TERM TO MID-TERM):

High Priorities
1 Common Item Bank
2 Computer-administered Tests
3 Coordinated Multiple Measures

Other Priorities
4 Multi-tiered Continuous Assessment Systems
5 Power Objectives
6 Staff Development in Assessment Development and Use for Teachers
7 Consistent Criterion-referenced Interpretive Methodology for Classroom Assessment
8 Teacher Certification in Assessment Knowledge
9 Individually-administered Clinical Tests to Measure State Academic Standards
10 Improved Early Childhood Tests

This electronic document is the full version of the printed, abridged report.

Introduction

Over the last decade, testing has played an increasingly prominent role in the campaign to improve our nation's public schools. Parents, employers, policymakers, and, increasingly, educators have moved steadily toward a greater focus on results in exchange for every education tax dollar—including better results for groups that historically have not achieved at high levels. Testing is the instrument that has allowed this shift to occur. Currently, nearly all the states and territories have instituted some form of assessment program for accountability.

New challenges, however, have emerged with the increased emphasis on testing. Because educational policy is not nationalized and is generally reserved to the states, the states have proceeded independently of each other in developing and implementing many testing programs that are redundant with testing programs in other states. Given the diversity of the nation and the poor track record of previous educational reform initiatives, decentralization of educational policy has provided the benefits of innovation and avoidance of big mistakes, but it has also resulted in strain for the industry and high costs for states. In many cases, the testing industry has struggled to keep up with state expectations regarding the volume of new test development, scoring, and program delivery. These demands will only increase under the requirements for additional testing in the *No Child Left Behind Act of 2001* (Public Law 107-110).

The testing industry and the states are dependent upon each other for the success of state assessment programs and educational improvement. It is therefore appropriate for states to voluntarily engage in collaboration on testing innovation priorities with the benefit of industry input and participation. To identify such priorities, the Education Leaders Council and AccountabilityWorks, Inc., have assisted participating states in developing these *State Innovation Priorities for State Testing Programs* in order to define clearly their expectations for innovative testing practices that should achieve widespread adoption and implementation. This effort included an intensive Testing Summit held on February 20-21, 2002 in Austin, Texas, attended by state education and testing industry leaders, where participants provided input and suggestions regarding this document. (A list of attendees is included at the end of this document.)

This document presents needs for the future of state testing, especially in light of P.L. 107-110. It is designed as a catalyst for discussion between the leaders of the testing industry and the leaders of statewide education systems. In a free enterprise economy, an important market need will result in innovations from the providers of products and services to meet that need. In this case, an established need of the states should result in the testing industry's development of innovative products and services to improve customer satisfaction. Further, some of

the testing innovations described here will not result from new or improved products from industry but from enhanced cooperation between states. After a number of years of experimentation with testing, states are at a point where they can intelligently discuss ways to work together that strengthen their ability to implement accountability systems efficiently while also accelerating reform.

This document is organized in two parts. First, a table summarizes the innovations by A) timeframe for implementation and B) whether each innovation is a high priority. The table also succinctly describes key benefits and challenges for each innovation. Second, an extended discussion of the innovations is organized according to state needs: Test Content, Technical, Logistics, Interpretation, Application, and Business. As the needs are presented, the discussion identifies the innovations that could help meet them. Some innovations (e.g., computerized testing) address multiple sets of needs.

This document presents 25 innovations, divided into two time periods. Innovations in time period 1, labeled 1.01 through 1.15, address pressing needs for which the necessary technology, personnel, or other infrastructure are ready for immediate implementation. Innovations in time period 2, while similarly important, are on a longer timetable due to technological, logistical, or statutory obstacles that must be overcome for their widespread implementation. Development efforts on these short- to mid-term innovations—such as the Common Item Bank (2.01)—need to be intensive already or to begin immediately if they are to be ready for implementation in the short- to mid-term. Apart from the division of innovations into those that are "high priorities" and those that are "other priorities," no judgment about the relative importance of the innovations should be inferred from the order in which they are listed. It should be noted that the quality and effectiveness of these innovations, particularly those that are to be implemented now, will improve over time. Thus, innovations that are implemented immediately might only produce a modest effect in the short term, but have a substantially greater effect as the innovation is researched and refined.

It is also worth noting that several testing companies, research organizations, and state agencies are creating or have implemented some of the innovative practices that are described herein. The reader should assume that there is a degree of knowledge and experience with many, if not all, of the innovations described. As a matter of fact, some of these practices have been available for years (e.g. randomized forms and customized norm-referenced tests). Those practices are "innovations" for the purpose of this document only in that they are not yet widely applied and can stand enhancement through either technology or research.

Throughout this document, the terms "testing companies" and "industry" apply generically to refer to all providers of test content, printing, scoring, validation and other testing-related services, whether for-profit or not-for-profit, public or private. The term "state" applies to the educational enterprise of the fifty states, territories, and other appropriate jurisdictions (such as the Department of Defense Dependent Schools) and includes state education agencies (SEAs), state boards of education, and other official educational entities.

Innovations Table

TIME PERIOD 1: IMMEDIATE

Innovation	Description	Category	Benefits	Challenges
HIGH PRIORITY INNOVATIONS				
1.01: Computerized or Web-based Reporting	A paperless system of providing real-time accessibility to reports either on a school computer network or on the Internet. Ideally these reports would be interactive, allowing cross-referencing and detailed investigation and explanation of scores for multiple types of audiences.	Information Technology	■ Decreases time and cost of paper report handling, assuming some functions are replaced ■ Allows interactive reports that are more informative and tailored to the individual needs of the reader ■ Allows more rapid reporting	■ Requires computers, networks and infrastructure. ■ Requires system maintenance ■ Requires security controls ■ Not accessible to all, especially parents of some subgroups
1.02: Coordinated Diagnostic Tests	Systems of diagnostic, classroom-based assessments that are keyed to the state standards and assessments. These tests allow teachers to do formative evaluations throughout the year to improve student performance.	New Products	■ Involves teachers more deeply in state assessment system ■ Monitors student progress more frequently ■ Provides for practice in test taking skills ■ More timely results for diagnosis and remediation	■ Require staff development in their appropriate use ■ Need to be integrated with teaching time in a way to add and not detract from instruction ■ High cost

TIME PERIOD 1: IMMEDIATE

Innovation	Description	Category	Benefits	Challenges
HIGH PRIORITY INNOVATIONS				
1.03: Collaborative Resource on State Assessment	An organization serving as a resource to states on a range of challenges related to assessement: a) the legal and psychometric defensibility of assessments; b) timelines for testing program schedules; c) state/vendor contract templates; d) public defense and communication on state testing; e) "safe harbor" model accommodations classification system for special needs students, with options for combining scores; f) similarly, standard procedures for bias studies; g) voluntary system of describing results using standards-based achievement levels h) other issues as needed.	Collaboration	■ Benefit from the diverse past experiences of states ■ Cost savings in solving several of the challenges ■ All states benefit from top national experts ■ Accelerates implementation ■ Greater range of experiences allows for faster and better revisions where needed ■ Allows for "united front" in states' efforts to publicly defend testing programs	■ Some differences in state law and regulations add complexity ■ Who pays?

TIME PERIOD 1: IMMEDIATE

Innovation	Description	Category	Benefits	Challenges
HIGH PRIORITY INNOVATIONS				
1.04: Reliable System for Linking Test Records to Students	A set of affordable methods to reliably attach test results to students without unreasonable administrative effort, and even when students are mobile.	Information Technology	■ Meets requirements of law ■ Allows for student tracking and thus better student instruction when information is properly applied ■ Allows for more effective program evaluation ■ Allows for more effective and fair teacher evaluation ■ Lower administrative cost	■ High initial investment in developing system ■ Requires cooperation of many agencies ■ Possible privacy threats must be dealt with ■ May be politically or legally difficult in some states ■ Pre-gridding adds to time before the test
1.05: Computerized Scoring of Open-ended Responses	A method by which open-ended questions are scored entirely or partially by computer with little or no human scoring.	Information Technology	■ Reduces one of the highest cost components of the assessment system – hand scoring ■ Reliability and accuracy has been shown to be equal to or better than human readers in some studies ■ Extremely sophisticated logic can be applied to many innovative item types ■ Allows the cost-effective development of multiple measures ■ Can be used alone or in combination with human reading ■ Increased speed of scoring if used alone	■ Competing systems use different artificial intelligence models, so a model must be chosen first ■ Possible credibility problems with the general public and with teachers ■ Possible for students to learn techniques to "fool" some of the systems ■ Requires either computer-administered tests, or conversion of handwritten responses into computer readable text (not just images) ■ Works better for some types of open-ended questions (e.g., essays) than others (e.g., short answers)

TIME PERIOD 1: IMMEDIATE

Innovation	Description	Category	Benefits	Challenges
HIGH PRIORITY INNOVATIONS				
1.06: Integrated Test Readiness Programs	Programs that simultaneously teach meaningful content around the state academic standards while improving test-taking skills.	New Products	■ Protects instructional time while still providing test preparation ■ Makes scores more valid ■ Improves equity	■ More costly than simple test prep ■ Requires carefully integrated development of tests and test readiness curricula ■ Requires staff development to be used properly

TIME PERIOD 1: IMMEDIATE

Innovation	Description	Category	Benefits	Challenges
OTHER INNOVATIONS				
1.07: Customized Off-the-shelf Tests	Publishers' norm-referenced or criterion-referenced tests that are supplemented with sets of items to round out the measurement of state academic standards.	Content Sources	■ Less expensive ■ Faster to develop ■ Provide normative data for measuring gains and other purposes ■ High technical quality of core test	■ Measurement of some standards (those on shelf test) is preset to the publisher's item types ■ Not as unique to the state as a totally custom test. ■ Would not allow for release of items. ■ Some potential psychometric problems with norm-referenced scores if publisher's test is changed enough
1.08: Randomized Forms	Tests in many alternate forms administered across students so that students in the same class receive different forms. While this is best done in computer administration, it can be accomplished to some degree in paper and pencil administrations.	Assessment Structure	■ Cost efficient way of improving security, depending upon the difference between forms ■ Allows efficient try-out of new items	■ Psychometric issues with comparability must be addressed.

TIME PERIOD 1: IMMEDIATE

Innovation	Description	Category	Benefits	Challenges
OTHER INNOVATIONS				
1.09: Local Scoring Option	A system whereby LEAs can score tests locally for immediate results prior to sending answer media to the test scoring contractor for official statewide scoring and reporting.	New Products	■ Allows rapid scoring and generation of reports needed for instructional intervention and program selection (e.g., summer school)	■ Requires an investment in local scoring capability, including staffing and training ■ Would not help turnaround time with items requiring hand scoring ■ Could be more difficult with complex programs requiring coordination of scoring of multiple booklets ■ Security problems ■ Scannable forms sometimes scan differently the second time because of carbon transfer ■ Redundant scoring is not cost effective
1.10: Integrate Important Benefits of Norm-referenced Scores into State Standards-based Assessments	Data providing comparisons of performance—within-state or national—that add to the informative power of state assessments, as well as key design elements such as vertical equating, without replacing or compromising the criterion-referenced interpretation system.	New Products	■ Allows better understanding by stakeholders ■ Allows for value-added assessment	■ Requires overcoming some philosophical objections to comparisons ■ For national comparisons, requires some elements of standardized testing or NAEP items

TIME PERIOD 1: IMMEDIATE

Innovation	Description	Category	Benefits	Challenges
OTHER INNOVATIONS				
1.11: Enhanced Quality Control Procedures	Multiple improvements in quality control in printing and scoring tests, including: automated methods of monitoring quality control in scoring and detecting errors before they affect reporting; application to the testing industry of established industry standards in manufacturing, project flow, software and communications.	Information Technology	■ Efficient methods to improve scoring accuracy without expensive, and fallible human intervention ■ Applications of accepted industry standards, such as ISO 9000 and others	■ Not foolproof and does not eliminate the need for careful planning, execution, and state review ■ Certification in industry quality control standards is very expensive for testing companies and costs could be passed on to customers
1.12: Distributed Scoring	A method for scoring open-ended responses more efficiently by either (a) capturing the student's written response using imaging technology, or (b) having the student enter responses directly into a computer, and sending the responses to trained scorers "at large" who can work from their homes or offices on their own schedules. The scorers score the items on their personal computers and relay the scores, which are then reconciled with the second scorer and accepted or referred for a resolution scoring. The same quality control procedures are used as for hand scoring in scoring centers.	Information Technology	■ Reduces cost of hand scoring by allowing individuals to score from their homes and offices on their own schedule, which often means having to pay a lower rate for the convenience ■ Allows the use of highly trained and qualified scorers who otherwise would not be willing to do the work, thus improving quality ■ Allows a larger number of scorers in a more flexible manner, thus decreasing turnaround time	■ Still requires centralized training ■ Greater requirement for monitoring scoring drift and agreement since the scorers are not under the visual supervision of management ■ Requires both imaging technology and scanning (adds some cost) or computer administration of the test. (Image scanning might not be an added cost if the documents are scanned anyway for other purposes)

TIME PERIOD 1: IMMEDIATE

Innovation	Description	Category	Benefits	Challenges
OTHER INNOVATIONS				
1.13: Valid Program Evaluation Models	Testing industry should cooperate with researchers working on models for using state assessment programs for program evaluation.	Collaboration	■ Makes test results more valuable for determining what works ■ Allows for more valid decisions than simple score observations	■ Requires careful research ■ Some models require more elaborate, expensive programs with more types of data provided.
1.14: Standard Procedures for Bias Studies	Agreement on methodologies for bias studies (i.e., which bias analyses should be performed) that are known and funded by law-making bodies.	Collaboration	■ Creates a set of model standards, saving states from original work each time ■ More legally defensible ■ Funding should be easier	■ Requires agreement on which bias analyses should be performed ■ Requires sign-off by authorities ■ States would be choosing to follow established procedures rather than inventing their own
1.15: Reports of Teacher Effectiveness	Valid reports of student gains while under the instruction of an individual teacher, with spurious data eliminated. Reports would aggregate data from multiple groups of students taught by the same teacher.	New Products	• Provides fairer measure of teacher performance than current simplistic models • Helps teachers improve their instruction • Helps administrators work with teaching staff to improve student learning	• Requires good longitudinal tracking and collection of demographic information. • Models need further research. • Impacts early planning of entire assessment system so that the proper data is obtainable

TIME PERIOD 2: SHORT-TERM OR MID-TERM

Innovation	Description	Category	Benefits	Challenges
HIGH PRIORITY INNOVATIONS				
2.01: Common Item Bank	Sets of calibrated and researched items and test elements tied to a massive set that is the combination of participating state academic standards. Work would begin immediately, with the first products available in three or four years. The items and test elements would be kept in a computerized content management system for efficient and secure access by qualified individuals who could assemble tests as needed.	Content Sources	■ More cost-efficient, whether the items are released annually or kept secure and reused ■ Less wasteful than developing all new items in each state to measure, substantially, the same objectives ■ Higher technical quality of tests by using items with more research	■ Requires some entity to monitor access and manage the system ■ Effectiveness depends upon the degree to which same items can be used for multiple states with different standards; increased cooperation at earlier stages, such as development of standards, would increase efficiency and savings
2.02: Computer-administered Tests	State assessments administered to every student on computers without the need for printed tests.	Information Technology	■ Lower cost (once initial investment is made) ■ Allows many other innovations ■ Increases speed of development, scoring and reporting ■ Allows for new item types impossible with paper and pencil	■ Initial cost is considerable (more computers, networks, development and training) ■ Requires technological maintenance in states, LEAs and schools ■ Some believe that, like other technology innovations in the classroom, many of the potential benefits will never be realized ■ Fairness to minority students who lack equal access to computers at home or during other school hours might create opportunity to learn challenges

TIME PERIOD 2: SHORT-TERM OR MID-TERM

Innovation	Description	Category	Benefits	Challenges
HIGH PRIORITY INNOVATIONS				
2.03: Coordinated Multiple Measures	Measures of academic standards using more than one measurement methodology, e.g., performance assessments or technology assessments, but coordinated with the primary assessment for fairness and comparability. These could include non-assessment data, like grade point average.	New Products	■ Provides coverage of some standards that are hard to assess ■ Meets a requirement of P.L. 107-110	■ High cost ■ Requires research for fairness

TIME PERIOD 2: SHORT-TERM OR MID-TERM

Innovation	Description	Category	Benefits	Challenges
OTHER INNOVATIONS				
2.04: Multi-tiered Continuous Assessment Systems	Integrated sets of assessments consisting of a standardized, statewide, every-student testing program for high stakes and mid-stakes and/or low-stakes assessments of the same or related standards that are administered by classroom teachers in informal programs. The results are integrated to create a comprehensive picture of student achievement and reported in both integrated and separated forms. The scoring and recording of student performance would be integrated throughout the year into a database.	Assessment Structure	■ Increases content coverage without increasing test size, potentially providing more useful diagnostic information ■ Involves teachers more deeply in state assessment system ■ Monitors student progress more frequently ■ Provides for practice in test taking skills ■ More timely results for diagnosis and remediation	■ Not all standards would be covered on end-of-year statewide test, though all major categories of standards would be covered ■ Students who are substantially below grade level, or substantially above, may not be studying content in the grade-level standards. ■ Less consistency in assessment of standards that are measured under non-standardized conditions ■ Requires extensive staff development ■ To combine scores across assessment tiers, requires complex data integration for which some schools do not have the technology ■ Requires either computer-administration or complex performance maintenance systems ■ Conditions for administration are less standardized ■ Tests are less secure for high stakes

TIME PERIOD 2: SHORT-TERM OR MID-TERM

Innovation	Description	Category	Benefits	Challenges
OTHER INNOVATIONS				
2.05: Power Objectives	Higher-order objectives that subsume lower-order or enabling objectives. In order for a student to be proficient in a power objective, he or she must be proficient in all of the subordinate objectives.	Assessment Structure	■ Decreases test size by measuring only higher order objectives ■ Supports multi-tiered assessment systems ■ Maintains highest standards for statewide assessment	■ Requires supplementary classroom assessment system, which adds cost ■ Requires identification of power objectives ■ Requires complex data integration for which some schools do not have the technology.
2.06: Staff Development in Assessment Development and Use for Teachers, especially online.	Exemplary programs that train teachers to develop, use, and understand assessments using e-learning techniques, either alone or in concert with instructor-led training.	New Products	■ Better trained teachers will use the assessments effectively for student learning ■ E-learning provides flexibility of time and place ■ Lower cost than traditional methods ■ High initial cost of development	■ Teachers need access to the computers and Internet for online component ■ Some teachers (like some students) would not learn well online
2.07: Consistent Criterion-referenced Interpretive Methodology for Classroom Assessment	A uniform, understandable way of grading and reporting on informal assessments that is coordinated with the state assessment system.	New Products	■ Allows better understanding by stakeholders ■ Provides consistency between classroom and state assessment interpretation	■ Requires agreement among professionals on a system.
2.08: Teacher Certification in Assessment Knowledge	A portable certification of teacher competence in the development, use and interpretation of assessments.	New Products	■ Teachers would know what they know and do not know ■ Administrators could plan for staff development more effectively ■ Portability of certification would help teachers if they move	■ Teachers' unions might object to certification requirements if they are imposed on current teachers (rather than limited to new teachers); limiting to new teachers would reduce impact

TIME PERIOD 2: SHORT-TERM OR MID-TERM

Innovation	Description	Category	Benefits	Challenges
OTHER INNOVATIONS				
2.09: Individually-administered Clinical Tests to Measure State Academic Standards	Clinical instruments coordinated with the state assessment program objectives and information system that can be administered to students who are unable to take the standardized group-administered state assessment.	New Products	• Allows makeup tests and more valid tests of special populations • Results can be integrated with the state assessment program • Clinically more appropriate than administering the group-assessment to a single individual	• Individual administration is costly • Requires highly trained administrators and pull-out from the classroom • Possible measurement issues with comparability of scores between group- and individually administered tests, even of same content.
2.10: Improved Early Childhood Tests	Tests designed for administration in Pre-Kindergarten through Grade 2 that provide valid and reliable measurement of emerging skills and readiness without the disadvantages that many current forms of standardized tests have at those grades.	New Products	■ Allow measurement of important skills in the early grades that inform instruction and program evaluation before the compulsory testing at grades 3-8 ■ Better prepare young students for testing at the required grades	■ Controversial area of testing ■ Fear that forcing young children into standardized methods is inappropriate

Discussion of Innovations

According to State Needs

TYPES OF NEEDS

The needs of the states fall into six general areas: test content, test technical characteristics, logistics of testing programs, interpretation and dissemination of results, application of results, and the business of running testing programs.

Test Content needs deal with what the tests measure and how it measured.

Technical needs deal with the tests as measurement instruments: the quality of the data that tests yield and the research that backs up the integrity of the test scores.

Logistics needs deal with test development, dissemination, administration, scoring and reporting.

Interpretation needs deal with the understanding of the measurement provided by the tests by the audiences who need to understand the results.

Application needs deal with turning the test results into improved education.

Business needs deal with choosing vendors and partners, financing test programs and getting value for the public.

TEST CONTENT

Content validity is often cited as the most critical quality measure for an assessment program. Most states go through painstaking processes to define what the schools should be teaching, and want tests that validly measure the state academic standards. This is an expensive, time consuming process that often involves significant debate.

Most state testing programs in the past ten years have been standards-based testing programs. The new ESEA authorization (P.L. 107-110) mandates that states implement annual reading and math assessments for grades 3-8, which is the key principle of President Bush's ambitious plan for holding states and local school districts using federal funds accountable for improving students' academic achievement. This will be followed by other mandated assessment grades and subjects. Although states can select and design assessments of their choosing, they must be aligned with state academic standards. Many states test in additional subjects and/or additional grades.

Historically, some states adopted nationally-normed achievement tests from publishers who did rigorous analyses of curriculums and gleaned from their analyses a common core curriculum on which students could be compared, regardless of the state in which the student went to school. Since those national tests from the publishers measured only a core curriculum that was common across the nation, many schools and states supplemented the publishers' tests with criterion-referenced tests of their own that would either be integrated with the norm-referenced test or administered separately. The shift towards

standards-based assessment accompanied the movement toward high-stakes tests of content that seek to ensure student mastery of state learning goals.

In general, states need to make sure that the content of their tests is appropriate and that they have coverage of the state academic standards, allow for comparability from year to year, be of objective knowledge, and avoid assessment of personal family beliefs and attitudes.

There is a set of needs relative to improving the development of content standards that are not addressed in this document because, in most cases, they are preliminary to the involvement of the testing industry. For the purposes of this discussion it is assumed that the state academic content standards have already been developed, and the issue is representation and coverage of those standards in the state assessment program.

Need: Better coverage of standards without adding testing time

The current state of affairs with regard to content coverage in state assessments is that most states struggle with appropriate emphasis, comprehensive coverage, and appropriate types of items that are fair to all students. States must constantly weigh content coverage against cost and testing time, which combine to constitute "test size." For example, if the state academic standards are detailed, providing greater guidance for teachers and curriculum developers, coverage must focus on critical skills, sampling, or a combination of the two.

Numerous methods exist for balancing content coverage with test size. One method is matrix testing, in which each student is presented with a subset of the state content. This method was used in the past by states that emphasized group scores over individual performance scores. Matrix testing fell into disfavor some time ago, and would certainly be difficult to implement appropriately under NCLB, which requires individual student information. Various innovations have been implemented over the past twenty years to blend matrix testing with student census testing, but none have proven practical in use.

Another method for increasing content coverage without increasing test size is to share the burden of testing all of the academic standards between multiple tiers of assessment from high stakes to low-stakes *multi-tiered continuous assessment systems*[1]. For example, the high-stakes accountability test could focus on a smaller set of academic standards while less formal lower-stakes tests could be added into the assessment program via coordinated classroom assessments handled by teachers.

Continuous multi-tiered assessment could be delivered on a teacher-selected schedule allowing the teacher to choose the order in which instructional components are delivered, but still requiring that all of the continuous assessment components be used within the year. Continuous assessment could reduce the amount of time needed for the state end-of-year tests, since all elements of the standards would be tested in short tests during the year, thus allowing the state test to be shorter by only selecting the core elements for final accountability. Each year, states could choose different core elements for the high stakes test, eliminating the temptation for teachers to only teach the core elements, since continuous assessments would cover all of the standards and be recorded for diagnos-

[1]Innovation 2.04: *Multitiered Continuous Assessment Systems:* Integrated sets of assessments consisting of a standardized, statewide, every-student testing program for high stakes and mid-stakes and/or low-stakes assessments of the same or related standards that are administered by classroom teachers in informal programs. The results are integrated to create a comprehensive picture of student achievement. The scoring and recording of student performance would be integrated throughout the year into a database.

tic and evaluation purposes.

Continuous assessment would also allow for a systematic record of delivery of instruction and incremental achievement that can be helpful to schools and states in demonstrating the presence of opportunity to learn for all students. One caution is that, while multitiered continuous assessments may be useful in driving improvements in curriculum and instruction, their value as valid and reliable external evaluations (assessments) of student progress must be carefully examined.

A number of states have implemented these multi-tiered programs. The challenge is to determine which objectives are worthy of placement on the high-stakes test that every student in the state takes under standardized conditions. One method for doing this is the identification of *power objectives*[2]. Power objectives are summative in nature and cannot usually be accomplished by students without proficiency in subordinate objectives that are prerequisite to the power objective. For example, for a student to understand how to solve a simple story problem in mathematics that involves subtracting two numbers, the student must know how to subtract two numbers and how to solve simple story problems. Therefore, the inclusion on the task of only the higher level standard will show proficiency in both of the lower level standards as well as the ability to integrate them into a higher task. The problem with only testing power objectives is that students who fail to achieve the power objective might still have proficiency in one or more of the relevant subordinate objectives. If the test doesn't measure those lower-level objectives, that diagnostic information is not available to the educators about that student. If the information on attainment of the lower level standards is essential to some high-stakes testing purpose, it must be included on the high-stakes test. If it is not relevant to the high-stakes interpretation (e.g., accountability or adequate yearly progress evaluations) but it is relevant to the instruction of the student, that need can be filled by a classroom assessment administered by the teacher in a low-stakes setting. The issue of diagnostic assessments coordinated with the state assessment system is addressed later on in this document.

In order for this type of multi-tiered testing to meet the needs of the states, there would have to be an established system of coordinated tests that are low-stakes and high-stakes. There would also have to be an acceptable system for determining what the power objectives are and which subordinate objectives are subsumed inside of power objectives. This would require research as well as development of an appropriate assessment system and coordination between the information systems of the high-stakes and low-stakes tests.

Need: More customization of items and subtests

Another need is for more customization of test content. This need arises for those states interested in keeping costs low by avoiding the development of custom tests from scratch. (For the purpose of this document, *custom assessments* are defined as new test built entirely from scratch for a specific purpose, such as a state assessment of academic standards; customized assessments are defined as existing tests that are modified to make them appropriate for a specific purpose.)

One innovation would be to have a library of test content from which to choose items and assemble state tests

[2] Innovation 2.05: *Power Objectives:* Higher-order objectives that subsume lower-order or enabling objectives. In order for a student to be proficient in a power objective, he or she must be proficient in all of the subordinate objectives.

of academic standards, or a c*ommon item bank*[3]. Rather than publishers providing an intact test form, they would provide small test-lets or even individual items. This would allow states to administer targeted custom tests that measure their state content standards at a lower of cost with acceptable research backing up the tests.

A common item bank would save the states considerably in the cost of developing tests, but so would using the same the test. The issue is to examine the trade-offs between independent development of standards and assessments by the states and the high cost of developing over 50 complete, multi-subject, mutigrade assessment systems within the limited resources of the assessment profession. To the extent that states are willing to agree that some or many of their standards are essentially the same, and are willing to agree to similar test item types to measure those standards, a common item bank can be very helpful.

Test items must be constructed to meet strict test specifications and item specifications. Items are not necessarily interchangeable since an item written in one state may not meet the specifications desired in another. In order for an item bank to be credible and helpful to the states, they must contain information on where the items originated, what objectives they measure, what specifications they match, and the statistical performance of the items.

Modern computerized item banking systems make up for many of the shortfalls of past solutions. They store graphic images and allow the development of print quality copies of the items and the graphics, thus making it possible to assemble useable test forms.

Another innovation would be to start with an existing test, such as a published norm-referenced test, and customize it to provide accurate and comprehensive measurement of state academic standards[4]. This need arises for those states interested in phasing in a custom state assessment while starting with an off-the-shelf test as well as those states interested in saving funds by avoiding ever having to develop full custom tests. One disadvantage of this approach is that it prevents states from releasing the off-the-shelf test items, which are owned by the test developer, substantially increasing test security risks.

When a state sets out to customize an existing test, publishers often put some restrictions on the extent of customization they will permit in order to maintain the integrity of the test's technical data and norms. In one method, the publisher would develop a short form, or abbreviated version, of the standardized test. The short form would be the publisher's best effort at creating the briefest test to which it can attach national performance data. The state would then have a testing company create supplemental forms that would add content coverage to the test. Items from the abbreviated standardized test would be added to items in the supplemental form to create reliable measurement of a particular academic standard. This method has the advantage of high quality on most of the test and the provision of national performance data, which many states desire. It has the disadvantage of not providing states with the exact type of content

[3] Innovation 2.01: *Common Item Banks:* Sets of calibrated and researched items and test elements tied to a massive set that is the combination of all state academic standards. These items and test elements would be kept in a computerized content management system for efficient and secure access by qualified individuals who could assemble tests as needed. Such common items could be released annually or kept secure for reuse.

[4] Innovation 1.07: *Customized Off-the-shelf Tests:* Publishers' norm-referenced or criterion-referenced tests that are supplemented with sets of items to round out the measurement of state academic standards.

coverage and item types that they might want in a custom test. The technical quality of the norm-referenced scores are also dependent upon the degree to which the final test approximates the version that was standardized intact(the original form). There are other models for customization as well, each with advantages and disadvantages.

Need: Multiple measures to assess standards

A third need with regard to test content is the need for more *multiple measures*[5] that are comparable and fair. One of the criticisms of current state testing programs is that they do not allow for sufficient coverage of difficult-to-assess standards. For example, a language arts standard may call on students to develop effective debating or presentation skills. Such a standard is generally impractical to assess in a large-scale assessment environment, but needs to be assessed by teachers.

Another often-cited purpose for multiple measures is to allow for student variations in learning style and test-taking style across students. For example, some students claim to do better on open-ended questions like essays than on multiple-choice questions, while other students feel the opposite is true. There is little evidence, however, to support the notion that students without disabilities who posses the underlying skills can have difficulty demonstrating them on certain item types. Usually, such students lack at least some of the targeted skills; it is thus a policy question for states to determine which skills all students should be able to demonstrate, if any.

Some students who have disabilities, however, might be better tested via different items types and formats. P.L. 107-110 sets aside funds for the creation of "enhanced assessments," which would allow, among other things, the creation of multiple measures of student academic achievement from different sources. This raises the issue of accommodations in test administration, which is addressed below under *Logistics.*

Multiple measures can create problems with regard to consistency and fairness, and also can add considerable cost to a testing program. To the extent that the testing industry can provide an innovation that would permit multiple measures of state content, and still deal with the issues of fairness, it would be helpful to states.

Need: Better instruments for early childhood assessment

Testing in the early grades (Pre-K through Grade 2) has been controversial for years. While most educators and parents see the need for early detection of a child's strengths and weaknesses as well as evaluation of early grade education programs and pedagogy, the use of standardized tests for very young children is largely frowned upon. Rather than get involved in the debate about whether current testing instrumentation is appropriate for young students, it is important to acknowledge that there is room for improvement. An innovation that would be most welcome would be improved early school test instruments that are developmentally appropriate for the 4 to 7 year olds in these grades but also provide the kinds of standardized, valuable data that is obtainable on older students[6]. For example, computerized testing might allow the introduction of multimedia and more game-

[5] Innovation 2.03: *Coordinated Multiple Measures:* Measures of academic standards using a different measurement methodology, e.g., performance assessments or technology assessments, but coordinated with the primary assessment for fairness and comparability.

[6] Innovation 2.10: I*mproved Early Childhood Tests:* Tests designed for administration in Pre-Kindergarten through Grade 2 that provide valid and reliable measurement of emerging skills and readiness without the disadvantages that many current forms of standardized tests have at these grades.

like experiences for the young students that can still provide excellent standardized test results.

TECHNICAL

The second general category of needs for innovation deals with the technical quality and nature of the information provided by the test. The main purpose of assessment programs is to provide useful information and the extent to which tests provide scores that lead to instructional improvement enhances their utility for states. Technical needs can be divided into two subcategories: types of scores and technical quality.

Types of Scores

There are two general types of scores provided by tests: criterion-referenced and norm-referenced. The debate has raged for years in the state assessment community about the relative merits of these two approaches. Some feel that norm-referenced scores should be eliminated completely from state assessments and all interpretation should be criterion-referenced. Others feel that norm-referenced information is not only beneficial to public understanding of test results but also allows for interpretive schemas that are not possible with criterion-referenced scores alone. There is a clear need to get past the rhetoric of this debate and provide innovations in developing and using the appropriate test score types. Many do not realize that criterion-referenced scores can be obtained from norm-referenced tests and vice versa. While the test types are specifically structured to maximize the quality of one type of data or the other, many states have designed systems that can obtain the quality of data they need from hybrid examinations.

Need: Acceptable, understandable ways of measuring mastery at individual and group levels

Criterion-referenced scores are necessary for an assessment program to be in compliance with P.L. 107-110.

There are many types of scoring schemes used across the states that have been implementing standards based assessments. The variety of methods for providing mastery and proficiency scores for individual students and groups causes much confusion among the consumers of testing information. Teachers, school administrators, state policy makers, parents and the public-at-large all have a need to understand clearly how well schools are doing in educating students and how well students are doing in achieving proficiency or mastery of state academic standards. A useful innovation would be a *common system of criterion-referenced information*[7] that was acceptable and useful in meeting the general needs of test consumers.

In the past, many states have adopted the NAEP scoring scheme of labeling performance as basic, proficient, and advanced, and P.L. 107-110 specifically requires at least these labels, although states are free to add other levels of performance. The methods for determining these proficiency levels, however, have varied considerably from state to state, which makes it confusing for the public. To add even more confusion to interpretation, publishers have attempted to tie their norm-referenced tests to these NAEP proficiencies. When states have tried to reconcile differences between performance on publishers' tests, state custom assessments, and NAEP itself, the analyses have raised more questions than they have answered. The states have a need for a consistent criterion-referenced interpretive scheme to be created by the testing industry, certified as technically appropriate by the psychometric scientific community, and made available to states for all of their assessment programs.

[7] Innovation 1.03: *Collaborative Resource on State Assessment*

Not only does this criterion-referenced interpretive scheme have to be available for the major state assessment programs, it should also be made available for classroom assessments. At the least, a consistent interpretive methodology should be established for classroom assessments[8] so that teachers, students, and parents get used to understanding the interpretive scheme. Classroom teachers do not get very much, if any, instruction on classroom assessment when they are in their education college programs. Good in-service training that will improve classroom assessments to make them more academically sound would reduce teachers' current use of what is known as "hodgepodge grading." This is addressed later in this document.

In so many areas of life there is a common language to talk about measurement. For example, there is a common understanding of miles per gallon, degrees Fahrenheit, height in inches, the Dow Jones Industrial Average, and wind chill factors. Most people also generally understand the meaning of some norm-referenced scores, such as percentile ranks. Most parents understand the letter grade system of A, B, C, D and F, and they understand percentage correct, passing scores and honor rolls (notwithstanding the fact that the letter grades might not be comparable across teachers). The problem with the basic/proficient/advanced scheme is that it is inconsistently applied across states, and still debated among experts as to how it should be interpreted.

It is important to point out that it is not possible to accomplish absolute equating of educational standards across different tests in different states. Even if all states agreed to use a common set of labels, there would be inconsistency behind the meaning of the labels in operational terms. Unless the states agreed to the definitions of content for the various tests, the use of common definitions would be meaningless. If one state has a math test that is totally problem solving and another has a test that is totally computation, the terms "proficient" and "basic" have different meanings, even with the same operational definition.

Need: Ways to provide the public with comparisons and other benefits of norm-reference tests while still using criterion-referenced tests

Norm-referenced scores traditionally have aimed to provide information on student and group performance relative to that of other students and groups. While norm-referenced data in isolation is often not as informative as is needed for educational improvement and accountability, and would not satisfy the requirements of federal law, the public feels comfortable interpreting norm-referenced comparisons and will often create comparison interpretations informally when only criterion-referenced scores are provided. The perception that all aspects of traditional norm-referenced tests are antithetical to state standards-based assessment programs has evolved because of a narrow interpretation of the utility of norm-referenced data.

The Tennessee Value-Added Assessment System (TVAAAS), for example, which has been hailed as an innovative way to interpret student progress, relies on a norm-referenced test to compare student progress from year to year. This system has also been found to be useful for measuring the effectiveness of teachers and of schools and a number of states are considering whether to implement such a value-added model. Yet, unlike norm-refer-

[8] Innovation 2.07: *Consistent Criterion-referenced Interpretive Methodology for Classroom Assessments:* A uniform, understandable way of grading and reporting on informal assessments that is coordinated with the state assessment system.

enced tests, many standards-based tests lack an important design feature—"vertical equating"—that allows for their use in value-added analyses. It is important to note that it is not necessary for a test to be norm-referenced to accomplish the goals of TVAAS or other value-added systems. Theses systems do require regular testing and vertical scaling across the grades, which is routinely done for norm-referenced achievement batteries, but can also be done for custom state assessments.

A needed innovation is for the testing industry to develop methods for applying the best features of norm-referenced tests while avoiding misinterpretations[9].

Technical Quality

As the stakes for testing become higher, tests will be subjected to increased scrutiny by the public, the education community, the scientific community, and the courts. As a result, the need for proof of technical quality in the test will continue to increase. In our litigious society, it is likely that any testing program will be challenged by a lawsuit from a school, a parent, a special-interest group, or student. States have every right to expect that the vendors they choose to work with for testing programs provide the highest technical quality of assessment program that current technology and science permits, and the law now requires it.

The three major components of technical quality are reliability, validity, and fairness.

Reliability

Reliability is the extent to which assessment provides consistent information over occasions and forms. When states commonly used publishers' norm-referenced assessments, they did not need to be concerned about this because the reliability of published standardized tests that were bought off the shelf was consistently high. A publisher creating only one multilevel testing instrument can take great pains to develop the test carefully over a period of years and thoroughly research the instrument. The significant investment publishers made in each of these off-the-shelf tests (often exceeding $20 million) was deemed worthwhile because millions of copies of the tests would be sold over many states and multiple years. Reliability presents particular challenges, however, with custom built tests that are different for every state. The timelines for test development are extremely short, and the opportunity to recover the research and development costs over many years and millions of administrations are much more limited, especially in smaller states.

It is also the case that many custom state assessments use innovative item types that have not been subjected to the years of research that more traditional item types have. This also creates reliability challenges for the test developer.

The third challenge to the reliability of custom developed state assessments stems from their high-stakes nature, which puts additional pressure on security. One of the typical ways to deal with security threats is to use a new form of the test every year. The more forms of a test are developed, the more the issue of inter-form reliability becomes problematic. It is widely believed that drift occurs from form to form and year to year. This threatens the meaningfulness of the test results as well as the reliability.

[9] Innovation 1.10: *Integrate Important Benefits of Norm-referenced Tests for State Standards-based Programs:* Data using comparisons of performance and design features such as vertical equating that add to the informative power of state assessments without replacing or compromising the criterion-referenced interpretation system.

Need: Internal consistency and alternate-forms reliability so that results are credible

A priority for the states is the need for internal consistency and consistency across forms so that results are credible and so that good educational decisions are made on the basis of the results, and not on the basis of aberrations in performance that result from lack of reliability of hastily developed instruments. Innovations that lead to increased reliability without sacrificing security and customizability are needed by the states. It is unlikely that the nature of these innovations will be technical, since most technical improvements in psychometrics that would yield increased reliability have already been made. It is more likely that procedural and the technological innovations can help solve this problem. For example, the use of randomized forms[10] could allow for more security through scrambling of items across forms. Unfortunately, this is difficult to accomplish cost-effectively with printed tests. The use of randomized forms is already widespread in state assessment systems, because this is a convenient way to field test multiple new items. However, this use of randomized forms is accompanied by some significant technical challenges. It may or may not be a cost effective way of improving test security as that depends on the extent to which the test forms are different. Furthermore, research has shown that changing item placement in a test can affect student performance, thus possibly creating inequalities across the various test forms.

Also, the innovation in content of having common pools of items available across states (Innovation 2.01) could provide improved reliability without sacrificing the standards-based nature of most state assessments.

Validity

Validity is the extent to which the test measures what it purports to measure. Assuming that the content objectives of the test are well defined and established, it is important that the test items actually measure that content. This places restrictions on the types of items that may be used. For example, measurement of complex higher-order skills, while desirable and included in many states' academic standards, must be done very carefully because items used to develop these higher order skills can often be measuring multiple things, some of which are not the state standard. In addition, P.L. 107-110 specifically prohibits the evaluation and assessment of personal or family beliefs and attitudes.

Need: Legal defensibility for high stakes

One of the needs that states have is for greater legal defensibility of high-stakes tests. This goes to the validity question very directly. If the test is valid for this use, it should theoretically be appropriate for high-stakes measurement of academic standards. Right now, states are often left to their own resources to defend the validity of their testing programs and benefit little from each other's similar situations. One innovation desirable to the states would be the creation of a legal entity or agency specifically devoted to the legal defensibility of high-stakes assessments and a body of knowledge and case law that can be shared across states[11]. Currently this exists only informally.

[10] Innovation 1.08: *Randomized Forms:* Tests in many alternate forms administered across students so that students in the same class receive different forms. While this is best done in computer administration, it can be accomplished to some degree in paper and pencil administrations.

[11] Innovation 1.03:*Collaborative State Resource for Assessment Systems:* An organization serving as a resource on the legal defensibility of state assessments, possible timelines for testing programs and schedules, and state/vendor contractual techniques.

Fairness

The third major component of technical quality is fairness. Fairness is a broad category but includes at a minimum the absence of bias in a test and the accessibility of the test to all students. P.L. 107-110 specifically requires that tests be used for all students, including those who are traditionally difficult to test with standardized procedures, specifically mentioning economically disadvantaged students, students with disabilities, students with limited English proficiency, and migrant students.

Need: Proper evidence of bias-free tests

One of the greatest needs that states have is proper evidence that their assessments are free of bias. Because there are so many subgroups to be dealt with and complex analyses of test data, the bias studies must be carefully aimed at the right populations and subpopulations. This might not seem like a legitimate need, since bias study procedures are well established and documented in the technical literature. However, there is a remarkable inconsistency from state to state on the degree to which these studies are actually conducted, and they often cost more than the legislature has appropriated. While most state test development efforts deal with straightforward ideas of gender bias and ethnic bias, they often skip the elaborate studies of subgroup bias that would be necessary to prove fairness to all students. Some of these difficulties are cost and procedural, and some are technical. An innovation that is needed by the states is to have an agreed-upon set of procedures for bias studies[12] that are used by all states and that are clearly known to legislatures and lawmakers so that they can appropriate the necessary funds to do these studies. The extent to which the national item bank innovation (Innovation 2.01) is achieved will help meet this need if the item banks contain sufficient bias information about the items.

P.L. 107-110 requires the results in each state, LEA and school to be disaggregated by gender, each major racial and ethnic group, English proficiency, migrant status, students with disabilities and by economically disadvantaged students as compared to students who are not economically disadvantaged. Furthermore, P.L. 107-110 requires "evidence from the test publisher or other relevant sources that the assessment used is of adequate technical quality for each purpose for which the assessment is used [§1111(b)(3)(C)(iv)]," which evidence is consistent with nationally recognized technical and professional standards [§1111(b)(3)(C)(iii)]. The Joint Professional Standards require that the bias analyses done for any group or subgroup give evidence of the suitability of the test for each of its intended purposes. This puts a significant burden on states to meet the requirements of the law, and few states have the budgetary resources to fund this research properly. The innovation of creating standard procedures for bias studies would help if there were clear costs associated with the appropriate studies.

LOGISTICS

The next major set of needs deals with the logistics of the testing process. This includes test development, distribution, administration and scoring.

Test Development

Test development represents a high cost for states for two reasons: First of all, the resources to develop high quality assessments that are suitable and

[12] Innovation 1.14: *Standard Procedures for Bias Studies:* Agreed-upon methodologies for bias studies that are known and funded by law-making bodies.

defensible for high-stakes use are limited. Secondly, the increase in testing in the past few years, which will only get more intense with the passage of P.L. 107-110, has increased the demand for testing. Simple economics say that increasing the demand on a limited supply drives up prices. In the past, the term "standardized test" applied to five or six nationally-normed off-the-shelf instruments. The testing industry evolved to develop instruments like these every four to seven years, with suitable research. The addition of fifty state assessment systems, some of which are close in size to the nationally-normed tests, increases the demand for test development by a factor of ten or more. When consideration is given to the frequency with which the state assessments are expected to be revised, and the commensurate need for classroom-based systems that support and complement the state assessments, it is easy to see the potential for overtaxing the industry's test development capacity.

Need: Ways to develop more and better tests more quickly

States need innovations from the testing industry that increase the quality of state tests without increasing the cost to more than what the taxpayers can bear. They also need tests developed more quickly without compromising quality. The faster tests can be developed (i.e., items written and assembled), the more time there is for research that improves and maintains the psychometric quality of the assessments. It is also the case that some state education agencies are challenged by their legislatures to implement testing programs very quickly. Lawmakers sometimes do not allow enough time for states and the testing companies to do the best job.

There have been various proposals to meet this need. One is that states share items informally in a massive item pool. Rather than writing and buying new items the states can use common items that are also used on other states' assessments. The challenges to this are security and the appropriateness of items from one state to another when the state academic standards differ. Another challenge to this is the desire some states have to be different from others, and to have complete ownership of their testing systems.

The innovation that would best address this need is a secure common item bank (Innovation 2.01), maintained in a computerized system, accessible only by appropriate personnel with passwords. Items would either be released annually or would be maintained and updated with technical data from administrations in the various states and would be tagged with information about their appropriate use. In the case of a pool that is not released, states could pay the owners of the items a negotiated royalty dependent upon the use of the item in that state so that the item bank could be maintained and improved over time.

There are many obstacles to overcome in implementing this innovation. First of all, the issue would have to be addressed as to who maintains the item bank. Sophisticated computerized test item banking systems have been developed and are on the market, so the technology is only a small challenge. The biggest challenge is logistical. Would all the testing companies share access to the same item bank? Would item banks be proprietary to testing companies and only the states that worked with that testing company have access? Would some trusted entity create and maintain the bank and testing companies and states would have their owned items placed in the bank voluntarily and receive payment when their items were used (like the music industry is attempting to do with online music purchasing)? Should companies or agencies create item banks and share them by making busi-

ness deals? However this innovation occurs, it is a priority for the states to have faster development of high-quality, properly researched tests that provide multiple forms at an affordable cost.

Innovation 1.07: *Customized Off-the-Shelf Tests* can also go a long way to reducing the time and cost of test development. The challenges to overcome in implementing this innovation were discussed earlier.

Test Distribution

Innovations in test distribution would allow tests to be presented to students in a more timely fashion with the appropriate level of security without creating logistical challenges for teachers and school personnel.

Need: On-time test delivery

In recent years, some tests have been delivered to schools for administration later than scheduled, causing problems in the timeliness of testing and reporting. In some cases, the late delivery of printed test materials have resulted in major disruptions of the school calendar, issues with comparability of data to past and future years when students are tested at a different point in the instructional process, and resultant late scores reports, sometimes causing schools the additional expense of mailing reports home to parents after the school year has ended and problems with setting up registrations for summer school. The testing industry needs to find a way of more consistently delivering tests on time. The challenge relates to the test development and research issues discussed above. If the process for test development takes longer than the budgeted time, it places a burden on the production and printing of the test.

Part of the challenge remains with the testing industry, which must be able to meet deadlines to which it has agreed. But part of the solution also has to do with the states establishing schedules that can reasonably be accomplished. Since there has been so much experience in developing tests over the years, it should be possible to identify model project timelines so that states can plan better and can inform their lawmakers and policymakers of the minimum and optimum time to allot to test development and distribution. A useful innovation for the states would be a single source to access that would provide suggested schedules for testing programs.[13]

This is another area in which technology innovations can help. *Computer-administered tests*[14] do not need to be printed and shipped, thus saving a considerable amount of time. Computerization of test development through online item banking (Innovation 2.01) also can improve test development efficiency. The combination of these innovations can increase the likelihood of on-time presentation of the tests to the students.

Of all the innovations presented in this paper, computer-administered tests is the most powerful and most challenging. While it provides very many benefits, it provides significant challenges as well. First of all, the infrastructural demands of giving widely administered tests on a computer are considerable. If the tests require an Internet connection, it must be functional and fast enough to not distract from the testing setting. If the tests depend on a server-based local area

[13] Innovation 1.03: *Collaborative State Resource for Assessment Systems:* An organization serving as a resource on the legal defensibility of state assessments, possible timelines for testing program schedules, and state/vendor contractual techniques. This organization would be available to states and testing companies to assist them in the implementation of their programs.

[14] Innovation 2.02: *Computer-administered Tests:* State assessments administered to every student on computers without the need for printed tests.

network, it must be up and functional for all students. There are opportunity to learn challenges that would make it hard to justify equity in the scores from students who have frequent access to computer at home to students who do not. There are issues about the equating of scores between computer administrations and paper and pencil administrations if some students take the test in different modalities from others.

Despite these challenges, computerized testing offers enormous benefits, including instant turnaround time of scoring, lively multimedia questions, randomized forms, capturing of every keystrokes that allow for richer scoring algorithms, and ability to administer adaptive tests that provide more accurate scores with shorter testing times.

Computerized tests exist today, but the challenges to their widespread application in school testing are significant enough that they will probably not realize widespread use in high-stakes K-12 testing for many years. In the interim, they can be effectively used for makeup tests of students who miss the main testing administration window, of for retests on required exams for students who failed the test on an earlier paper and pencil administration and can retake the test at their convenience. In any case, continued research into this innovation is warranted because of the many benefits it can provide when the technological challenges are solved.

Need: Better security

Security breaches in the distribution of tests create many problems for the states. First of all they compromise the integrity of the results, undermining the entire assessment program. If the state officials do not know who has seen the test or portions of the test before administration, there is no way to make fair comparisons. A known security breach can also compromise a form of the test, taking it out of use, not only for that state, but perhaps for other states if they share items and tests.

One of the most effective measures to maintain security is to release the test items every year, though the cost of developing new items each year is substantial. Maintaining the security of test results can be expensive and logistically complex. For example, some state have their testing companies label every test booklet and packet, and log which school and classroom received those tests. Any tests that are not returned are noted and the security breach is investigated. Another security measure that adds cost is sealing of booklets in sections. The seal is broken by the student just before a section is administered, increasing the likelihood that no one can see the test content before administration. While these procedures have worked in many states, they are expensive and time consuming.

Nothing can completely protect the security of tests, but it can be improved. Innovations that would improve security are those that would decrease the opportunity to see the test before administration, and those that would increase the pool of items from which a form presented to a particular student could be drawn.

Computerized-testing with randomized forms, a combination of Innovations 1.08 and 2.02, is actually a two-edged sword with regard to security. On the one hand, it can be used to customize tests at the time of administration from an item bank, thus improving security. On the other hand, computerized testing can actually exacerbate the security risk if tests are administered over a longer period of time. Students' memory of a few test items shared with another student who is tested later in the year can be just enough to invalidate a test. This is certainly an area that merits investigation for procedural improvement through innovation.

Test security breaches can be unintentional or intentional. The most notorious cases of breaches are where the intent is to see test content in advance in order to unethically improve student scores. However, many test security breaches are actually unintentional and are the result of poor training and/or information provided to teachers and administrators as to what constitute appropriate test security measures. This issue is addressed in the section below on professional development needs.

Need: Logistical simplification for teachers and school personnel

Testing has always been a logistical burden for educators and the increased emphasis on tests has increased that burden. Not only does test administration take time, but the logistics of scheduling, distributing tests, collecting and preparing them for return for scoring, and the distribution of reports to students and parents are time consuming activities that do not add anything to the learning process. Sparing educators these administrative and logistical chores would be most welcome and would add to instructional time and school instructional management.

One innovation that would help in this regard is computer-administered tests (Innovation 2.02), to the extent that problems with technology and equipment do not make it even more difficult for teachers and administrators. Another innovation would be computerized or web-based reporting[15], which might decrease the amount of time that reports are handled and filed.

Test Administration

Once the tests are distributed, they must be administered. There are a number of needs of the states relative to test administration that could be better met by the testing industry, such as ways to handle absences, special situations and make-up tests, and ways to eliminate test taking skills as a factor in test scores.

Need: Better ways to handle absences and special situations

It is a fact of life that some students are absent on the day of standardized testing. These students need make-up tests administered, which creates logistical challenges and security challenges. The innovation of computer-administered tests (Innovation 2.02) can be ideal for this, provided that there can be assurance that the scores from a computer-administered version of the test have the same meaning as the scores from the paper and pencil version of the test. This will require research and development of appropriate tests that can be administered by computer, as well as network of computers that are capable of administering tests under standardized conditions.

Need: Practice tests and methods to eliminate test taking skills as a factor

One of the concerns that states have is the extent to which test-taking skills are differentially distributed in the student population, resulting in some students performing less well than they should. Theoretically, the effect of test taking skills should be unidirectional, that is, obtained scores are below true scores for students with poor test-taking skills, but obtained scores are not increased above true scores with students with good test taking skills. Thus, all students should be proficient in taking tests so as to make valid scores and comparisons. States need high-quality test preparation programs that do not artificially inflate scores by

[15] Innovation 1.01: *Computerized or Web-based Reporting:* A paperless system of providing real-time accessibility to reports either on a school computer network or on the Internet. Ideally these reports would be interactive, allowing cross-referencing and detailed investigation and explanation of scores for multiple types of audiences.

teaching an overly targeted curriculum, but bring all students to the same level of proficiency in test taking. An innovation would be an integrated program that improves test taking skills while also providing legitimate instruction, not just test prep[16].

Innovation 2.04, Multi-tiered Continuous Assessment Systems can also meet this need as it involves becoming familiar with test formats and structure in the context of instruction, as it should be.

Need: Special forms of tests for students with disabilities

P.L. 107-110 specifically requires that the state assessments provide for participation by all students with reasonable accommodations for students with disabilities necessary to measure their achievement relative to state content standards [§1111(b)(3)(C)(ix)].

Students who cannot take the test under standardized conditions need to be able to take an appropriate form of the test. Often the group-administered instruments designed for the general population cannot easily be adapted to students with disabilities while maintaining full validity. States need an innovation that would provide tests in different formats, such as individually administered or oral tests[17] that measure the same skills as the general test and need the research that proves this to the satisfaction of the stakeholders in the test. These individually administered tests would be clinical instruments specifically designed for their target population of students. This technique is viewed by school psychologists as more appropriate than simply providing an individual administration of the same test taken as a group-administered test by the general population of students.

The issue of accommodations has been a difficult one for the states and the testing industry for some time. On the one hand, there is the desire to provide every child the benefit of the state assessment system, and all parents the information they need on their children. On the other hand, efficient testing programs usually present assessment situations that some children cannot deal with and obtain valid scores. For example, a visually-impaired student might need a Braille or large-print version of the test. Other children might need the test read aloud to them. Still others might need someone else to fill in their answer documents. These are called "accommodations." Testing condition alterations and modifications compensate for extraneous factors and can change the meaning of test scores. A valuable innovation would be the creation, through collaboration, of a model system of classification of accommodations, and options for appropriately incorporating the scores from modified test administrations and alternate tests into state accountability ratings[18].

Note that accommodations and modifications are usually determined on an individual student basis based on needs. Each state's laws also come into play. For example, some states award full diplomas to studentswith disabilities if the student accomplishes the IEP written for him or her. Other states label such diplomas as exceptional in some way. This area requires much more discussion and research before these problems are solvable. Some states might even require additional legislation.

[16] Innovation 1.06: *Integrated Test Readiness Programs:* Programs that simultaneously teach meaningful content around the state academic standards while improving test-taking skills.

[17] Innovation 2.09: *Individually-administered Tests to Measure State Academic Standards:* Tests coordinated with the state assessment program objectives and information system that can be administered to students who are unable to take the standardized group-administered state assessment.

[18] Innovation 1.03: *Collaborative Resource on State Assessment*

Test Scoring

The last area of test logistics is test scoring. Since the major purpose of testing is to obtain the information in the score reports, the process of changing the student's raw responses into meaningful information is crucial. Interpretation and the creation of reports are addressed in the next section. The logistical challenges in test scoring are speed and accuracy. In each case the states need more speed and more accuracy.

Need: Faster turnaround of scores

One of the greatest needs of the states is fast turnaround of scores. Even though many states put extreme pressure on their scoring vendors to turn around results of tests in as little as two weeks, this still sometimes results in scores being returned to schools after school ends. Without the scores, schools have difficulty identifying which students need summer school and other tutoring. The problem is worse when there are deadlines missed either in test delivery or in scoring turnaround. Missing documents at the scoring center, make-up testing and transportation problems create even greater delays. Educators often ask why results cannot be returned the next day.

There are two possible innovations that can help in this regard. The first is computer-administered tests (Innovation 2.02) that can be scored instantaneously. The second is a method for scanning and scoring answer sheets locally within the school or district to provide immediate feedback to the schools[19]. The documents could then be sent to the scoring vendor for additional scoring and aggregation for summary reports. In order for this to be feasible, schools would have to invest in ways to score the tests locally without compromising the answer media for the official scoring. States would also have to deal with the increased costs of scoring the tests twice and the security risks that this would create.

Sometimes the testing program design makes it harder to turn around scores quickly. For example, items that need hand-scoring, like open-ended questions and writing samples, add considerable time to the scoring process. One method that can decrease the turnaround time for hand-scoring is computerized scoring of open-ended responses[20]. There are at least four programs currently available that can score essay questions with accuracy that approaches or surpasses that of human scorers. While these programs have been used in professional examinations and college testing, the extent to which the states and their stakeholders will trust these programs for K-12 statewide assessment will depend upon the research that backs them up and their continued use in actual programs without negative consequences. (A less radical option for those interested in computerized scoring of open-ended responses is to replace only one of the two human raters typically used for each response; this may still result in financial savings, but most if not all of the increased speed is lost.) The value of computerized scoring is different for short response items and longer essays. Most of the research on computerized scoring has shown it works well for longer sample of student work, in particular for writing passages. Short responses have not been accurately scored by computer systems. Different testing programs with different purposes may differ in whether computer scoring is a benefit or a detriment.

[19] Innovation 1.09: *Local Scoring Option:* A system whereby LEAs can score tests locally for immediate results prior to sending answer media to the test scoring contractor for official statewide scoring and reporting.

[20] Innovation 1.05: *Computerized Scoring of Open-ended Responses:* A method by which open-ended questions are scored entirely or partially by computer with little or no human scoring.

Another method that can decrease turnaround and also assist with reducing cost is distributed scoring[21]. In this method a student's open-ended responses are captured on a computer, either through use of an image scanner and software that isolates the image of the written response, or through computer-based test administration (Innovation 2.02). Once the student's response is isolated or captured in the computer, it is transmitted over the Internet to the scorer who has been trained to score that specific response. That scorer has access on the computer to the rubrics and sample responses. The scorer reads the response, enters a score on the computer and transmits it back to the computer at the scoring center. The program in the computer at the scoring center waits until the second reader's rating is submitted, reconciles the scores and either accepts the score (if it is within the tolerance set by the program) or sends the student's response to someone else for a resolution reading and scoring. The cost advantage of this method comes from the fact that the scorers often can use their own PCs at home or at work, thus saving the cost of renting or owning a scoring center facility that seats hundreds of hand-scorers. The responses are viewed on a computer screen which saves the cost of making multiple copies of the papers. The ability to work on a flexible schedule from home allows a greater variety of qualified individuals to do the scoring who otherwise would not be willing to do so, since they might have other jobs or responsibilities. It also allows the scorers to live anywhere, not just in the city where the scoring center is located. This flexibility provides two advantages: higher quality scorers and the ability to pay a lower rate (since a person is often willing to accept a lower rate to avoid leaving home). The scorers still need to be trained and that would often mean coming to a central location for at least the training, although net-meeting technology would also allow that to be done over the Internet. This technique exists today, but is still used in very few instances. Its wider acceptance would be a significant innovation.

Certain score types, like state norms or comparisons, cannot be generated for reports until a minimum number of students' tests have been scored. P.L. 107-110 requires student score reports to show how students served by the local educational agency perform on the statewide assessment compared to students in the state as a whole [§1111(h)(2)(B)(i)(II)]. It also requires information that shows how a school's students performed on the statewide assessment compared to students in the LEA and the state as a whole [§1111(h)(2)(B)(ii)(II)]. These data for reporting cannot be generated until there is some way of obtaining the state performance data.

There are two ways of dealing with this need. The first is to use last year's data to generate state statistics for a preliminary set of score reports and then to use this year's data for a revised, later set of reports. While this decreases turnaround time for the initial reports, it creates confusion and necessitates double reporting, and might not meet the requirements of P.L. 107-110 (This has not yet been determined). Another possibility is to leave state comparisons off of the initial wave of reports and place them on a later wave of reports. Any innovation that allows state comparisons or statistics to be calculated more rapidly would be welcomed by the states.

[21] 1.12: *Distributed Scoring.* A method for scoring open-ended responses more efficiently by either (a) capturing the student's written response using imaging technology, or (b) having the student entering responses directly into a computer, and sending the responses to trained scorers "at large" who can work from their homes or offices on their own schedules. The scorers score the items on their personal computers and relay the scores, which are then reconciled with the second scorer and accepted or referred for a resolution scoring. The same quality control procedures are used as for hand scoring in scoring centers.

Some innovations previously discussed might help in this regard (e.g., innovation 2.04, Multi-tiered Continuous Assessment Systems). The use of a good continuous assessment program, tied directly to the end of year components and the state standards, could make results available for summer school placement, fall class placement, and even as an aid in teacher self-evaluation without having to wait for the end of year state assessment. In this way teachers would know right away after delivering a unit if the students "got the lesson" or they need to repeat it, instead of wondering at the end of the year when the class does not do well on a particular area. It also helps create an incentive for the teacher to actually deliver the state standards and curriculum, as well as a possible record to demonstrate that the opportunity to learn was present in each classroom. A challenge might be requiring a particular form of ongoing assessment might be viewed as overly prescriptive.

Need: Greater accuracy of scoring

Few things are more frustrating to states than errors in scoring that are discovered after reports have been distributed. In the past, this rarely happened but the increased burden on the test scoring profession in past years has created more errors than at any time in the past. Scoring errors have many sources, some of which are incorrect scoring keys or rubrics, incorrect application of tabular data (norms or proficiency cutoffs), and human error in hand scoring. States and school districts have had to issue revised reports because of errors like these and in some extreme cases, students were retained in a grade, denied graduation, or sent to summer school based on incorrect score reports.

It is incorrect to assume that all scoring inaccuracies are the fault of the companies scoring the tests. In some cases, the states themselves or the test developers (sometimes different from the test scoring contractor) signed off on incorrect rubrics or keys. Regardless of the causes, states need to work with the testing companies to improve the accuracy of scoring to as close to 100% as possible.

Logistical changes can improve accuracy of scoring, but result in creating other problems. For example, adding time to the scoring vendor's schedule for additional checks and balances goes against the need for faster turnaround. Adding more staff to the process increases costs that many states already feel are straining their budgets.

Some of the innovations already proposed can help with scoring accuracy. For example, human error can be reduced by having computers score open-ended items (Innovation 1.05), but this introduces new challenges making sure that the computers are properly programmed and the right keys and rubrics are used. Decreasing the novelty of assessment programs by using common items, forms and item banks (Innovation 2.01) can also improve scoring accuracy because it reduces the need to train scorers and set up new programs for every single state and form. Recently, some scoring companies have developed sophisticated logic that can catch errors before they are reported by applying artificial intelligence and algorithms[22]. To the extent that this innovation can be expanded, states would be well served.

Another way that scoring quality might be improved would be the adoption in the testing companies of quality control standards that are widely accepted

[22] Innovation 1.11, *Enhanced Quality Control Procedures:* Multiple improvements in quality control in printing and scoring tests, including automated methods of monitoring quality control in scoring and detecting errors before they affect reporting and application to the testing industry of accepted industry standards in manufacturing, project flow, software and communications.

in other high-stakes industries. There are numerous standards that might be applicable, such as those of the International Organization for Standardization (ISO) and the Workflow Management Coalition (WMC). The former have promulgated ISO 9000 standards, which are generic management system standards, which provide the organization with a model to follow in setting up and operating the management system that is built on a firm foundation of state-of-the-art practices. There are also specific standards for software, for communication and workflow that can be adopted by testing vendors. To the extent that certain companies distinguish themselves by adoption and certification in accepted standards, states will gain confidence in the use of those vendors in their testing programs.

INTERPRETATION

A major category of need for innovation is in interpretation of tests. Interpretation starts with the score reports that are the main product of the entire assessment system. Innovations in reporting and interpretation are needed for all the uses of testing: understanding student performance, teacher performance, and school performance; evaluating programs; and diagnosis of strengths and weaknesses.

Reporting needs fall into two categories: report design (what information the reports contain and how it is organized) and report accessibility (how the reports are presented to their audiences). The states need innovations in both these categories in order to use their assessment programs for improving the education process.

Need: Student tracking regardless of mobility

Longitudinal reports can serve a number of information needs: describing student, teacher and school performance and program evaluation. Longitudinal reports describe progress from testing occasion to testing occasion, usually year to year and grade to grade. Tracking students and cohorts across years provides valuable information about student learning that cannot be obtained with a snapshot report.

In order to obtain useful longitudinal reports, the states need reliable ways of matching students, teachers and programs across years. This requires the testing program to collect or access demographic information without illegal invasion of privacy.

One of the innovations the states need is an easier way to track students, teachers and programs from year-to-year without placing unreasonable logistical burdens on teachers to fill in demographic information on answer sheets. Methods like pre-identification services have been used for some time, which match students' answer sheets with preloaded demographic files at the scoring center. Not all states use these systems, however, because of their high cost, and they still result in many unmatched students. If no student is to be left behind, states need reliably generated longitudinal reports from their testing companies.

Innovation is also need in tracking students regardless of mobility. If a student moves from one school to another in the same district, one district to another in the same state, or even from one state to another, there should be a way to track that student on some set of standards so that those responsible for his or her education can understand the student's progress. Furthermore, a student's progress in future years reflects back on the effectiveness of programs used to teach the student in earlier years and this information should follow the student wherever he or she goes. This is often difficult because different testing vendors are used. What the states need is an innovative way of tracking the stu-

dent's programs and performance that stays with the student, but is protected accordingly as private information, similar to medical information[23].

Challenges to overcome in this innovation are that creating such a system may be politically or legally difficult in some states and that any kind of pre-gridding operations will cause a time-lag prior to testing and during this time period, since students who move in and out of school or between schools make some of the records inaccurate on the day of testing.

Student performance

The fundamental goal of the state assessment program is improving student performance. To that end, the states need excellent reports of student performance that are accurate, insightful and easy to understand. This is also required by P.L. 107-110. The main consumers of student level reports are the students themselves, the parents, and the teachers. None of these consumers should be expected to be expert in understanding testing data and statistics. Therefore, student level reports must be simple to understand, without watering them down so they lack the valuable information the program can provide.

Need: Parent reports that make sense for parents

For years, test publishers have attempted to create better parent reports. Innovations such as the narrative report appeared in the late 1970s and have become more sophisticated over time. Narrative reports explain the student's scores in simple language. In the late 1980s, schools were given the opportunity to provide cartoon reports in which the speech balloons of the characters explained the test scores. As time went on, and printing technology improved, the reports contained more and more graphics. States with large percentages of students whose parents did not speak English began providing reports in other languages, such as Spanish and Navajo. At first these languages were only used for static interpretive information preprinted on the back of the page, but eventually became dynamic text that changed with the students' scores.

The need for better parent reports continues, especially in light of the more sophisticated assessments states are using today. As new concepts such as proficiency are introduced to parents, reports must provide the information in easy to understand fashion.

An innovation the states would like to see is the creation of interactive, web-based reports (Innovation 1.01) that allow the parent to delve as deeply as possible into the information about their child. Presentation of the data with links to interpretive information, graphs and even videos that explain the data would be helpful in drawing parents into the education of their child. The challenges associated with this innovation are accessibility to all parents, since some do not have computers with Internet connections, and protection of the confidentiality of the results. One possible solution to the former challenge is to have the results available on school computers for parents to come in and access. A solution to the latter is a sophisticated system of passwords and registrations similar to what is used today to access bank accounts and financial statements on the web. The interactive student reports can also be useful for teachers to better understand test scores and determine what to do about the results.

[23] Innovation 1.04: *Reliable System for Linking Test Records to Students:* A set of affordable methods to reliably attach test results to students without unreasonable administrative effort, and even when students are mobile.

Teacher performance

The decision to use student achievement data as a measure of teacher performance is still somewhat controversial and not all states have endorsed this approach. However, the states that have done so need to have innovations in the provision of test information that helps accomplish this goal.

Need: Fair and acceptable ways to measure teacher performance based upon results

One of the criticisms of the use of student achievement data to measure teacher effectiveness is one of fairness. While it makes intuitive sense that teachers should be evaluated on how much students learn in their classrooms, how do we know which part of the student's achievement is attributable to a particular teacher? Various models have been proposed for using student data to evaluate teachers and these need further research. Whatever models are adopted by the states, the testing companies should be able to provide the information that supports that model accurately and efficiently[24]. This might call for the creation of special longitudinal reports that only include students who have been with that teacher for the full school year. It might also include the matching of a teacher's students' data from year to year. Once we are able to accurately measure student achievement gains while under a teacher's instruction, and eliminate the spurious data, we can present that information effectively to educational leaders and teachers.

School and District performance

State assessments must be used to evaluate school and district performance. Since schools are often the unit of funding and administration for school districts, they are often the focus of rewards and sanctions that provide the drivers for educational improvement. Under P.L. 107-110, the school is the basic unit for which Adequate Yearly Progress is determined, though districts also face increased accountability. This calls for accurate reports of school and district performance from year to year that, like the teacher reports, eliminate spurious data that do not allow valid conclusions about school performance.

Need: Reports that make sense for educational leadership and the general community

Reports must provide the data that educational leaders need to evaluate school effectiveness according to the accepted models for evaluation. Reports that make sense to the general community are also needed since these data are often presented to the public in newspapers, on websites, and in mailings to the community. Information that is complex and garbled causes the community to distrust the schools. Information that is clear and meaningful causes the community to trust and want to help the schools.

Methods that states have used for reporting include school report cards that are mailed home or placed on websites, community meetings and press releases. Report cards for districts and states are also required under P.L. 107-110 [§1111(h)].

States need more innovations in this area to continue to improve the clarity and timeliness of reports to the community of school performance as well as district and state performance. Interactive websites are a good way to do this. Many businesses have developed very sophisticated websites that provide excellent information in an interactive self-paced manner.

[24] Innovation 1.15, *Reports of Teacher Effectiveness:* Valid reports of student gains while under the instruction of an individual teacher, with spurious data eliminated. Reports would aggregate data from multiple groups of students taught by the same teacher.

The e-learning profession has developed excellent interactive tools for individuals in schools and businesses to learn complex content. Assessments can be used to determine what the person already knows so that the materials can be tailored to his or knowledge level. States need to have access to tools like these for informing their constituency about educational performance.

Program evaluation

If test results are merely used to report the status of schools and students and are not used to improve educational practice, they fall far short of their purpose. Program evaluation involves determining the efficacy of educational programs and activities. Which textbook programs works? Which technology programs are most effective? Which organizational schemes make the most sense?

Need: Reports and schemas that provide actionable information of which programs and practices work

States need reports and schemas (methods for working with the information) that lead to action. If the results are not available in the proper format at the proper time, or the interpretive schema is inappropriate or poorly applied, the results will not be effectively used to improve student learning.

This calls for a close interaction between test developers, scorers, and educational researchers so that the needed actionable data are provided efficiently and effectively. Each type of program evaluation requires the testing program to provide appropriate data. For example, the model promoted by Just For the Kids and National Center for Educational Accountability requires statewide testing, standards, information on student demographics, exemptions and special education status, student-level data with linked records between databases, and access to the data. The requirements are actually more specific and can be found at http://www.just4kids.org/US/US_otherstates.asp.

The Education Value-Added Assessment System, developed by William Sanders of Tennessee, requires annual testing with norm-referenced or criterion-referenced tests that possess certain psychometric features. Information can be found at http://www.sasinschools/evaas/index.shtml

Another value-added model is one developed by Anthony Bryk of the Consortium on Chicago School Research (http://www.consortium-chicago.org/achvm.html), which also depends upon annual testing at every grade with a standardized test and numerous demographics.

The innovation needed is for testing companies to place cooperate with researchers on the continued development of good program evaluation models[25] and on reasonable ways to provide the information needed to feed these models.

Diagnostics

Most state assessment programs have provided a mechanism for obtaining diagnostic information, either from the state assessment, or from a coordinated classroom assessment system. Many publishers have provided products that diagnose students' strengths and weaknesses on the state academic standards. P.L. 107-110 requires that the state assessments produce individual student diagnostic reports to parents. Furthermore the assessments must be provided to districts, schools and teach-

[25] Innovation 1.13: *Valid Program Evaluation Models:* Improved models for using state assessment programs for program evaluation.

ers in a manner that is clear and useful for improving the educational achievement of individual students.

Need: Better diagnostic information that is actionable

Regardless of the specific requirements in P.L. 107-110, diagnostic information is required for educational improvement. One of the difficulties that states have faced in the past is the balance between providing detailed diagnostic information on objectives and test security. The more specific the information is about what the student can and cannot do, the more the items of the test are exposed. Releasing forms yearly drives up costs and puts additional strains on development resources. A more intrinsic difficulty is providing the type of detailed information useful to on the basis of a single, end-of-year assessment. Useful information can certainly be gleaned from such assessments, but there are limits.

States need innovative methods to provide detailed, targeted diagnostic information to students on the state academic standards without compromising the state assessment program's validity, security and cost. One possible method is a coordinated system of classroom-based assessments that is tied to the state standards and to the state tested objectives, but not to the state tested items[26]. This diagnostic assessment addresses the need for detailed diagnostic information in a timelier manner, more often (when teachers can intervene and correct instead of merely memorializing failure after the fact).

Numerous systems exist on the market today and are of varying quality and efficacy. Providers of these systems should be required to collect efficacy data on the effectiveness of those systems for improving student achievement.

The innovation of providing common banks of objectives, items and tests (Innovation 2.01) would go a long way to toward satisfying the need for diagnosis. The key to the effective use of these tests for diagnosis is the type of data they provide. If that data is actionable and understandable, it would assist greatly in the improvement of student achievement.

APPLICATION

All of the innovations in testing programs: content, information, logistics, and interpretation are only effective to the degree that the information is applied properly. Many of the aspects of application of the testing information to the educational process are beyond the scope of this document, because they go past the testing program per se and move into the areas of curriculum development, staff development and pedagogy. To a certain extent, however, some innovations from the testing community can assist greatly in the appropriate application of assessment results to educational progress.

Improvement in practices

Improvement in educational practice comes from finding out what works well and doing it.

Need: Ways to identify processes and programs that work

In order to identify what works, the assessment data must be carefully targeted and clearly applicable to answer the question. The testing providers must be prepared to respond to knowledge gained from the education community about effective practice—such

[26] 1.02: *Coordinated Diagnostic Tests:* Systems of diagnostic, classroom-based assessments that are keyed to the state standards and assessments. These tests allow teachers to do formative evaluations throughout the year to improve student performance.

as research on reading, in particular, and instructional design more generally—and inform their state clients of this and the innovations they are putting into their testing tools and systems to reflect this improved knowledge base.

Improvement in testing quality

There are two types of information provided by a testing program: tests scores and transaction data. Test scores have been addressed throughout this document. Transaction data consists of information on who took the tests, where, when, and how the test itself behaved. This kind of data provides for the improvement of the test instruments over time. Bad items and item types are eliminated, ineffective procedures are replaced, and misunderstood reports are revised.

Need: Ways to continually improve the quality of testing programs by learning

While publishers have long used transaction data to improve their published instruments, there has been less consistent application of transaction data to the improvement of custom state assessments. States need methods to learn from their successes and mistakes to improve the assessment system.

Some of the innovations discussed already can be of use in meeting this need. For example, computer-administered tests (Innovation 2.02) can capture and save every keystroke, which information can be used to improve the test and the testing experience. Common item banks (Innovation 2.01) can store useful information on items beyond the traditional item statistics. Information websites (Innovation 1.01) can provide anecdotal information in an organized fashion so that state educators learn from each other about what works in assessment. Developing a constant improvement process will enhance the state assessment programs significantly.

Staff Development

Staff development for improving teaching is beyond the scope of this analysis, but a key area of staff development related to assessment is not. Teachers need to be made better makers and consumers of tests. This will only occur if educational leadership in a state develops a consensus that teachers need assessment training and motivates teachers to want it. Currently, teachers do not take available assessment courses in college because there are "too many other program requirements." To be effective, training in assessment may need to be required as part of staff development or re-certification.

Need: Staff development programs that make teachers better makers and consumers of tests

While high-stakes state assessments need to be professionally developed, there are many types of assessments that teachers develop. Teachers vary significantly in their training to develop valid, reliable tests, and many products and services exist to make it easier for teacher to develop or select tests from item banks and test banks. These products need to be tied to the state academic standards, but more importantly, teachers need staff development on how to use these assessments properly.

Staff development is time consuming and expensive, but fortunately new methods exist for effectively training teachers on testing. E-learning systems include both synchronous (instructor-led) and asynchronous (self-paced) as well as blended (a combination of both) methods. These program need not be completely online, but can be a combination of online, video, CD-ROM and face-to-face programs. Testing companies and other entities should devel-

op more comprehensive and more accessible teacher staff development programs using whatever methods work[27]. A teacher educated about assessment techniques and information use is a prerequisite to educational improvement.

In addition to staff development, a system of learning verification or certification[28] would be very useful. Administrators would know that the teachers are properly trained in the things they need to know about assessment. Teachers will benefit from knowing how to use test results effectively. The certification program can be associated with rewards for the teachers who complete it (such as a stipend) and can be accompanied by informal pre-certification tests that the teachers can take on a computer, in private, to prepare for the official certification. This method is widely used in other professions.

BUSINESS

The last set of needs surrounds the business of state assessment programs. States almost always contract with test providers – publishers, testing service companies, training companies, and technology companies – to deliver parts of the state assessment program. States have a responsibility to their taxpayers to conduct their business efficiently and effectively, and providers have similar commitments to their stakeholders. Without customers, business cannot survive, and without providers of services, states cannot run testing programs. This interdependence motivates all parties to develop better business systems so that both groups can thrive.

Contractual arrangements

Currently, each state contracts with testing companies to provide products and services. State business laws require differences in contractual provisions and details from state to state, and testing companies are used to this and have systems to deal with this method of doing business.

Need: Ways to share business arrangements that work, rather than reinventing the past

The decentralized nature of the business arrangements between states and testing companies often results in wasted efforts. Each state goes through the same process of negotiating with the testing companies that another state has just completed. Testing companies may be frustrated that they have to explain to each state individually the costs and logistical requirements of running testing programs. An innovative way of doing business that protects the independence and sovereignty of states, but allows them to learn from each other's contractual dealings so as to write fair contracts efficiently would be very beneficial to the states and to the publishers as well.

An obstacle to this is federal laws on competition that do not permit business to discuss or set prices with their competitors. One innovation that would be worth pursuing is some way that states could share a body of contractual techniques and information[29], including methods of fair pricing, so that they could concentrate on running testing programs and not spending taxpayers' money on contract negotiations and legal fees.

[27] Innovation 2.06: *Staff Development in Assessment Development and Use for Teachers:* Exemplary programs that train teachers to develop, use, and understand assessments using e-learning techniques, either alone or in concert with instructor-led training.

[28] Innovation 2.08: *Teacher Certification in Assessment Knowledge:* A portable certification of teacher competence in the development, use and interpretation of assessments.

[29] Innovation 1.03: *Collaborative State Resource for Assessment Systems:* An organization serving as a resource on the legal defensibility of state assessments, possible timelines for testing program schedules, and state/vendor contractual techniques.

Increasing value

The burden for increased testing faces limited funds at the state level. States need to get more for their money, i.e., greater value for their testing expenses. This includes reducing costs and increasing utility.

Need: Better ratio of utility to cost

Cost reduction is relative. Increasing the number of grades tested or subject areas and increasing the size of state tests by testing more academic standards and ordering more score reports will certainly increase total cost, but states need a way to decrease the unit costs of items and of student scores. As the ratio of utility (what the program provides for the state that leads to educational improvement) to cost increases, so does value.

Innovations that might increase value include the following:

- Item banks that reduce the amount of novel item and test development for which a state must pay (Innovation 2.01)
- Replacement of paper and pencil test (which require paper purchase, printing, handling, shipping, distribution, collection, scanning and storage or destruction) with computer-administered tests, when computer-administered tests become financially feasible (Innovation 2.02)
- Automated scoring of questions that currently require trained human readers (Innovation 1.05)
- Web-based reporting that saves the printing and distribution costs of reports (Innovation 1.01)
- E-learning programs that provide staff development more efficiently, particularly for remote locations (Innovation 2.06)
- Longitudinal tracking systems that allow more information to follow a migrant student, rather than discarding expensive test results because they cannot be matched (Innovation 1.04)

Each of these innovations was discussed elsewhere in this document and could result in greater cost-effectiveness in testing programs.

Protection

When assessment programs are high stakes, they create legal exposure for the states, districts and schools, as well as the testing companies. States and testing companies must work together to minimize the exposure to costly lawsuits and judgments.

Need: Better protection of the state education agency from litigation and liability due to errors and omissions of others

It is appropriate for each party to a contract to be responsible for errors and omissions of their own, but they should be protected from errors and omissions of others. Because the state assessment efforts are expanding so quickly, new areas of exposure and new types of errors and omissions are constantly being discovered. States, in particular, need to be protected against any error or omission of others. By sharing contractual language regarding liability and protection (included in Innovation 1.03), the states and testing companies can concentrate on improvement of the content and logistics of programs and not waste time and resources dealing with legal issues.

Acknowledgements

We thank the following individuals for their advice and input regarding this document, primarily through their participation at the Education Leaders Council Testing Summit, held February 20-21, 2002 in Austin, Texas. By including their names here, we acknowledge their contributions; this list should not be interpreted as necessarily implying any endorsement of the ideas contained in this document by the individuals or organizations listed herein. This document is a product of the participating states listed on the cover, the Education Leaders Council, and AccountabilityWorks.

Michaele Alcaro
Policy Specialist
Pennsylvania Department of Education

Dr. Arturo Almendarez
Deputy Commissioner for Programs and Instruction
Texas Education Agency

Leslye Arsht
President
StandardsWork

Virginia Belland
Vice-President, Sales
Eastern Division
EdVision

Sondra Bisig
Director of Curriculum Implementation
Lightspan, Inc.

Larry Bosley
Vice-President Territory Sales
Vantage Learning

Stacey Boyd
President & CEO
Project Achieve, Inc.

Merri Brantley
Research and Policy Advisor
Office of the State Superintendent of Schools
Georgia Department of Education

Benjamin Brown, Ph.D.
Executive Director
Evaluation and Assessment
Tennessee Department of Education

Ron Carriveau
State Assessment Director
Arizona Department of Education

David Chayer
Vice President, Research and Test Development
Data Recognition Corporation

Maye Chen
Vice President of Operations
Project Achieve, Inc.

Mitchell D. Chester
Assistant Superintendent
Office of Assessment
Ohio Department of Education

Dr. Criss Cloudt
Associate Commissioner for Accountability
Texas Education Agency

Wilmer "Bill" Cody
NCES/Westat
Education Policy Studies

William Cox
Standard and Poors

Richard Cross
Director of Research
AccountabilityWorks

Keith Cruse
Managing Director
Texas Education Agency

Joel Dando
Manager Public Affairs
Renaissance Learning Inc.

Jill Dannemiller
Associate Director of School Accountability
Ohio Department of Education

Alicia Diaz
Education Consultant
Ohio Department of Education
Policy Research and Analysis

Richard Dobbs
CTB McGraw Hill

Chrys Dougherty
Director of Research
National Center for Educational Accountability/Just 4 Kids

Dennis Doyle
Co-founder and Chief Academic Officer
Schoolnet, Inc.

Brad Duggan
Interim President and Executive Director
National Center for Educational Accountability/Just For Kids

Susan Engeleiter
President and Chief Operating Officer
Data Recognition Corporation

Thomas H. Fisher, Ed. D
Educational Testing & Assessment Administrator
Florida Department of Education

Steve Fleischman
Executive Director
Education Quality Institute

Andrea Fuller
Vice President, Sales & Marketing
SMARTHINKING, Inc.

Matthew Gandal
Executive Vice President
Achieve, Inc.

Mark Gedlinske
Chief Technology Officer
Data Recognition Corporation

Christopher Gergen
President/Chairman and Co-founder
SMARTHINKING, Inc.

Dennis Gormley
ETS K-12 Works

Maureen E. Grazioli
Vice President, Sales and Marketing
Riverside Publishing

Ruth Grimes-Crump
Office of Elementary and Secondary Education
U.S. Department of Education

Dr. David J. Harmon
Director of Research, Evaluation, and Testing
Georgia Department of Education

Carolyn Haug
Policy Assessment
Colorado Department of Education

Dr. Hugh Hayes
Deputy Commissioner for Initiatives and Administration
Texas Education Agency

Mark Heidorn, Ph.D.
Director of Program Operations-
Northern Region
CTB/McGraw-Hill

James M. Hill, III
Director, Computer-based Testing
Marketing
Harcourt Educational Measurement

Steve Hodas
The Princeton Review

James "Jim" Horne
Secretary, Florida Board of Education

Shannon Housson
Director of Analysis and Reporting
Texas Education Agency

Steven Hoy
Vice President
Lightspan, Inc.

Gary Huggins
Executive Director
ELAC

Jeremy Hughes
Director, Office of Educational
Assessment and Michigan Merit
Award
Michigan Department of Treasury

Brian Jones
Vice President for Communication and
Policy
Education Leaders Council

Debra Jones
Senior Intelligence Advisor
SAS inSchool

Dr. Margie Jorgensen
Vice President
Harcourt

John Katzman
CEO
The Princeton Review

Lisa Graham Keegan
CEO
Education Leaders Council

Richard Kohr
Measurement and Evaluation
Supervisor
Pennsylvania Department of Education
Division of Evaluation and Reports

Carolyn Kostelecky
Assistant Vice President of Educational
Services
ACT, Inc.

Steve Kromer
Vice President
NCS Pearson

Stephen Kutno
The Princeton Review

Jackie Lain
Associate Director
Standard and Poors

John E. Laramy, Ph.D.
President
Riverside Publishing

Cait Ngo Lee
Evaluation and Grants Coordinator
Lightspan, Inc.

Lenny Lock
Education Associate
Pennsylvania Department of Education
Division of Evaluation and Reports

Dr. Kathleen Madigan
Executive Director
National Council on Teacher Quality

Gary Mainor
President
NCS Pearson

Dewayne Matthews
Vice President
Education Commission of the States

Patricia McDivitt
Vice President, Educational
Measurement & Test Development
ETS K-12 Works

Thomson W. McFarland
Research Associate
AccountabilityWorks

Brian McGee
Regional Vice President, Southern Region
Harcourt Educational Measurement

Georgia McKeown
Chief of Staff - Jim Horne
Florida Board of Education

Dr. Ron McMichael
Deputy Commissioner for Finance and Accountability
Texas Education Agency

Michael McNally
Vice-President of Business Development
Vantage Learning

Chad A. Miller
Policy Analyst
Education Leaders Council

Libbie L. Miller, Ph.D.
Statistical / Research Analyst
Lightspan

Dr. William J. Moloney
Commissioner of Education
Colorado Department of Education

Dr. Mark Moody
Assistant State Superintendent for Planning, Results, and Information Management
Maryland State Department of Education

Judith Morgan
Executive Assistant/Press for Faye Taylor
Tennessee Department of Education

Ina Mullis
Professor and Co-Director of the International Study Center
Lynch School of Education, Boston College

Dean H. Nafziger, Ph.D.
President
Harcourt Educational Measurement

Dr. Beverly Nash
Assistant Division Director
Research, Evaluation, and Testing
Georgia Department of Education

Jeff Nellhaus
Associate Commissioner for Student Assessment
Massachusetts Department of Education

Jim Nelson
Texas Commissioner of Education
Texas Education Agency

Jo O'Brien
Assistant to the Commissioner
Colorado Department of Education

Dr. Billie Orr
President
Education Leaders Council

John Oswald
Consultant
AccountabilityWorks

Shaundra Overmyer
Manager of State Assessment Programs
Data Recognition Corporation

Terrance (Terry) D. Paul
Co-Chairman
Renaissance Learning Inc.

Dr. Ron Peiffer
Assistant State Superintendent of Schools
Maryland State Department of Education

Heidi Perlman
Massachusetts Department of Education

Robert B. Peterson
President
C4SI, Inc.

Linda Pfister
President and CEO
ETS K-12 Works

Susan Phillips
Adviser and Consultant
Accountability Works

Susan Pimentel
Standardswork

Kelly Powell
ELC Consultant

Debbie Ratcliff
Senior Director of Communications
Texas Education Agency

Theodor Rebarber
President
AccountabilityWorks

Dr. June Rivers
Assistant Manager, Value-Added Assessment and Research
SAS in School

Marilyn Roberts
Michigan Department of Education

Ed Roeber
Vice President for External Relations
Measured Progress

Dr. Bill Sanders
Research Fellow, University of North Carolina
Manager, Value-Added Assessment and Research
SAS in School

Gary A. Schaeffer, Ph.D.
Director of Research Projects
CTB/McGraw-Hill

John Schilling
Chief of Staff
Education Leaders Council

Linda C. Schrenko
State Superintendent of Schools
Georgia Department of Education

Amanda Seals
Director of Media Relations
Office of the State Superintendent of Schools
Georgia Department of Education

Ramsay W. Selden
Vice President & Director, Assessment Program
American Institutes for Research

Anne Smisko
Associate Commissioner for Curriculum, Assessment, and Technology
Texas Education Agency

Malbert Smith
President
MetaMetrics, Inc

Russell R. Smith
Managing Director
Rothary Landis Group

Larry Snowhite
Vice President, Government Relations
Houghton Mifflin Company

Dr. Lew Solmon
Senior Vice President
Milken Family Foundation

Patricia Spence
Director of Psychometric Services
Data Recognition Corporation

Bernia Stafford
Vice President of School Marketing and Evaluation
Lightspan, Inc.

Jack Stenner
Chairman and CEO
MetaMetrics, Inc

John Stephens
Executive Director
National Assessment Governing Board

Phyllis Stolp
Director of Development and Administration
Texas Education Agency

Karen Stroup
Office of Management Services
Colorado Department of Education

Circe Stumbo
President
West Wind Enterprises

Leonard Swanson
Vice President, Technology
ETS K-12 Works

Robert W. Sweet, Jr.
Professional Staff Member
Committee on Education and the Workforce

Faye Taylor
Commissioner of Education
Tennessee Department of Education

Blake Thompson
Data Analysis Manager
GreatSchools.net

Cassandra Thompson
ETS K-12 Works

William "Bill" Tudor
Chairman & CEO
EdVision

Cynthia Ward
Vice President, Contract Management and Proposals
ETS K-12 Works

John Winn
Assistant Secretary
Florida Board of Education

Wendy Yen, Ph.D.
Vice President, Research
ETS K-12 Works

Dr. Michael Yoes
NCS Pearson

Philip B. Young, Ph.D.
Director, State Programs
Riverside Publishing

Victoria Young
Director of Instructional Coordination
Texas Education Agency

Jodie Zalk
Assessment Specialist
Massachusetts Department of Education

Molly Zebrowski
Manager of Business Development
NCS Pearson

Charles Zogby
Secretary of Education
Pennsylvania Department of Education