RENNIE CENTER
for Education Research & Policy

PACE

MAY 2011

The Road Ahead for State Assessments

Research commissioned and report produced by Policy Analysis for California Education (PACE) and RENNIE CENTER for Education Research & Policy

David Plank, *Executive Director,* PACE

Jill Norton, *Executive Director,* Rennie Center

Corinne Arraez, *Communications Director,* PACE

Ivy Washington, *Operations Manager,* Rennie Center

Design

Tanya Lazar, lazar design

Cover photograph by Kate Samp

Acknowledgements

This publication was made possible by a grant from Carnegie Corporation of New York, and we are grateful for their support. We are also grateful to the authors of the three research papers: Mark Reckase, Robert Linquanti, Chris Dede and Jody Clarke-Midura. We extend our thanks to the individuals who reviewed drafts of the report for their feedback and insight: Pendred Noyce, Lisa Famularo, as well as to several anonymous reviewers who provided feedback on the individual papers.

About PACE

Policy Analysis for California Education (PACE) is an independent, non-partisan research center based at the University of California—Berkeley, the University of Southern California, and Stanford University. PACE seeks to define and sustain a long-term strategy for comprehensive policy reform and continuous improvement in performance at all levels of California's education system, from early childhood to post-secondary education and training. PACE bridges the gap between research and policy, working with scholars from California's leading universities and with state and local policymakers to increase the impact of academic research on educational policy in California. For more information, please visit www.edpolicyinca.org.

About the RENNIE CENTER

The Rennie Center's mission is to develop a public agenda that informs and promotes significant improvement of public education in Massachusetts. Our work is motivated by a vision of an education system that creates the opportunity to educate every child to be successful in life, citizenship, employment and life-long learning. Applying nonpartisan, independent research, journalism and civic engagement, the Rennie Center is creating a civil space to foster thoughtful public discourse to inform and shape effective policy. For more information, please visit www.renniecenter.org.

PACE and the Rennie Center are supported by generous contributions from:

The James Irvine Foundation

The Noyce Foundation

The William and Flora Hewlett Foundation

The statements made and the views expressed in this report are solely the responsibility of the authors, and do not necessarily reflect the views of PACE, the Rennie Center, or our funders.

Suggested Citation

Policy Analysis for California Education and Rennie Center for Education Research & Policy. (May 2011). *The Road Ahead for State Assessments.* MA: Rennie Center for Education Research & Policy.

Table of Contents

Introduction i

Computerized Adaptive Assessment (CAA): The Way Forward
Mark D. Reckase, Michigan State University 1

Strengthening Assessment for English Learner Success: How Can the Promise of the Common Core Standards and Innovative Assessment Systems Be Realized?
Robert Linquanti, WestEd 13

Next Generation Assessments for Measuring Complex Learning in Science
Jody Clarke-Midura and Chris Dede, Harvard University
Jill Norton, Rennie Center for Education Research & Policy 27

Closing Remarks 41

Glossary of Key Assessment Terms 41

Introduction

The adoption of the Common Core State Standards presents states across the nation with an unprecedented opportunity to enhance the educational opportunities they provide students. States that have adopted the Common Core State Standards are now in the early stages of revising curriculum frameworks, adopting new instructional materials, developing new systems of assessment, and providing professional development for teachers to prepare them to deliver instruction aligned to the new standards. This process has the potential to fundamentally transform public education for the majority of U.S. students. It is therefore essential that policymakers and education leaders take full account of the issues and challenges that lie ahead as early as possible in the implementation process.

New assessments can be a key driver of the successful implementation of the Common Core State Standards, providing support for deeper learning and holding educators accountable for their students' progress toward true college and career readiness. When they are well-designed and well-used, assessments can motivate students and teachers, and focus their attention on the knowledge and skills that really matter for student success. Conversely, though, assessments that are poorly designed and implemented can narrow the curriculum, impoverish instruction, and undermine students' enthusiasm for learning.

Recognizing the transformative importance of assessments in realizing the promise of the Common Core, the U.S. Department of Education has funded two consortia of states that will work together to develop new assessments aligned to the new standards in English-language arts and mathematics. The Partnership for the Assessment of Readiness for College and Careers (PARCC) comprises 25 states including Massachusetts and California, while the SMARTER Balanced Assessment Consortium (SBAC) comprises 29. (Some states participate in both consortia.)

The two consortia are beginning to address the challenges of next generation assessments. But even as they strive to expand the frontiers of current knowledge and practice, the constrained financial resources of most states and the short timeline for implementation (new assessments are to be in use by 2014) makes the prospect of radical changes daunting. The essential end game is to develop rigorous assessments that effectively and efficiently serve the twin purposes of accountability and supporting better instruction.

As the two consortia begin their work, our hope is that they aim to build a system that responds to the immediate challenge of measuring student performance against the Common Core State Standards, and that establishes a firm foundation

for assessments that can continue to evolve toward ever-greater precision and utility over time. As the consortia move to develop a system of assessment that measures student performance in core academic subjects against national standards, the goal is to ensure that the systems are sufficiently robust and adaptive that they can provide the information needed to assess diverse student populations and more complex and challenging subject matters.

This report includes three papers that address critical "next generation" issues in assessment policy that can help to guide the choices made about system design: computer adaptive assessments, assessment of English learners and assessing science. None of these topics has received the attention that it deserves in the current debate on assessment policy. These three papers cannot provide a definitive or comprehensive plan for the next generation of assessments, but they do describe some of the critical attributes of a better system. The common conclusion in all three papers is that assessment policy will have to take full advantage of new technologies to provide useful and timely information to students and teachers about the quality and effectiveness of teaching and learning. The authors' provide a vision of new assessments that goes beyond the horizon of current practice.

The first of these three issues is computer-adaptivity. Both PARCC and SBAC are committed to developing computer-based assessment systems, but SBAC plans to develop a computer adaptive system, while PARCC does not. This is the most significant difference in the strategies of the two consortia, and could have lasting implications for the next generation of state assessments. The practical challenges that must be overcome to develop a state-wide computer adaptive system are substantial, but such a system may establish the platform that will enable states to solve some fundamental assessment problems, including the assessment of career readiness and the assessment of English learners. Mark Reckase of Michigan State University discusses the state-of-the-art in computer adaptive assessment, and identifies the costs and consequences of an immediate move toward the implementation of a computer adaptive system.

The second issue is the assessment of English learners. This is an urgent issue in California, where more than half of school-aged children come from homes where English is not the primary language, but it is also an issue of growing importance in states across the country. The fundamental question is how to design assessments that accurately measure students' mastery of academic content and not simply their mastery of Standard English. On the one hand, this requires the development of better assessments for evaluating students' English skills, as acknowledged in a separate federal grant program for the development of English-language proficiency assessment systems. On the other hand, though, it will also require the development of instruments that are simultaneously robust and flexible enough to assess the performance of English learners in all subject matters, and not just in English. Robert Linquanti of WestEd surveys the critical issues in the assessment of English learners, and points the way toward a system that ensures fairer and more accurate assessment for all students, including English learners.

The third issue is assessment in science, where the inadequacies of traditional assessments have been particularly troubling. Multiple-choice tests generally do a poor job of assessing students' knowledge and skills in science. They are especially ineffective at determining how well students are developing sophisticated inquiry skills—a key capability for science, technology, engineering, and mathematics (STEM) careers. Chris Dede and Jody Clarke-Midura of the Harvard Graduate School of Education describe the potential for next generation state assessments in science that utilize computer technology to deliver and score assessments, and to report the results to teachers. The authors argue for new types of assessments that deepen students' understanding of core science concepts over time. They illustrate their argument with a case study of virtual performance assessments (VPAs) in science that are currently in development at Harvard.

The work that PARCC and SBAC are doing marks a big stride forward in assessment policy, but this is only the beginning of a long journey. Our hope is that the assessments that both consortia are developing will not only help to address the challenges posed by the implementation of Common Core State Standards, but also put us on a path toward assessments that more accurately and effectively measure and support students' learning, and their progress toward readiness for college and careers far into the future.

Computerized Adaptive Assessment (CAA): The Way Forward

Mark D. Reckase, Michigan State University

The Educational Context for Large-scale Assessment

Large-scale assessments have always had the goal of helping improve the educational process, be it through accountability, helping to focus instruction, providing diagnostic information, or evaluating educational programs. In the past, large-scale assessments were purchased from vendors by schools or districts and the criterion for selection was sensitivity to local curriculum and instructional practice. More recently, large-scale assessments have been mandated by state or federal law and are custom developed to match the needs of the states or the country. Kifer (2001) provides a description of the educational environment for large-scale assessment with a somewhat critical view, while others, such as Phelps (2005), describe the current environment for large-scale assessment in a more positive light. These authors, and the others writing on large-scale assessment issues (e.g. Koretz, 2008; Rothman, 1995; etc.), agree that such assessments are now tools of educational policy and that it is important to understand how these tools function and the legitimate uses for them. There is general agreement that large-scale assessment is an important component of any accountability system and that an additional desired use is to foster good educational practices. There is less agreement about how to develop assessments to support these uses.

Even as the proper uses of large-scale assessments are being debated, the design of the assessments and the mechanisms for delivery are changing. It is becoming clear that the major form of communication in the future will be through a technology (computer) format. Smart phones are now widely available, e-readers are top gifts, and the print industry is going through a transition from paper to electronic presentation of text. Further, almost all professional writing is done on a computer keyboard. A 2006 survey indicated that 95 percent of first year college students wanted to use computers for writing activities (Kennedy et al., 2008). It is not surprising that in our current technological environment there is a desire to use the power of technology to administer and score tests. This is seen most clearly in the proposals by both the SMARTER Balanced Assessment Consortium (SBAC) and the Partnership for the Assessment of Readiness for College and Careers (PARCC) to use computerized procedures for their assessment delivery systems.

Given the wide availability of computers, the desire for computerized administration of tests extends beyond using the computer to present material and collect responses, although there are many advantages to such applications by themselves, it also extends to putting some "intelligence" behind the ways that test tasks are presented for administration and the ways that they are scored. Within the testing industry, adding intelligence to the selection of test tasks is called "computerized adaptive testing" when the selection of tasks is related to the characteristics of each individual person (Wainer, 2000). The test is adapted to each person according to what the test developer determines are important characteristics of the individual. SBAC plans to use this type of testing in their proposal for innovative assessment, and they have given it the label "computer adaptive assessment" (CAA).

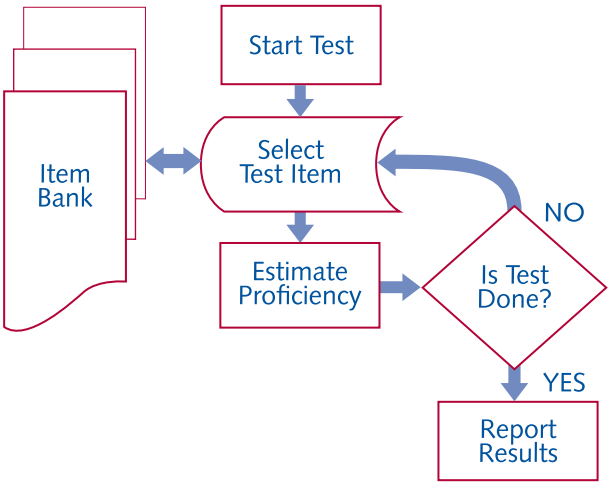
The Use of Computers for Educational Testing

The use of computers for the testing of achievement was first given practical consideration in the 1970s. These efforts came both from a perspective of making testing more efficient (i.e. Lord, 1970) and that of getting instructionally relevant diagnostic information from the test (i.e. Brown & Burton, 1978).

The most common way that computers are used to adapt the test to the characteristics of the person is to match the difficulty of test tasks (often called test items) to the estimated level of ability of the person being tested. Current tests are designed to measure the achievement of large groups of students at the same time using the same set of test items. Getting good measurements of all examinees has many technical challenges. For example, to measure students who have high levels of achievement well, difficult items must be included on a test. These difficult items are not appropriate for those students at the other end of the achievement continuum. Yet, for practical reasons, all students get the same booklet of test items. Similarly, tests must contain easy items so that low achieving students can be well measured, but these will be much too easy for high achieving students. Computer adaptive assessments may provide a solution to this problem, as the computer can be programmed to select the items that are appropriate for each student based on his or her current estimate of level of achievement. A well designed test of this type gives each student a different test that is matched to his or her estimated level of achievement. This type of testing is especially appropriate when there is a wide range of student capabilities as is the case when administering statewide assessments (Wainer, 2000, Chapter 1).

A key idea behind this approach to testing is that student achievement is measured along a continuum in the same way that we measure temperature, time, etc. The goal of the assessment is to select test items that give the best information about the location of a student along this continuum. This is different from traditional ways of scoring tests by summing the number of correct responses. Instead, a model is proposed that connects the difficulties of the test items and the responses of the examinee to a location on the continuum. The models used for this purpose are called item response theory (IRT) models and they relate the probability of getting a specific test item correct to the location of an examinee on the continuum of achievement (Lord, 1980). The use of such models frees the testing process from the requirement to administer the same set of test items to every examinee. There is a cost to this, however. Items must be calibrated to the achievement continuum before they are used for operational scoring of the test. "Calibration" amounts to estimating the functional characteristics of the test items by administering them to a sample of examinees from the target examinee population (see Parshall (2002) for specifics about calibration). The sample must be of sufficient size to get good estimates of the characteristics of the test items. It is the calibration information that allows the use of different sets of items to determine the locations of different students on the same achievement continuum.

Figure 1. Schematic diagram of the functioning of a CAA.



Advantages of Computerized and Computerized Adaptive Assessments

Tests can be administered by computer without being adaptive. These are fixed tests that use the computer as a means for presenting test items. This simple approach has several advantages independent of the advantages from adaptation. Computer administration puts the test in a mode that is becoming more and more familiar to some students. It is probably already the case that requiring students to write using pen on paper puts some of them at a disadvantage because they seldom write that way. As a result, hand written essays may yield underestimates of what students can actually do. Computer administration is becoming the more familiar format for academic work. In fact, the plan for the 2011 NAEP Writing Assessment is to have students use a word

processor on a computer to prepare their responses to prompts (National Assessment Governing Board, 2010).

A more important advantage is that the computer format allows more creative ways of assessing students' skills and abilities. The multiple-choice format is a result of the need to test large numbers of persons and quickly score the results. If computers can score tests as the responses are being made, then any computer scorable item format is equally good. With the advances in computerized scoring of open-ended test items, the possibilities for creative test formats are endless (see Shermis & Burstein (2003) for examples related to essay scoring).

Another advantage is that every response that an examinee makes can be recorded along with the time that it took to make it. This gives much more information from each test task. The challenge is to use all of that information in ways that enhance the results of the testing process. There is active research in this area, and no doubt useful procedures for using this expanded response information will become available in the near future (Drasgow & Mattern, 2006; van der Linden, 2006).

The above advantages can accrue to all tests that are administered by computer. A well designed CAA adds significantly to those computerized tests that have a fixed set of items (van der Linden & Glas, 2000). The CAA can either have the same measurement accuracy as a fixed test using fewer items, or greater accuracy than the fixed test using the same number of items. By targeting the difficulty of the test items to each student, more precision per test item is obtained. In a sense, the items on the traditional test that are too easy or too hard for specific examinees are wasted because they do not provide much useful information about the location of the examinees along the achievement continuum. By eliminating those items for each examinee, greater testing efficiency can be obtained. This efficiency can be used to get either shorter tests or more precision by administering additional well targeted items. The greater efficiency can also be used to get more diagnostic information about students' areas of difficulty.

A CAA can also result in much more efficient decision making than a fixed test in either computer presented or paper-and-pencil form. The CAA can select items for optimal decision making and stop the test once an accurate decision has been made. There is no need to give more items just because an accurate decision has not been made for others in a large group of students.

Components of a Computerized Adaptive Assessment

Like other types of technology, CAAs are complex. While they can achieve the advantages listed above, CAAs must have a number of well developed parts to function properly. The first and most important part is a well-constructed item pool. An item pool is the full set of test items along with technical information about how each item functions, saved on a computer storage medium. As with any other test, the quality of the results depends on the quality of the test items. The best of technological innovations cannot overcome poor test item writing. And, a CAA requires more test items than the traditional test form even though it does not administer all of the items to each examinee. More items are required so that the full range of observed achievement can be measured at an equal level of precision. A traditional paper-and-pencil test does not do that. Examinees at the extremes of the achievement continuum are measured less well than those in the middle range of the continuum.

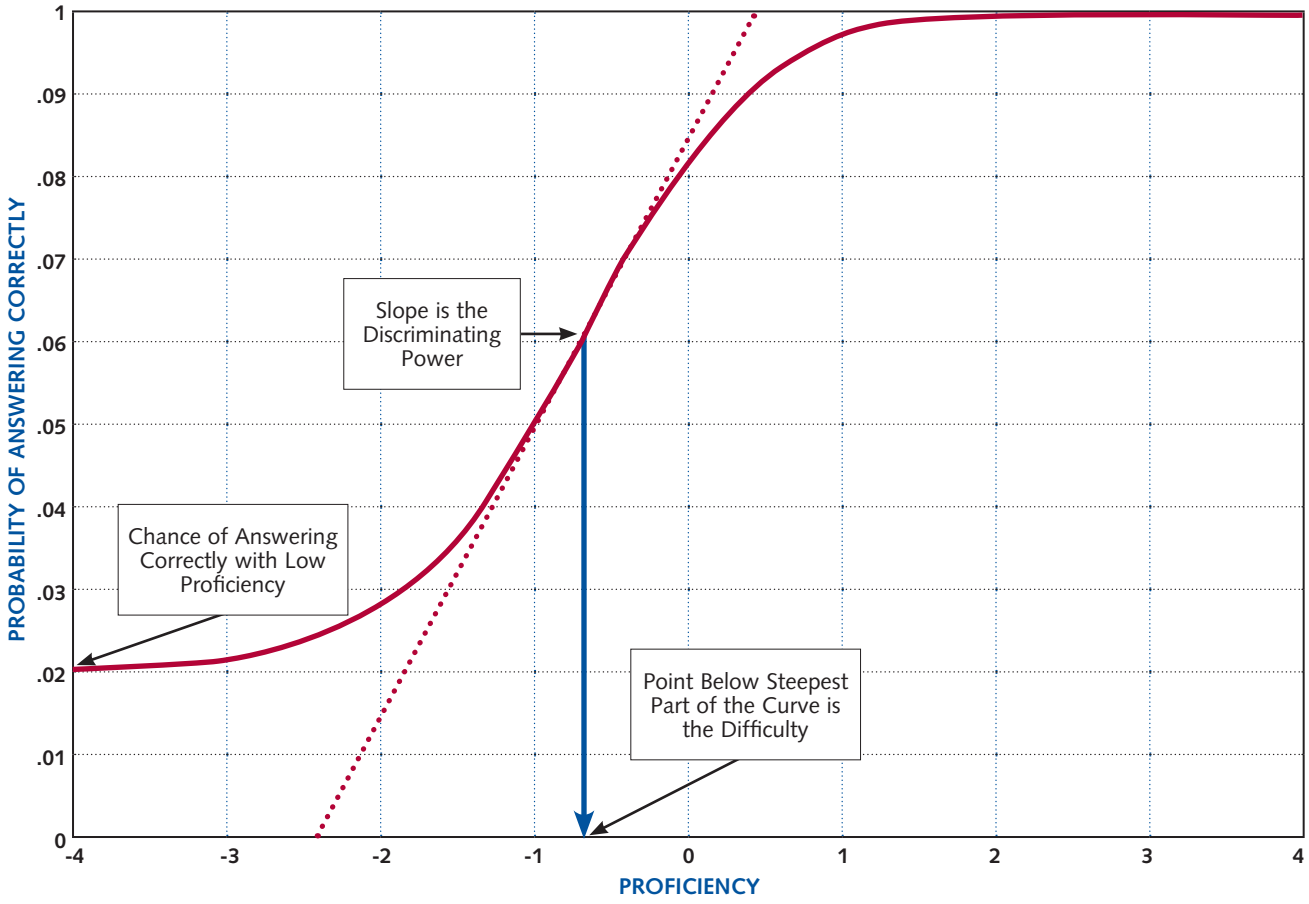
The required size of the item pool is dependent on a number of factors (see Parshall, Spray, Kalohn & Davey (2002) for a summary of the factors). The first is the precision of estimate of student achievement that is desired. Highly precise estimates require bigger item pools than rough estimates. A second factor is the range of achievement to be measured. If the range is very broad, a larger item pool is needed because items with a large range of difficulty are required. A third factor is the level of stakes associated with the test. If the test is very high stakes, there is also high motivation to cheat on the test to get high scores. This requires large item pools so that each examinee receives quite a different set of test items. In the technical language of the testing industry, the number of common items that are seen by two examinees is called overlap and the proportion of times that an item is seen by the full population of examinees is called exposure. Large item pools are needed to minimize overlap and exposure in a high stakes testing environment. If the test is low stakes, smaller item pools can be used and

issues related to cheating are not of a major concern.

Technical information about the test items in an item pool is obtained through the process of calibration. In the process of calibration all of the items to be included in an item pool are administered to a sample of examinees in order to estimate each item's level of difficulty,[1] discriminating power (its power to discriminate between examinees at different levels of achievement),[2] and the likelihood of getting a multiple-choice item correct even with very low knowledge of the subject matter.[3] In some cases only the level of difficulty is estimated. Estimates of difficulty are based on the assumption that the chances of answering correctly increase with students' level of knowledge.

Figure 2 shows the assumed relationship between proficiency and the chances of answering correctly for

Figure 2. Item calibration.



1 Difficulty refers to the level of challenge provided by the test item. In classroom contexts difficulty is usually indicated by the percentage of students who answer a test item correctly. Values near 100 percent indicate easy items and those near 0 percent are very hard items. Most operational tests have test items averaging in the 60 to 70 percent range.

2 Discriminating Power (or discrimination) indicates the capability of the test item to distinguish between low and high performing examinees. When a test item has high discriminating power those who answer the test item correctly are clearly higher performers overall than those who do not answer the test item correctly. Low or zero discriminating power means that a test item is very poor at distinguishing level of performance. Low discriminating items are generally screened out before constructing operational tests.

3 Chances of Low Performing Examinees Responding Correctly—When multiple-choice items are used, there is some likelihood that examinees will select a correct answer even when they do not know the required subject matter. In some cases, they choose the correct answer as the result of random guessing. In most cases, however, examinees engage in a more complex process of eliminating alternatives. Generally, the likelihood that low performing examinees will answer difficult test items correctly is less than if they randomly guessed. In the technical literature, this value is called a pseudo-chance level to indicate that it is not the same as random guessing.

one test item. The values obtained by the calibration process are indicated by the text boxes. Without calibration information a CAA cannot yield the promised advantages.

The second part of a CAA is an algorithm for selecting the test item to administer to an examinee based on what is known about the examinee and the characteristics of the items in the item pool. A number of algorithms have been developed for CAAs (see van der Linden & Pashley (2000) for a summary), but they all share some common features. One feature is a way to define an optimal test item for the current examinee given what is known about the examinee. Optimal is usually defined by a combination of statistical criteria for minimizing error in the final reported score and content information needed to balance the content, cognitive processes, and item types administered to an examinee. Other criteria that are often used have to do with the frequency that an item is selected for administration. When test security is of concern, test items cannot be administered to a high proportion of the examinees. Item selection algorithms contain rules on the rate of use of items called "exposure control" to reduce the frequency that items are seen by examinees. The combination of an optimization technique to reduce error and exposure control to support test security defines the item selection algorithm.

The third part of a CAA is an algorithm for estimating the location of the examinee on the continuum that is the target of the assessment. This estimate is based on the responses to the set of test items that have been administered at that point in the testing process and the calibration information that is available for the test items. The estimation procedures are well-known statistical methods, but they are computationally intensive. Early CAAs needed to take into account the computational load on the computers because of slow computational speeds in the 1970s. Now, any notebook computer can easily meet computational speed requirements. It does mean, however, that the estimates of location on the reporting score continuum for a CAA cannot easily be computed by hand. Estimation of location is not simply summing the number of correct responses. Instead, the CAA algorithm makes use of the information obtained from the examinee's pattern of responses to administered items to locate the examinee on the performance continuum (van der Linden & Pashley, 2000).

The final part of a CAA is deciding when to stop the test. Paper-and-pencil tests typically have a fixed number of test items. The flexibility of computer administration makes many other options possible. In general, these are called stopping rules. For example, test items can be administered until a desired level of precision of measurement is reached. This is a popular approach because all examinees are measured with the same level of accuracy. That is not the case for fixed tests. Using this approach, examinees typically receive different numbers of items. Examinees who are very consistent in their pattern of responses might be administered fewer items than an examinee who sometimes misses easy items and then answers difficult ones correctly. If such a test administration plan is used, the CAA is labeled as variable length. Another type of variable length CAA is to administer items until a desired level of accuracy of classification is reached. A third possibility is to administer special diagnostic items to examinees that show evidence of some level of misunderstanding of content. Of course, another option is to give all examinees the same number of test items. This is a fixed length CAA (Thissen & Mislevy (2000) provide a summary of stopping rules).

This discussion of the component parts of a CAA should make it clear that there is not a single design for a CAA, but rather a number of options that are selected depending on the purpose of the assessment and the stakes that are attached to that purpose. Many design decisions need to be made before the development of a CAA can begin and before implementation.

The Costs of Developing a CAA

The costs of developing a CAA are similar to those of developing a paper-and-pencil version of a large-scale assessment of educational achievement, but the costs are apportioned somewhat differently. These cost drivers will be described in some detail to emphasize the amount of work that needs to be done to enable implementation.

Item Pool Development

The first step in the development of a CAA is to write or select from existing items the test items that will be used in the CAA. If these are items that are scored correct/incorrect, then at least 200 acceptable items are needed if the test is low stakes. High stakes tests require more. One operational licensure examination uses item pools that are 1600 to 2000 items in size. The number of test items required when there are more than two score categories is not well researched. Such items provide more information than those with two score categories so the tests can be shorter. Open-ended items tend to be more memorable, though, so their use raises more security concerns.

A low stakes CAA can use the same item pool for months or years, but for high stakes use, item pools need to be changed regularly—for example, every three months. Davey and Pitoniak (2006, pages 548-549) provide a discussion of the factors that influence the required size of an item pool.

A second part of item pool development is the calibration of the items. This means administering the test items to several hundred to several thousand examinees for the purpose of empirically determining how the items function—that is, to determine the difficulty and discriminating power of the items (see Figure 2). The data from these administrations are analyzed using an IRT calibration program to get the information that is used for item selection (see Figure 1).

Calibrating the item pool has a challenging data collection aspect. It is not likely that all of the items in the item pool can be administered to the same group of individuals at the same time. The most common solution is to produce a number of paper-and-pencil test booklets with overlapping sets of test items. Each booklet is designed to be administered in a reasonable period of time. The data from all of the test booklets are analyzed together to yield the calibrated item pool.

It is also possible to administer all of these sets of items by computer. Since the test will be on computer, it is actually better to do it in this way because items sometimes function differently in paper-and-pencil and computerized form. But, this requires having all of the computer acquisition and interface design work done early in the process. Sometimes it is not practically possible to calibrate all of the test items in computer form.

Once a CAA system is operational, new test items can always be administered by computer for calibration. The new test items can be inserted into the test that is in current use, with the examinees' responses to the new items excluded from the operational score. These responses are saved along with the estimates of proficiency for later use in calibrating the new test items. When sufficient data have been collected, other new test items can be inserted into the test. By collecting item calibration information in this way, the CAA item pool can be self-sustaining. That is, special studies will not be needed to collect data for item calibration.

The actual costs for item pool development include: cost of producing test items, cost for administering the test items for calibration, cost of recruiting the calibration sample, cost of analyzing the data, and the cost of entering the items and calibration into a computer accessible form. This last point requires that there is a database that is available to store the items. If there is not, there is also a cost for developing the item pool storage system. Of course, the cost is not only in dollars; it is also in time. The development of a calibrated item pool will take at least a year, and longer if many items need to be written and edited and if an item pool design needs to be produced. The item pool is not a random selection of test items. It must be carefully designed to meet the purposes for the CAA.

Computer Access

The biggest administration problem for a CAA is getting access to the number of computers needed for administration to the examinee population. The first issue is determining the types of computers that are acceptable for use. This depends on the speed of processing required and the screen characteristics needed to present the test items. If test items require detailed analysis of diagrams or photos, or the reading of long texts, then small screens put the examinee at a disadvantage. If all test items are only a few lines of text long, screen size may not be an issue. Issues of fairness might require that all examinees use the same computer configuration. For large scale assessments, this may not be possible, but the computers at least need to be comparable.

One way to look at the cost issues related to computer access is the amount of time that is needed for each test. This is often discussed as "seat time" in front of the computer. Suppose, for example, that a state has 100,000 students at a grade level to be tested. Suppose also that the test of one content area takes one hour and that three content areas need to be tested. This means that seat time per student is a minimum of three hours—four hours will probably be needed to deal with all administrative issues. That means testing that grade level will require 400,000 hours of seat time in front of a computer. If it is possible to test two students per day on a computer, that means 200,000 computer/days are needed. If all students need to be tested within a two-week window, that means that 20,000 computers are needed to meet the demand for that grade level in that state.

Of course, the number of computers needed can be reduced by shortening the testing time or widening the testing window. These options have implications for test security and for the accuracy of measurement. Deciding on the correct balance of all of these factors is a difficult design issue. In any case, getting access to the needed number of computers will be a major cost.

One approach to dealing with the seat time issue is to have each examinee use his or her own computer. This eliminates the issue of cost of access, but it raises other issues related to security, fairness and equity. Resolving the issues of computer access is one of the most challenging parts of implementing CAA. Of course, these issues also apply to non-adaptive computer delivered tests as well.

Administration Software

Assuming that the computers are available for use, there is still a need for a software system to select the items from the item pool, present them on the computer screen to the examinee, record the response to the item, score the item, estimate the location of the examinee on the continuum, determine when to stop administering items, store the results in a database, and report results. This is a complex computer software system, the cost of which can be dealt with in a number of ways. One alternative is to purchase off-the-shelf or custom-designed systems from the private-sector, with costs negotiated through contracts with the vendors.

Another alternative is to produce custom software to meet the needs of the CAA. This is an expensive proposition and includes time and cost needed to produce a design and the cost of developing and testing the resulting code. Such systems have been developed for states, but sometimes there are problems with meeting development schedules and bugs in the software that are found during initial implementation.

Either using existing software or developing new software will be expensive. Forcing the implementation of a CAA to fit existing software, or underfunding the development of new software, will likely result in serious problems later in the implementation of the CAA. The CAA software is more elaborate than typical software that can easily be purchased. It is important that the developers have experience in the area of psychometrics. Any new implementation of software also takes time. Problems will likely occur if the software is used for operational testing before it has been fully checked under the level of demand needed for implementation.

Long-term Implementation

If there is an expectation that the CAA will be in place for years, other cost issues arise. These issues relate to technical support and maintenance, and refreshing item pools. As with any other part of our technological society, the technology for implementing the CAA will change. Computers and software systems will become obsolete. New computer hardware will come into com-

mon usage. Old hardware will breakdown and it will need to be maintained. Computer interface software may change, requiring revisions to the way that items are presented.

At the current rate of computer innovations, computers are obsolete after about five years. Often these old computers will not run newer versions of software. It is important that plans be in place to deal with these issues if there is an expectation that the CAA will be in place for longer than five years. These issues are raised here because this is an additional cost of implementation. Resources are needed to renew and replace software, hardware, and item pools as time passes. These costs can be substantial.

Is CAA Technology Ready for Application?

The previous section of this paper can be very discouraging when considering the development of a new CAA program. It raises questions about whether or not it is practically possible to develop and implement a large-scale CAA. The short answer to this question is "yes," it is possible, and there are a number of operational implementations already underway. In this section, we will consider current systems to determine if they are good models for implementation in a broader educational setting.

The first of the large-scale CAA programs to be implemented was for the Armed Services Vocational Aptitude Battery (ASVAB). This is the testing program that is used by the U.S. Military to determine the capabilities of persons entering the various branches of the Armed Forces. This testing program is high stakes for the examinees in that it determines whether they will be enlisted at all, as well as the type of assignment they will receive after enlistment. Being the first large scale system of its type, it was very expensive to develop, and years of research were done to work out the details. The full history of the development of this CAA is given in Sands, Waters and McBride (1997). This implementation of a CAA is a model for others, but it probably would not have come into existence without the resources of the Pentagon behind it. Now it is widely regarded as a successful example of CAA implementation, and large numbers of examinees are tested each year. It is notable that it is not a single test, but a full battery of aptitude measures. Separate scores are reported for each part and the different branches of the Armed Forces form different composites of those scores to make acceptance and job placement decisions.

A second example of ongoing implementation of a CAA is for placement into postsecondary courses. Both ACT, Inc. (COMPASS) and College Board (Accuplacer) have developed CAAs to place students in entry level courses at two- and four-year colleges. This program is low stakes for the examinees in that being inaccurately placed is easily corrected and cheating to get higher scores only puts the student in a course that might be beyond the student's current capabilities. These programs currently test large numbers of students through the placement services at postsecondary educational institutions. These programs were developed over relatively short periods of time as new initiatives by the respective companies. The developers dealt with the issue of seat time in front of computers by using the computer facilities at postsecondary institutions, but did produce the software systems and item pools.

A third example is in the area of licensure and certification in professions. One of the longest running programs of this type is for licensing nurses. The NCLEX program is run by the National Council of State Boards of Nursing (NCSBN), but the actual administration is done by a vendor, Pearson View. Pearson View supplies the computer hardware and the testing sites and manages the scheduling of the examinations. It also stores the item pool and works with NCSBN to develop the item pools. This is a particularly high stakes testing program because a person cannot be employed as a nurse unless they pass this examination. Because of the stakes that are involved, this CAA uses very large item pools and frequently changes the item pools. Despite the need for strong security measures, this CAA has been very successful, and it has been in place for many years.

A final example that is especially relevant to state achievement testing programs is the CAA that is implemented by the Northwest Evaluation Association (NWEA). This CAA is designed to assess student growth over grades two to ten. It includes tests in the subject matter areas of reading, mathematics and science. NWEA reports testing over 24 million students over 30 years. This CAA is designed for school-based testing, and it has been in place since about 1977. It is a moderate stakes examination program that takes advantage of computer resources in the schools. NWEA has developed the software system and the training programs to support the implementation of the CAA. School-based staff members carry out the actual implementation.

Although there are a number of success stories for CAAs, there are some notable areas where there have not been successful implementations. Most prominent of these are the two college admission testing programs in the U.S., the ACT and SAT programs. The organizations that run these testing programs have been investigating computer administration for quite a few years, but the investigations have not led to operational administration. One reason for this is the issue of seat time in front of a computer. Each of these testing programs administers over 2,000,000 examinations each year under secure conditions. It is extremely difficult to determine how to get sufficient time in front of computers to administer this many assessments while maintaining high security. These programs may yet move to computer administration in the future, but no plans have been announced at this time.

Common Themes for Successful Implementation

The review of these CAAs yields a number of common themes that led to their successful implementation. All of them took the time in the beginning to do careful planning for development and implementation. Often there were research studies to support the design process. These studies were designed to answer specific practical questions, such as: does the size of the computer screen affect the difficulty of test items? The planning processes for these CAAs often took years. One reason for such long planning times was because CAA developers were charting unknown territory, but there were also many practical issues to be resolved. Even with the experience gained from these CAAs, a reasonable planning period is needed to design a new CAA. It is difficult to imagine a period less than six months to a year.

A second theme was that all of these programs started with pilot testing implementations. These were relatively small scale implementations that could be used to identify problem areas and develop solutions to the problems. Then there was a scaling up of the implementation. As the number of administrations increases, the problems of scale become more evident. How are thousands of data records sent to a central system without overwhelming that system?

When new CAAs are phased in rather than implemented at one time, there is always an issue of whether paper-and-pencil and computerized versions of an assessment can be implemented at the same time and be expected to yield comparable results. The issue of comparability is a very challenging one. There are few or no operational programs that have successfully implemented parallel paper-and-pencil and computerized tests that give comparable scores. The general advice is to avoid trying to run parallel programs. This approach is expensive and the scores from the assessments are not sufficiently comparable to be used interchangeably.

Recommendations

Developing and implementing a CAA is the equivalent of the work required for any other type of technology driven product. These development activities must take into account the capabilities of the technology and how they will change over the life of the product. The development plans also need to anticipate possible problem areas and have the resources available to deal with them when they occur. CAAs are complex systems and there will be challenges to overcome. These challenges may have minor consequences if they are anticipated, or they

may create a major roadblock if they are a surprise, and there are no resources to deal with them. With these challenges in mind, the following steps are recommended for those who plan to develop a CAA.

1. Bring together a development team with expertise in computer hardware, computer software, educational data systems, school administration, and the psychometrics of CAAs. It will take all of these individuals working together to produce a viable design.
2. Develop a thorough plan for design and implementation of the CAA. This plan should include a description of the design process with goals for the level of detail needed in the design and detailed steps for creating all of the components included in the design. Enough time should be allocated to this process that it can be given a critical review to look for possible problem areas.
3. As early as possible in the process, conduct a feasibility study. Will it be possible to get the seat time needed in front of the computers? Are the channels for sending information sufficiently large and reliable that the system will not be overwhelmed by operational use?
4. Develop a prototype system and test it on a small scale to make sure all the component parts work before moving to large scale implementation.
5. Do a larger scale pilot test of the system to determine how it will function with more examinees. If this is a school-based system, make sure the pilot test uses multiple schools and it should include sites that were not involved in design and development.
6. Develop a full-scale plan for implementation with cost estimates and timelines. Until the realities of full implementation are clear, it is not possible to determine if all the component parts will work as planned.

It is possible to design, develop and implement CAAs on a large scale. This is a difficult task, and it should not be undertaken without the resources and expertise that is needed. Yet, most of the challenges to implementation also exist for tests that do not make full use of the computer capabilities—those that are not adaptive. If computer access can be made available, it would seem logical to make use of the full potential of the computer by adding intelligence to the testing process rather than using computers simply to present material.

Final Comments

There is little doubt that computer technology will continue to be used in an educational context. It is more a matter of *when* and *how* computers will be used for assessment rather than *if*. It is important to note, however, that the merger of computers and assessment does not have to be adaptive. There are already many computerized tests that are either fixed tests that have been transferred to computer or tests that are random selections of items from an item pool. These are viable intermediate steps between traditional paper-and-pencil and fully adaptive tests. Simple computerized tests of these types require the development of data systems and provisions to ensure the necessary number of computers, but they sidestep the psychometric complexities of adaptive testing. Of course, they sacrifice the advantages of CAAs as well. The purpose of making this point is not to discourage the development of CAAs, but rather it is to show that the full implementation can be phased in through stages of development.

If there is the luxury of having reasonable computer resources in every classroom, the possibility of totally embedding assessment in instruction becomes quite practical. Assessment tasks and instructional activities become more similar. The test items can include multiple steps that are scored for the information they provide about different types of skills and knowledge. Work on the development of such tasks is being done by a number of researchers (see Bennett (2010) for one example). If all of students' day-to-day classroom work can be collected through computer systems and scored using intelligent evaluation software, the need to have a separate, stand-alone CAA is no longer present. As the development of CAA systems progresses, the goal of many educators—to have instruction and assessment be one and the same—can be met.

References

Bennet, R. E. (2010). Cognitively Based Assessment of, for, and as Learning (CBAL): a preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary research and perspectives,* 8(2-3), 70-91.

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

Davey, T. & Pitoniak, M. J. (2006). Designing computerized adaptive tests. In S. M. Downing and T. M. Haladyna (Eds.) *Handbook of test development.* Mahwah, NJ: Lawrence Erlbaum Associates.

Drasgow, F. & Mattern, K. (2006). New tests and new items: Opportunities and issues. In D. Bartram and R. K. Hambleton (Eds.) *Computer-based testing and the internet:* Issues and advances. West Sussex, England: John Wiley & Sons.

Kennedy, G. E., Judd, T. S., Churchward, A., Gray, K. & Krause, K. (2008). First year students' experiences with technology: Are they really digital natives? *Australasian Journal of Educational Technology,* 24(1), 108-122.

Kifer, E. (2001). *Large-scale assessment: dimensions, dilemmas, and policy.* Thousand Oaks, CA: Corwin Press.

Koretz, D. (2008). *Measuring up: What educational testing really tells us.* Cambridge, MA: Harvard University Press.

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.) *Computer-assisted instruction, testing, and guidance.* New York: Harper & Row.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

National Assessment Governing Board (2010, September). *Writing Framework for the 2011 National Assessment of Educational Progress.* Washington, DC: U.S. Government Printing Office.

Parshall, C. G. (2002). Item development and pretesting in a CBT environment. In C. N. Mills, M. T. Potenza, J. J. Fremer, and W. C. Ward (Eds.) *Computer-based testing: Building the foundation for future assessments.* Mahwah, NJ: Lawrence Erlbaum Associates.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing.* New York: Springer.

Phelps, R. P. (2005). *Defending standardized testing.* Mahwah, NJ: Lawrence Erlbaum Associates.

Rothman, R. (1995). *Measuring up: Standards, assessment, and school reform.* San Francisco: Jossey-Bass.

Shermis, M. D. & Burstein, J. (2003). *Automated essay scoring: A cross-disciplinary perspective.* Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D. & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer *Computerized adaptive testing: A primer.* Mahwah, NJ: Lawrence Erlbaum Associates.

van der Linden, W. J. (2006). Model-based innovations in computer-based testing. In D. Bartram and R. K. Hambleton (Eds.) *Computer-based testing and the internet: Issues and advances.* West Sussex, England: John Wiley & Sons.

van der Linden, W. J. & Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice.* Dordrecht, the Netherlands: Kluwer Academic Publishers.

van der Linden, W. J. & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. In W. J. van der Linden and C. A. W. Glas (Eds.) *Computerized adaptive testing: Theory and practice.* Dordrecht, the Netherlands: Kluwer Academic Publishers.

Wainer, H. (2000). *Computerized adaptive testing: a primer (2nd edition).* Mahwah, NJ: Lawrence Erlbaum Associates.

Strengthening Assessment for English Learner Success: How Can the Promise of the Common Core Standards and Innovative Assessment Systems Be Realized?

Robert Linquanti, WestEd

Introduction

Educational assessment policy must produce measures of performance that are fair and accurate for all students, in order to convey clear, helpful information to educators, parents, and the students themselves. Achieving these objectives is especially challenging when it comes to the nation's five million K-12 public school English learners (ELs). English learners—linguistic minority students not sufficiently proficient in English to be able to benefit adequately from regular classroom instruction and demonstrate their knowledge and abilities using English—constitute one of the fastest-growing student groups in K-12 public schools. At the national level ELs now constitute 10.8 percent of the K-12 population. Approximately 80 percent of them are Spanish-speaking, with the rest speaking a multiplicity of different languages. Just over half are U.S.-born, and almost two thirds are from low-income homes (National Clearinghouse on English Language Acquisition, 2010; Capps et al., 2011; Migration Policy Institute, 2010). In California more than half of the children now entering public schools come from households where the first language is not English.

ELs need to accomplish two key objectives in their schooling: language proficiency in English and achievement in grade-level subject matter across the curriculum. Civil rights statutes and case law including the Supreme Court's 1974 Lau v. Nichols decision affirm educators' dual obligations to these students: 1) Ensure they develop academic English-language proficiency, and 2) Ensure they have meaningful access to grade-level content via appropriate instruction. These are *interrelated* and therefore simultaneous (not sequential) obligations.

English learners are expected to meet two sets of standards—those for academic content under ESEA Title I, and those for English-language proficiency (ELP) under Title III. Title III of the No Child Left Behind Act of 2002 specifically requires each state to have in place ELP standards for ELs that are aligned to the state's content area standards, including the proficient academic performance standard. The federal government's just-announced grant program for next-generation ELP assessment systems—based on ELP standards that properly correspond to the Common Core State Standards (CCSS)—clearly reinforces this need for alignment. In fact, the ELP summative assessment is intended to indicate the point at which ELs attain a level of academic English "necessary to participate fully in academic instruction in English and meet or exceed college- and career-ready standards" (Federal Register, 2011, page 21978).

The interrelationship of the two goals of proficiency in academic content and the English language poses significant challenges to current assessment and accountability policies. First, academic assessments that fail to take account of ELs' English-language proficiency level will likely inadequately measure their content area knowledge and skills, both individually and as a subgroup. Students with limited proficiency in English often underperform on assessments of academic content, reflecting not a lack of knowledge but a lack of fluency, which may unfairly depress their scores (Abedi & Gandara, 2006). Second, current accountability

policies distort the overall EL cohort's academic performance and obscure long-term outcomes by removing those who achieve English proficiency from the EL subgroup. As a result, reporting of subgroup academic performance is limited to those EL students who are *by definition* the lowest performing.

The adoption and implementation of CCSS and the launching of the two comprehensive assessment consortia, Partnership for the Assessment of Readiness for College and Careers (PARCC) and SMARTER Balanced Assessment Consortium (SBAC), present a major opportunity to dig deeply into the challenges involved in fairly and accurately assessing ELs' academic performance, but the road ahead is long and, for now, only poorly mapped. The issues are complex, and the assessment technologies that might help to address them are relatively young. Nevertheless, the stakes are enormously high for these students, and the challenges must be engaged now. How these systems support (or neglect) ELs will depend largely on how we develop and implement next-generation assessment systems, and relate them to instruction, professional development and accountability.

This policy brief describes how next-generation state assessment and accountability systems can be made more responsive to the needs and strengths of English learners. Specifically, the brief argues that innovation must be grounded in a clear understanding of the EL population, as well as of English-language proficiency and its relationship to academic subject matter learning and assessment. It describes some key considerations in assessing ELs, and in using assessment to strengthen teacher pedagogical practice with ELs. It then makes a case for how comprehensive assessment systems can be more responsive to ELs' needs. Finally, the brief recaps lessons learned from EL access and accommodations research, sketches emerging technologies, and offers suggestions for coordinating ELP and academic assessment development efforts in order to improve their validity and utility.

Fundamental Considerations in Assessing English Learners

Though often referred to as if they were a monolithic group, English learners are diverse in ways that have important implications for both instruction and assessment. Over 80 percent speak Spanish, while the rest speak many different languages, and ELs exhibit a wide range of language proficiencies in both their primary language and in English. Some EL students have beginning-level English-language proficiency, while others are at intermediate or more advanced levels. Their linguistic skills often vary by language domains and functions, with students frequently demonstrating greater proficiency in listening and speaking[1] than in literacy (reading and writing) skills. Most have been in U.S. schools since kindergarten (Migration Policy Institute, 2010), but many arrive much later and with varying levels of prior formal schooling and first language literacy. Some come as refugees with interrupted or limited formal education, while others come from highly privileged backgrounds with advanced academic knowledge. Still others come as linguistic and cultural minorities from their home countries, and their socio-cultural distance from common U.S. schooling practices may be substantial. A large majority of ELs are also low-income, which may further affect their opportunities to learn and their linguistic and academic performance (Capps et al., 2007; Hakuta et al., 2000).[2]

What Constitutes English Learner Status

English learner status is expected to be *temporary*. Students are supposed to be removed from the category

1 At least as defined in current ELP standards; however, the Common Core standards call out much more rigorous academic language capacities in these domains.

2 Hakuta et al. (2000) note that ELs from lower income backgrounds progressed more slowly in their academic language development relative to their higher income EL peers.

as a result of effective, specialized language instruction and academic support services that they are required to receive. No other subgroup of students is defined in this temporary, instructionally-dependent way.

In practice, EL status is operationalized through various processes, instruments, and performance standards. Typically, students identified as language-minority (from a home where a language other than or in addition to English is spoken) take a brief screener assessment to determine their level of English-language proficiency. Those identified as ELs are provided instructional services to develop their English proficiency and support their access to grade-level content until they are deemed ready to reclassify and exit the category. Exit from EL status may involve more than language proficiency, however. A recent national review of EL classification and reclassification practices reveals that while all states use ELP measures to determine English-language proficiency and readiness to exit the EL category, three quarters of the states also use one or more academic performance criteria in their exit decisions (Wolf et al., 2008). Complicating matters further, several states allow local school districts to determine exit criteria, meaning that EL definitions vary *within* these states (National Research Council, 2011; Ragan & Lesaux, 2006).[3]

The most linguistically and academically accomplished students exit the EL category over time, while those not making sufficient progress remain and are joined by newly entering ELs who are by definition at lower levels of language proficiency. NAEP and most state-level assessments typically ignore this "revolving door" phenomenon, and therefore systematically underestimate EL subgroup results by excluding the higher-performing reclassified ELs. In partial recognition of this problem, NCLB regulations allow states to count students as part of the EL subgroup for accountability purposes for up to two years after they exit EL status. Yet this adjustment does not solve the problem. Since many students exit at upper elementary or in early secondary grade levels, they are absent from cohort analyses examining longer-term outcomes such as graduation rates and college and career course taking. These revolving door practices wrongly stigmatize the EL subgroup, demoralize students and teachers, distort the performance picture, and prevent examination of long-term outcomes for the cohort. Accurate representation of EL cohort performance would include the performance of all students ever included in the category for as long as they remain in the system.

A related underlying issue is that assessment and accountability systems generally treat the EL category as binary (a student is EL or not), when in fact EL students exhibit language competencies on a continuum that extends from the lowest levels of English proficiency through exit-level performance standards and beyond. Moreover, language proficiency becomes increasingly complex as students move through school. Proficiency in English at age six is very different from proficiency at age sixteen, and the language demands of academic subject matter increase substantially with grade level. Even reclassified ELs may need opportunities and support to continue developing their academic language proficiency and content area knowledge and skills.

Policy Implications of ELP's Relationship to Academic Performance

English-language proficiency has been defined as "language ability across relevant modalities (i.e. listening, speaking, reading, writing) used at sufficient levels of sophistication to successfully perform all language-related school tasks required of students at a specific grade level (given adequate exposure and time to acquire the second-language)" (Bailey & Heritage, 2010, pages 2-3).

Students also need discipline-specific academic language competencies that directly affect their ability to demonstrate academic subject area knowledge and skills. For example, effective argumentation in science—considered fundamental to mastery—carries with it very specific vocabulary, grammatical forms, and collaborative discourse patterns that teachers need to explicitly highlight and model and students need to practice through carefully structured interactions (Osborne, 2010; Schleppergrell, 2004).

It takes time for ELs to learn the academic English skills they need. Their progress will vary based on many factors, including their initial English proficiency and

3 These states include California, Florida, and Texas.

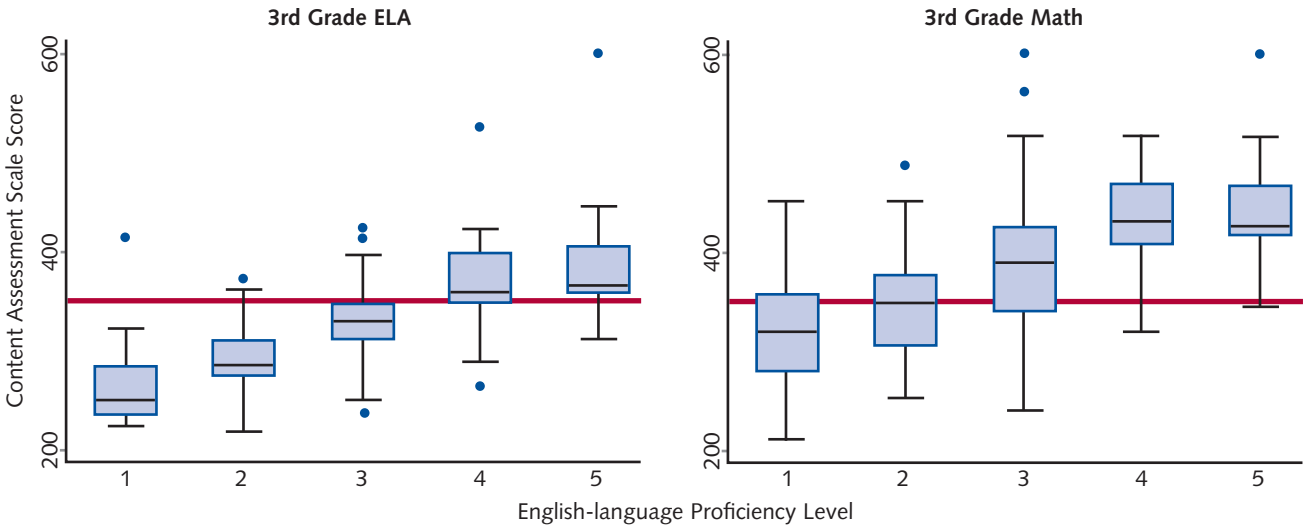
age on entry, opportunity to learn (i.e. the instructional delivery system and conditions for learning), and the language demands inherent in the content area and grade level. The best available empirical evidence suggests that ELs require roughly four to seven years to gain the academic language competencies needed to successfully handle complex, grade-level content demands without specialized support services (Hakuta et al., 2000; Cook & Zhao, 2011). Because of the extended time required to learn sufficient English, ELs are likely to be inaccurately assessed on their content knowledge as measured by standardized academic assessments.

An EL's level of English-language proficiency affects his or her ability to learn academic content taught in English, and to demonstrate academic knowledge and skills on assessment events carried out in English. In a pattern widely seen across states, Figure 1 (from Thompson, 2011) illustrates the distribution of a California school district's EL student performance on the state's third-grade English-language arts and math assessments as a function of EL students' English-language proficiency level.[4] EL students' academic performance clearly increases with increasing English proficiency. The level of language proficiency needed to demonstrate grade-level performance varies by subject, with more students attaining the state's grade-level performance standard in mathematics at a lower level of English proficiency when compared to English-language arts. In addition, the overlap in performance of ELs at different ELP levels suggests that English proficiency is necessary but not sufficient to explain academic performance. Students bring (and are provided) different resources that also affect their scores on assessments.

These facts have two clear implications for assessment policy. First, English-language proficiency is foundational to ELs' academic success. Since so many academic tasks are mediated by language, academic language skills are central to performing sophisticated content area tasks, and should be defined, taught, and measured explicitly. *All* teachers must be prepared to model, teach and provide students opportunities to develop the language of their academic disciplines, including the specialized vocabulary, sentence-level structures, and discourse patterns that second language learners must master.

Second, it is essential to disaggregate ELs' academic performance by their English-language proficiency level, and to examine their English-language proficiency growth over time in the education system. If a student is performing poorly on an academic content assess-

Figure 1. Distribution of EL student performance on academic assessments in English-language arts and math by English proficiency level, third grade.



4 Analyses of EL academic performance by ELP level have been performed across several states and yield similar performance patterns. (Francis & Rivera, 2007; Parker et al., 2009; Cook et al., forthcoming)

ment, we need to understand whether this is due to: 1) insufficient academic language proficiency to demonstrate content knowledge; 2) a lack of content knowledge or opportunity to learn content; 3) construct-irrelevant interference (e.g. unnecessarily complex language in the assessment); or 4) other sources of bias or error (e.g. cultural distance, dialectal variation, rater misinterpretation, etc.).

Policy Implications of the Common Core State Standards for ELs

The Common Core State Standards (CCSS) in English-language arts specify, to an unprecedented degree, the kinds of academic language that students need to use to demonstrate subject-matter mastery, and ultimately readiness for college and careers. For example, the K-5 reading standards require students to manifest their knowledge and comprehension through explaining, describing, comparing and contrasting, arguing, giving definitions, giving recounts, summarizing and paraphrasing, and explaining cause-and-effect. Third-graders are expected to "recognize and observe differences between the conventions of spoken and written Standard English," while eleventh-graders are expected to "propel conversations by posing and responding to questions that probe reasoning and evidence" (Common Core State Standards Initiative, 2010, pages 29, 50).

The standards also define discipline-specific literacy standards for history/social studies, science, and technical subjects at the secondary level. Figure 2 highlights what students in ninth and tenth grades are expected to do in understanding and generating scientific texts. There are clear academic literacy and language demands associated with these abilities, including analyzing and summarizing hypotheses and explanations; following directions exactly; inferring relationships among terms, processes and concepts; and comparing and contrasting information.

The new standards set a high bar for demonstrating content knowledge and skills, which has two powerful implications for the instruction and assessment of English learners. First, it is clear that sophisticated academic language competencies are necessary and central to performing content area tasks called for in the standards. Educators, curriculum developers, and test developers will need to clearly map out the language

Figure 2. Reading and Writing Standards for students in ninth and tenth grades (Common Core State Standards Initiative, 2010, pages 62, 64).

Reading standards for literacy in science and technical subjects:

3. Follow precisely a complex multistep procedure when carrying out experiments, taking measurements, or performing technical tasks attending to special cases or exceptions defined in the text.
7. Translate quantitative or technical information expressed in words in a text into visual form (e.g. a table or chart) and translate information expressed visually or mathematically (e.g. in an equation) into words.
9. Compare and contrast findings presented in a text to those from other sources (including their own experiments), noting when the findings support or contradict previous explanations or accounts.

Writing standards for literacy in history/social studies, science, and technical subjects:

1. Write arguments focused on *discipline-specific* content.

a. Introduce precise claim(s), distinguish the claim(s) from alternate or opposing claims, and create an organization that establishes clear relationships among the claim(s), counterclaims, reasons, and evidence.

b. Develop claim(s) and counterclaims fairly, supplying data and evidence for each while pointing out the strengths and limitations of both claim(s) and counterclaims in a discipline-appropriate form…

c. Use words, phrases, and clauses to link the major sections of the text, create cohesion, and clarify the relationships between claim(s) and reasons, between reasons and evidence, and between claim(s) and counterclaims.

d. Establish and maintain a formal style and objective tone while attending to the norms and conventions of the discipline in which they are writing.

demands these standards entail, and define meaningful benchmarks of academic language performance that correspond to levels of mastery in the academic content standards themselves. Making these academic language expectations explicit will help teachers identify the language skills they need to foster in their EL students. It will also help professional development providers, curriculum developers, and assessment developers to support teachers in providing students with opportunities to develop these language competencies, and in knowing where ELs are on a trajectory toward mastering them.

Second, existing English-language proficiency standards must be revisited and aligned to the academic language demands articulated in the Common Core standards. This will help to strengthen linkages between the ELP assessment system and the academic content assessment system. Specifying the breadth, depth, and complexity of academic language skills and functions that need to be taught, learned and assessed is essential for ELs' success. Explicitly developing and testing models that address these alignment and linkage challenges will help to strengthen assessment validity and utility for ELs.

How Comprehensive Assessment Systems Can Be More Responsive to EL Needs

Both multistate consortia are expected to develop comprehensive academic assessment systems that 1) deliver *summative assessments* of cognitively complex, college- and career-ready knowledge and behaviors to be used for program review and accountability purposes; 2) provide timely, useful *interim benchmark assessments* at key intervals during the school year to help predict outcomes and guide interventions; and 3) directly inform, support and enhance teacher practice and student learning through *formative assessment* practices, tools and processes.

A judicious balance among these three dimensions will be required to make these assessment systems responsive to EL strengths and needs. The opportunities and challenges associated with each dimension follow, beginning with the dimension traditionally least utilized, yet most promising for improving EL outcomes.

Formative Assessment: Assessment *For* and *As* Learning

Formative assessment is "a *process* used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes" (Heritage, 2010, page 9). Formative assessment occurs *within* instruction through informal observations, conversations, and other carefully planned, instructionally embedded methods that allow teachers to gather evidence of student learning, spur student reflection, and offer hints and cues to move students forward in understanding and acquiring skills and knowledge. Formative assessment practices have enormous potential to strengthen teachers' capacities to developmentally stage or "scaffold" ELs' language and content learning.

Developing formative assessment *for* and *as* learning is the most important element of an effective assessment system for ELs. It makes little sense to enhance standardized large-scale assessments for ELs without simultaneously investing in and fostering their learning.

The inequitable distribution of instructional resources to address EL instructional needs has been amply documented (Taylor et al., 2010; Gandara et al., 2003). The preparation, coaching, and ongoing professional development of all teachers of ELs are the most important investments to be made in improving educational outcomes for ELs (Wong-Fillmore & Snow, 2002). Correspondingly, formative assessment is the most important aspect of the comprehensive assessment system to get right for English learners because it is the *most instructionally relevant.*

Fortunately, there has been a steady evolution of EL-relevant formative language assessment practices and tools including pilot academic content learning progressions and associated language learning targets; prototype performance tasks and instructional supports linked to

the tasks; and professional development models that systematically build teachers' capacities to evaluate whether EL students have access to and are accomplishing language and content objectives that indicate progress toward larger instructional goals.[5] Integrating these into future assessments is essential if we are to accurately and fairly assess the academic performance of ELs while they develop English-language proficiency.

Interim/Through-course Assessments: Matching Intended Uses with Best Practices

In the design of interim or through-course assessments, it is critical to identify the language learning targets that correspond to the curricular material to be taught and ensure that students receive instruction that corresponds to these targets in the time covered by the assessment. Since EL students' language competencies develop throughout the school year, there may be differential opportunities to learn and demonstrate subject matter knowledge occurring *within* the school year that are manifested in interim assessment outcomes, particularly those from earlier in the academic year. How these interim assessments are weighted could misrepresent EL student results because of the potential mismatch between the results-aggregation method and how EL students are actually learning (Wise, 2011).[6]

Summative, Large-scale Assessment: Strengthening Signals, Managing Expectations

The No Child Left Behind Act mandated the participation of ELs in states' large-scale testing systems. The law requires that EL students be "assessed in a valid and reliable manner and provided reasonable accommodations... including, to the extent practicable, assessments in the language and form most likely to yield accurate data on what such students know and can do in academic content areas, until such students have achieved English language proficiency" (Elementary and Secondary Education Act, Sec.1111 (b) (3) (c) (ix) III).

NCLB got well ahead of the capacity of existing content assessments to accurately measure what students know and can do in a subject area when tested in a language they are still learning. This disjunction between policy requirements and the field's capacity to meet them has provoked serious issues regarding the validity and reliability of ELs' scores, and produced unintended negative consequences for students and educators (Abedi, 2004; Linn, 2005; Hakuta & Linquanti, 2011). The next generation of assessments will have to address these concerns by greatly expanding and refining the portfolio of EL-responsive access and accommodation practices. Even incremental validity gains are valuable.

What Have We Learned from EL Accommodations Research?[7]

Accommodations are changes to a test or testing situation intended to improve student access to tested content (e.g. students' understanding of what is being asked, or their opportunity to convey what they know) without altering the construct measured and the validity of inferences drawn from the test. An EL-responsive accommodation must therefore be *effective* with ELs (i.e. demonstrate substantial performance improvement), and also *valid* (i.e. demonstrate an absence of advantage for non-ELs) in order to help level the linguistic/cultural playing field (Kieffer et al., 2009).

Efforts to more effectively accommodate English

5 See the FLARE (Formative Language Assessment Records for English-Language Learners) project at http://flareassessment.org ; WestEd's Quality Teaching for English Learners program at http://www.wested.org/cs/tqip/print/docs/qt/home.htm; Bailey & Heritage, 2010; Heritage, 2008; Sato, 2008.

6 While educators may want to use interim assessment results formatively (e.g. to change instruction, to identify students in need of additional support), these can have predictive and evaluative purposes not easily reconciled with formative purposes. See Goertz et al., 2009.

7 This accommodations research review draws from Abedi et al., 2004; Duran, 2008; Kieffer et al., 2009; and Pennock-Roman & Rivera, (in press).

learners on large-scale tests have faced increasingly stringent expectations for demonstrating the validity of inferences made using accommodated scores, and current practices fall far short of the necessary standard. A National Research Council committee recently concluded that existing research on accommodations for ELs and students with disabilities was inadequate to demonstrate the comparability of inferences from scores derived from accommodated versus standardized tests, or to demonstrate which accommodations produce the most valid estimates of performance (National Research Council, 2004).

Disturbingly, less than 40 percent of the accommodations used by states with ELs actually address these students' linguistic needs. The effectiveness of the single most commonly provided English language accommodation for ELs—reading instructions aloud—is unproven (Shafer Willner et al., 2008). Distinguishing and deploying accommodations that provide direct and indirect linguistic support to ELs in testing situations is a critical priority for both policy and practice (Rivera et al., 2006).

English learners are heterogeneous in ways that can measurably influence the effectiveness of particular accommodations. Using a decision algorithm to assign configurations of accommodations tailored to ELs' linguistic and socio-cultural characteristics may yield better performance results than providing all available accommodations or no accommodations (Kopriva et al., 2007). A recent meta-analytic review by Pennock-Roman & Rivera (in press) offers insights on the most prominent and promising accommodations by EL characteristics and testing conditions, and these findings are briefly reviewed next.

English-language Linguistic Supports

English dictionaries/glossaries clarify key vocabulary not directly tied to the construct being measured. Paper-and-pencil versions have been found to be effective when ELs are given extra time, while pop-up English glossaries (used in online testing) are more helpful under restricted time constraints (Pennock-Roman & Rivera, in press).

Plain English (also known as simplified English) attempts to reduce the language load from test items and tasks by removing construct-irrelevant language complexity. This accommodation approach has yielded mixed results, sometimes with effectiveness, at other times with validity (Duran, 2008). There are several reasons why this could be the case. For example, many test developers are now more sensitized to unnecessary linguistic complexity in item development, which may be reducing the gains produced by plain-English-accommodated versions of items. Moreover, non-EL comparison groups may include former ELs with ongoing linguistic needs who also benefit, which could influence comparison statistics. There may also be limits to this strategy because academic language is inextricably connected to more complex content knowledge, and necessary academic language skills contribute to EL/non-EL achievement differences much more than does the construct-irrelevant language addressed by the accommodation (Kieffer et al., 2009).

Primary-language Related Linguistic Supports

Several challenges exist in using primary-language related accommodations. Apart from equity concerns raised by not addressing the large number of different languages spoken by non-Spanish-speaking ELs, these include substantial linguistic and psychometric issues with test translation, which—notwithstanding well-designed, simultaneous test development in two languages—usually undermines construct validity. They also include the existence of multiple varieties of a given primary language (including Spanish), which display notable differences in vocabulary and syntax, along with cultural factors intertwined with language (Solano-Flores & Trumbull, 2008; Solano-Flores, 2006).

To strengthen validity, the language of assessment should match the language of instruction. So students instructed bilingually may need to be accommodated based on their ELP level as well as on the goals of the instructional program in which they are enrolled. ELs vary in their primary-language proficiency for academic purposes, and not all can better demonstrate knowledge on primary-language assessments. Moreover, the goals of the academic instructional program are crucial to consider. For example, students in developmental bilingual education (which aims for academic achievement in two languages), may need to be tested in different

languages depending upon the subject matter and their expected language proficiency level. Students in a transitional bilingual education program (where the primary language is used only temporarily to teach and learn academic content until English-language proficiency is sufficiently developed) may need to be assessed based on academic subjects taught in their primary language in one grade, then in English in another grade per language transition expectations and timeframes.

Primary-language versions of test items are promising, but only for native speakers at low ELP levels including recent immigrants, or for those who are receiving academic instruction in their home language while learning English. This accommodation has been found to be more difficult than the original test in English for ELs at intermediate ELP levels and for those receiving academic instruction in English.

Dual language formats (e.g. parallel bilingual versions of the test booklet), and **bilingual glossaries** (providing primary-language equivalent terms without definitions) show promise if there are generous time limits given to help students utilize and benefit from additional materials and to address situations in which students have developing and not always overlapping capacities in two languages.

What EL-relevant Assessment Innovations Are on the Horizon?

Multi-semiotic Approaches

Substantial assessment development work is being done in online formats with an eye toward increasing access in *conveying information to* and *receiving information from* ELs at lower ELP levels. For example, multi-semiotic approaches (which use interactive schema, graphic/representational models, animations, and computer simulations of real-life contexts) appear promising for accessing the science and math knowledge of students at the lowest ELP levels (Kopriva, 2011).

Such approaches attempt to minimize language as the primary conveyor of meaning (often in conjunction with additional accommodations such as English or bilingual pop-up glossaries) for the purpose of measuring cognitively complex academic constructs. There is some evidence that these approaches may reliably minimize language interference and provide better estimates of students' conceptual and procedural knowledge and skills in math and science.[8] This is encouraging not only for the possibility of more accurate and valid information on what students at beginning ELP levels know, but also because it signals to educators that ELs at *all* stages of language development can learn and be assessed in academic subjects.

The key educational objective is still to ensure that EL students develop the academic language proficiency needed to fully engage with grade-level tasks in different academic disciplines, and to help them to accelerate academically while doing so. Such "language-minimizing" accommodations must therefore be understood and utilized as temporary strategies to better measure EL students' knowledge while they develop the language competencies required by the CCSS. Otherwise, particularly in states allowing accommodations in assessment that are provided during instruction, they may signal to teachers that ELs' language development is not essential to their learning and demonstrating academic content knowledge, and contribute further to ELs' instructional marginalization.

Computer Adaptive Testing

In computer adaptive testing, online testing formats present students with test questions of a level of difficulty that is continually adjusted based on how the student has answered previous questions. EL assessment experts have expressed hope that such technologies may

8 A prominent example of their use with ELs in science and math assessments, with accompanying research evidence, can be viewed at www.onpar.us.

be able to more accurately estimate EL students' content knowledge while also increasing the efficiency of assessment (i.e. reducing testing time), and reducing the stigma and demoralization that occurs for students when they are unable to answer many questions beyond their current performance level.

Some thorny issues arise for English learners facing these test formats, however. For example, the language load of a given question may not be adequately determined relative to its content complexity. (The question could include multiple embedded clauses, complex vocabulary, or passive voice constructions that obscure the meaning of what is being asked.) If an EL student answers the question incorrectly due to limited language proficiency or the item's construct-irrelevant linguistic complexity, and not for lack of content knowledge, the algorithm may shift to easier items and systematically underestimate the student's content knowledge. Since the goal is to more reliably and efficiently estimate the student's content knowledge, it will be particularly critical to review the language load of items of equivalent content complexity, and to ensure that the English learner's ELP level (particularly on literacy) is known and becomes part of the algorithm that assigns test questions. In this way, test items of equivalent construct difficulty, but with different levels of linguistic complexity, can be assigned to EL students at different ELP levels.

Test developers have also been working to create automated scoring routines to enable computer scoring of short essays and constructed responses. These artificial intelligence engines are trained on exemplars at various levels of performance. A concern emerges if the exemplars do not reflect the full range of writing features, including those characteristic of English learners at various levels of English-language proficiency. ELs at different ELP levels may exhibit "inter-language" grammatical or vocabulary errors that are typical of different stages of second language acquisition. They may also use different narrative and discourse patterns, and varying sentence and paragraph lengths, all of which could be misinterpreted in automated scoring methods. Addressing this concern may require the development of specialized scoring routines for use with ELs, trained to recognize common inter-language features, and provide more careful analysis of performance by students' ELP levels.

Where (and How) Do ELP Assessment Systems Fit?

As argued throughout this brief, the next generation of English-language proficiency assessments will need to be based on ELP standards that sufficiently specify the target academic language competencies that ELs need in order to progress in and gain mastery of the Common Core standards. This is a critically important goal, but it will be made more challenging as funding for ELP assessment is relatively low and restricted in scope. For example, no ELP enhanced assessment grant funds may be used to develop ELP standards better aligned to the Common Core, and the proposed ELP assessment systems are only required to include "diagnostic" tools (i.e. screeners and placement instruments) and a summative assessment. There is no requirement to include formative assessment. The Common Core State Standards are assumed to provide the link between the academic assessments being developed by PARCC and SBAC and the ELP assessment development initiatives, but there is no explicit framework for how these efforts will be coordinated. An additional challenge is that states participating in any consortium—PARCC, SBAC and/or an ELP consortium—are required by the federal government to adopt a "common definition of English learner" (Federal Register 2010, page 18177; Federal Register 2011, page 21978).

The failure to recognize the integral connections across these assessment initiatives will inevitably generate a number of difficult policy issues. For example, assuming that there are at least two ELP assessment consortia, how comparable will their respective ELP standards and assessment results be? What happens if states in the same ELP assessment consortium participate in different academic assessment consortia (PARCC or SBAC), or vice versa? Will states be permitted to use academic achievement results from PARCC or SBAC in their common EL definition? Will states using different ELP assessment

results in their common EL definition be able to demonstrate sufficient alignment and linkage to the academic assessments? Will ELP assessment results be expected to have comparable predictive validity with respect to the academic performance standards of both PARCC and SBAC? What path will federal accountability policy for ELs take while these issues are worked out? Will it constrain or support innovative improvements? Addressing these questions will require close collaboration across and within all assessment consortia, as well as careful coordination with state and federal policymakers.

Moving Forward

Developing fair and accurate assessments for English learners poses serious challenges for education policy. These challenges must be engaged *from the very beginning and at every stage in the development* of the comprehensive assessment system, from clarifying the validity arguments to be made through item and task development to field testing, educator professional development and technical assistance.

There are some immediate steps that policymakers can take to help ensure that assessments of ELs' academic content knowledge move toward greater accuracy and fairness.

1. The U.S. Department of Education can require and ensure close collaboration among the federally funded academic and ELP assessment consortia. Such collaborative efforts can strengthen communication, experimentation, EL subpopulation data collection and analysis, and prototyping ELP and academic assessment tasks to yield more aligned, coherent and useful information for all stakeholders in the system.
2. Accommodations, while not a panacea, must be strengthened. The academic assessment consortia states can define access and accommodation strategies in relation to EL students' specific linguistic capacities and needs. This includes configuring sets of accommodations better tailored to meet the needs of ELs with different profiles; and developing computer adaptive technologies to better measure ELs' cognitive processes and content knowledge at different ELP levels.
3. The consortia can invest heavily in formative assessment processes and practices, tools and tasks to support content area teacher practices with English learners. This requires carefully mapping out key academic language competencies and target language uses to meet the CCSS at different levels of academic performance, and to ensure these academic language competencies are articulated in language learning progressions reflected in ELP standards and in ELP assessment specifications.
4. As ESEA is reauthorized, lawmakers can work to ensure that ELP progress expectations are related to students' time in the system, and that academic progress is examined in relationship to ELP progress over time. The relationship between English-language proficiency development and academic progress and attainment is nuanced and interconnected. Our educational assessment and accountability policies and practices should be as well.

Acknowledging and overcoming the challenges involved in fairly and accurately assessing ELs is integral and not peripheral to the task of developing an assessment system that serves *all* students well. This is clearly true in states where linguistic-minority students constitute a major presence in public schools, including California, but it is equally true in states across the country. The academic language demands inherent in the CCSS will challenge many students, not just non-native English speakers. Insights gained in developing and implementing comprehensive assessment systems that are responsive to ELs will benefit all students and teachers. Treating the assessment of ELs as a separate problem—or, worse yet, as one that can be left for later— calls into question the basic legitimacy of assessment systems that drive high stakes decisions about students, teachers, and schools. These issues should be at the center and not on the margins of debate about assessment policy.

Acknowledgements

The author thanks Jamal Abedi, Diane August, Alison Bailey, Gary Cook, Patricia Gandara, Kenji Hakuta, and Jennifer O'Day for helpful feedback; Charlene Rivera for sharing in-press research; Karen Thompson for providing Figure 1; and David Plank for his sound guidance and editorial support. All errors and omissions remain those of the author.

References

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

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

Abedi, J., & Gandara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practice,* 25 (4), 36-46.

Bailey, A. & Heritage, M. (2010). *English-language proficiency assessment foundations: External judgments of adequacy.* Evaluating the validity of English-language proficiency assessments. An Enhanced Assessment grant (EVEA, CFDA 84.368). Washington DC: EVEA Project.

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

Common Core State Standards Initiative (2010). *Common core state standards for English language arts & literacy in history/social studies, science, and technical subjects.* Retrieved from: http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf.

Cook H.G., & Zhao, Y.G. (2011). *How English language proficiency assessments manifest growth.* Paper presented at annual meeting of the American Educational Research Association, New Orleans, LA.

Cook, G., Linquanti, R., Chinen, M., & Jung, H. (forthcoming). National evaluation of Title III implementation: *Exploring approaches to monitoring English learner linguistic and academic progress and attainment.* Washington, DC: American Institutes for Research.

Duran, R. (2008). *Assessing English-language learners' achievement. Review of Research in Education,* (32) 292-327.

Elementary and Secondary Education Act. PL. 107. Sec.1111 (b) (3) (c) (ix) III (2001).

Federal Register 75(68), 18171-18185. April 9, 2010.

Federal Register 76(75), 21978-21984. April 19, 2011.

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

Gándara, P., Rumberger, R., Maxwell-Jolly, J. & Callahan R. (2003). *English learners in California Schools: Unequal Resources; Unequal Outcomes.* Educational Policy Analysis Archives. Retrieved from: http://.epaa.asu.edu/epaa/v11n36/.

Goertz, M., Nabors Olah, L., & Riggan, M. (2009). *Can interim assessments be used for instructional change?* Policy Brief RB-51. Philadelphia, PA: Consortium for Policy Research and Education.

Hakuta, K., Butler, Y., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* University of California Linguistic Minority Research Institute Policy Report 2000-1. Santa Barbara, CA: University of California Linguistic Minority Research Institute.

Hakuta, K. & Linquanti, R. (2011). *Improving educational outcomes for English language learners: Recommendations for ESEA reauthorization.* Policy Brief (March 25). Palo Alto, CA: Working Group on ELL Policy.

Heritage, M. (2008). *Learning progressions: Supporting instruction and formative assessment.* Washington DC: Council of Chief State School Officers.

Heritage, M. (2010). *Formative Assessment and Next-Generation Assessment Systems: Are We Losing an Opportunity?* Washington DC: Council of Chief State School Officers.

Kieffer, M.,Lesaux, N., Rivera, M. & Francis, D. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analyses on effectiveness and validity. *Review of Educational Research,* 79(3), 1168-1201.

Kopriva, R. (2011). *The promise of demonstration-based interactive test task environments for struggling readers and English learners.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Kopriva, R., Emick, J., Hipolito-Delgado, C., & Cameron, C. (2007). Do proper accommodation assignments make a difference? Examining the impact of improved decision making on scores for English language learners. *Educational Measurement: Issues and Practice,* 26(3), 11-20.

Linn, R. L. (2005). Conflicting demands of No Child Left Behind and state systems: messages about school performance. *Educational Policy Analysis Archives,* 13(33).

Migration Policy Institute. (2010). *ELL information Center.* Retrieved from: http://www.migrationpolicy.org/news/2010_8_17.php.

National Clearinghouse for English Language Acquisition (2010). The growing numbers of English learner students (1998/99-2008/09). Washington DC: National Clearinghouse for English Language Acquisition.

National Research Council (2004). *Keeping score for all: The effects of inclusion and accommodation policies on large-scale educational assessments.* Judith Koenig and Lyle Bachman (Eds.). Washington DC: National Academies Press.

National Research Council (2011). Allocating federal funds for state programs forEnglish-language learners. Panel to review alternative sources for the limited English proficiency allocation formula under Title III, Part A., *Elementary and Secondary Education Act,* Committee on National Statistics and Board Testing and Assessment. Washington DC: National Academies Press.

Osborne, J. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *Science,* 328, 463-466.

Parker, C., Louie, J., and O'Dwyer, L. (2009). *New measures of English language proficiency and their relationship to performance on large-scale content assessments.* (Issues & Answers Report, REL 2009-No. 066). Washington DC: U.S. Department of Education, Institute of Education Sciences.

Pennock-Roman, M. & Rivera, C. (in press). Test Accommodations for English Language Learners: A Meta-Analysis of Experimental Studies. *Educational Measurement: Issues and Practice.*

Ragan, A., & Lesaux, N. (2006). Federal, state, and district level English language learner program entry and exit requirements: Effects on the education of language minority learners. *Education Policy Analysis Archives,* 14(20). Retrieved from: http://epaa.asu.edu/epaa/v14n20/.

Rivera, C., Collum, E., Shafer Willner, L., & Sia Jr., J. (2006). An analysis of state assessment policies addressing the accommodation of English language learners. In C. Rivera, & E. Collum (Eds.) A national review of state assessment policy and practice for *English language learners.* Mahwah, NJ: Lawrence Erlbaum Associates.

Sato E. (2008). *Language for achievement: Language demands and complexity taxonomy.* San Francisco, CA: WestEd.

Schleppergrell, M. (2004). *The language of schooling: A functional linguistics perspective.* Mahwah: Lawrence Erlbaum Associates.

Shafer Willner, L., Rivera, C., & Acosta, B. (2008). *Descriptive study of state assessment policies for accommodating English language learners.* Arlington, VA: George Washington University Center for Equity and Excellence in Education.

Solano-Flores, G. (2006). Language, dialect, and register: Sociolinguistics and the estimation of measurement error in the testing of English language learners. *Teachers College Record* 108(11), 2354-2379.

Solano-Flores, G., & Trumbull, E. (2008). In what language should English-language learners be tested? In R. J. Kopriva (Ed.), *Improving testing for English Language learners.* New York: Routledge.

Taylor, J., Stecher, R., O'Day, J., Naftel, S. & LeFloch, K.C. (2010). *State and Local Implementation of the No Child Left Behind Act,* Volume IX—Accountability under NCLB: Final Report. Washington, DC: U.S. Department of Education.

Thompson, K. (2011). Personal communication (April 27).

Wise, L. (2011). *Picking up the pieces: Aggregating results from through-course assessments.* Princeton, NJ: Center for K-12 Assessment & Performance Management at ETS.

Wolf, M.K., Kao, J., Herman, J., Bachman, P., Chang, S., & Farnsworth, T. (2008). *Issues in assessing English-language learners: English-language proficiency measures and accommodation uses.* Practice Review. CRESST Report 732. Los Angeles, CA: UCLA.

Wong-Fillmore, L., & Snow, C. E. (2002). What teachers need to know about language. In C. T. Adger, C. E. Snow, & D. Christian (Eds.). *What teachers need to know about language.* Washington DC: Center for Applied Linguistics.

Next Generation Assessments for Measuring Complex Learning in Science

Jody Clarke-Midura and Chris Dede, Harvard University
Jill Norton, Rennie Center for Education Research & Policy

One thing I never want to see happen is schools that are just teaching the test because then you're not learning about the world, you're not learning about different cultures, you're not learning about science, you're not learning about math. All you're learning about is how to fill out a little bubble on an exam and little tricks that you need to do in order to take a test and that's not going to make education interesting.

President Barack Obama, March 28, 2010

Introduction

Assessments can either drive or constrain innovation in education. The development of new assessments that measure student performance against the Common Core State Standards offers a powerful point of leverage in the effort to ensure that the new standards deliver on their promise of enhanced educational opportunities. New assessments can help to focus teachers' and students' attention on essential knowledge and skills, and provide more timely and useful information for states, schools, teachers, parents and students themselves. It is vital that these new assessments measure student achievement in more complete, authentic, and meaningful ways.

Currently, the Common Core State Standards exist only for English-language arts and mathematics, but work is underway to develop national K-12 standards for science. The National Research Council (NRC), Achieve, Inc., National Science Teachers Association (NSTA), and American Association for the Advancement of Science are all engaged in efforts to develop new national science standards. While numerous papers have summarized research findings and made recommendations to both assessment consortia on critical assessment issues, none has focused specifically on the possibilities for next generation assessments for assessing science content and inquiry skills.

The limitations of current state assessments are well documented, and are the fundamental reasons for the creation of the two national assessment consortia. Paper-and-pencil item-based tests are intrinsically incapable of providing authentic measurement of the complex intellectual and psychosocial performances that are essential for 21st century work and citizenship. The inadequacies of traditional assessments have been particularly troubling in science. Typical multiple-choice tests are insufficient for determining how well students are developing sophisticated inquiry[1] skills in science—a key capability for science, technology, engineering, and mathematics (STEM) careers. Many reports and studies have documented that higher-order thinking skills related to sophisticated cognition (e.g. inquiry

1 For this paper, we use two definitions of inquiry. The first, is based on White, Collins, and Frederiksen's (in press) definition, which centers on theorizing, questioning & hypothesizing, investigating, analyzing, and synthesizing. This detailed definition is placed in the context of the National Science Education Standards general definition of scientific inquiry as: *"…the diverse ways in which scientists study the natural world and propose explanations based on the evidence derived from their work…also…the activities through which students develop knowledge and understanding of scientific ideas, as well as an understanding of how scientists study the natural world."* National Research Council (1996). National Science Education Standards. Washington, DC: National Academies Press.

processes, formulating scientific explanations, communicating scientific understanding, strategies for resolving novel situations) are difficult to measure with multiple-choice or even constructed-response paper-and-pencil tests (NRC, 2006). Research has shown that these tests also are incapable of showing whether or not science instruction is effective in helping students learn inquiry (Quellmalz et al., 2007). The development of more valid assessments of science inquiry and related skills holds the potential to ensure that students are not only better prepared for the STEM fields that are essential for the nation's economic viability, but also better prepared for a broad range of work and citizenship responsibilities.

This paper describes the potential of next generation state assessments for science that utilize modern technology for delivery, scoring, and reporting. Its argument also makes the case for new types of assessments aligned to the NRC's vision for state science frameworks that encourages "students to actively engage in science practices in order to deepen their understanding of core ideas" over time. The paper illustrates these possibilities through a case study of a research project at the Harvard Graduate School of Education focused on developing virtual performance assessments (VPAs) in science. The approach described here aligns with the National Assessment of Education Progress science framework, which emphasizes 1) a paring down of content, 2) assessing students' conceptual knowledge and 3) performance assessments.

While discussions about using computer adaptive technology and about revising assessments to better assess English learners are ongoing, the development of new assessments for science is in an earlier stage. This paper seeks to help in shaping formative conceptions of science assessments. The hope is to unbind notions of science assessments from what *is* and to embolden new ideas about what these assessments *could be* if we truly mean to prepare all students for college and the careers of the 21st century.

Background and Context

Numerous reports on economic development in the context of globalization state that inquiry and complex reasoning are critical skills for competing in our knowledge-based, global economy (NRC, 2006; President's Council of Advisors on Science and Technology (PCAST), 2010). The worker of the 21st century must have science and mathematics skills, creativity, fluency in information and communication technologies, and the ability to solve complex problems.

Yet, data indicate that the U.S. is failing to promote student mastery of crucial science knowledge, skills and abilities, including inquiry practices and complex reasoning. For example, the 2009 National Assessment of Education Progress (NAEP) results reveal that a majority of the U.S. students who took the test received scores below proficient. Similarly, on the 2007 Trends in International Mathematics and Science Study (TIMSS), only 10 percent of U.S. eighth graders reached the advanced benchmark on the science portion—demonstrating a weak grasp of complex and abstract scientific concepts. On the 2009 Program for International Student Assessment (PISA), only 29 percent of students in the U.S. demonstrated ability to complete higher-order tasks such as those involving scientific explanations. That current instruction in science does not focus on these skills explains some of this lackluster performance; teachers center students' learning on what is measured in summative tests (Darling-Hammond, 2010). Re-thinking science assessments to better measure inquiry and complex reasoning will help refocus instruction on these crucial skills.

The Nature of Inquiry

Scientific inquiry is hypothesized to be the method by which scientists study the world. In order to promote scientific reasoning, the NRC argues that students must participate in authentic practices of science. As discussed in White, Collins, and Frederikson (in press), while detailed definitions of inquiry can be complex, at its core this suite of processes centers on theorizing and investigating. For example, Kuhn, Black, Keselman, and

Kaplan (2000) define inquiry learning as investigations where students individually or collectively investigate a set of phenomena (virtual or real) and draw conclusions about it.

Below are two of the related skills we measure in the particular virtual performance assessment (VPA) described in this paper:

- Student develops a scientific explanation of what is happening in the virtual world that includes: 1) a claim about the phenomena, 2) the evidence (either empirical or observations), and 3) reasoning that links claims with evidence.
- Student gathers data that help explain or provide evidence to justify the claim being made.

Later, we discuss how our VPA elicits and measures skills such as these in detailed ways more similar to the work of scientists than do multiple-choice, short answer, or essay questions.

Current Challenges in Measuring Inquiry

The NRC recently released a public draft of their conceptual framework for new science education core standards. A primary emphasis in this new framework is that learning about science and engineering involves the integration of knowledge of scientific explanations (i.e. content knowledge) with the practices and skills needed to engage in scientific inquiry and engineering design. Similarly, other national and international science frameworks, such as the Science Framework for the 2011 NAEP, the College Board Standards for College Success, and the Programme for International Student Assessment (PISA) of the Organization for Economic Co-operation and Development all place emphasis on integrated science practices and performances.

The mastery of discrete science facts as measured on most standardized tests is an inadequate representation of whether students know or understand integrated science practices. In response, for over two decades, researchers have sought to develop alternative assessments for measuring science that involve tasks in an authentic or real-life context or that mirror the workplace or other real-life contexts. For example, in the 1990s Maryland used hands-on performance assessments in science as part of their state testing program. Numerous studies were conducted on performance assessments to assess the psychometric properties of these alternate assessments. These studies also focused on the feasibility (i.e. cost effectiveness and practicality) of using these types of measures on a large scale. Research findings indicated that these alternate assessments were more aligned to the knowledge, skills and abilities (KSAs) being measured and more valuable for providing feedback to teachers about ongoing student attainment than multiple-choice tests.

However, there were several limitations to the use of hands-on performance assessments as summative assessments for accountability. Research on hands-on performance assessments found:

- Students performed differently on similar tasks that were supposedly measuring the same construct; ideally, a student would perform consistently on various tasks assessing a construct (Shavelson et al., 1993);
- Students performed differently on identical tasks on different occasions (Cronbach et al., 1997);
- Hands-on performance assessments are cost-prohibitive when compared to multiple-choice tests (Stecher & Klein, 1997); and
- Hands-on performance assessments still have limited validity, despite their authenticity, compared to multiple-choice tests (Linn et al., 1991).

Affirming this last point, a recent study found, "Even the hands-on performance tasks in these large-scale science tests are highly structured and relatively short (15-40 minutes), truncating the investigation strategies that can be measured" (Quellmalz et al., 2007, page 1).

New Opportunities in Technology-based Assessment

Fortunately, since the performance based assessment studies of the 1990s, three advances have taken place that potentially enable online performance assessments capable of validly measuring the full complexity of scientific inquiry: 1) advances in cognitive science, 2) advances in statistics and measurement, and 3) advances in information and communication technologies. To illustrate the power and potential of these new types of performance assessments, this paper describes our

current research on one such model: immersive virtual environments (IVEs). IVEs are three-dimensional (3-D) environments, either single or multi-user, where participants' digital personae (avatars) engage in virtual activities and experiences. These immersive, interactive media have become commonplace in many people's lives, through gaming (e.g. World of Warcraft and America's Army), social interactions and learning (such as Second Life, Club Penguin), and recreation (e.g. The Sims Online, Webkinz). Part of the attraction of IVEs is that they can simulate complex real-world settings in which participants can enact the types of complicated processes that underlie various real-world workplace roles. Both military and medical education have benefited from this capability (Fletcher, 2009; Kneebone, 2005).

Research has established that, when well-designed, IVEs can aid students in learning authentic, sophisticated inquiry practices (Mayo, 2009). Our studies of virtual environments as curriculum provide an example. For almost a decade, our research team has studied the feasibility and practicality of using IVEs to increase student achievement in scientific inquiry (Dede, 2009). In this research, we studied how virtual environments enable students to do authentic inquiry and engage in the processes of science. Our first series of studies, funded by National Science Foundation from 1999-2009, were on *River City*. The *River City* curriculum was a multi-user immersive virtual environment (MUVE) designed to teach middle school science (Clarke et al., 2006). The curriculum was centered on skills of hypothesis formation and experimental design, as well as on content related to national standards and assessments in biology and ecology.

We were able to implement MUVE-based curricula in a wide range of schools in a manner that teachers and technology coordinators found practical and scalable. We worked with over 200 teachers and more than 20,000 students. We conducted a series of quasi-experimental design studies to determine if virtual environments can simulate real-world experimentation and provide students with engaging, meaningful learning experiences that increase achievement in scientific inquiry. Using conventional paper-and-pencil, item-based measures, our results from a series of research studies showed that these virtual environments enable students to engage in authentic inquiry tasks (problem finding and experimental design) and also increase students' engagement and self-efficacy (Clarke & Dede, 2007; Clarke et al., 2006; Ketelhut, 2007; Nelson, 2007). This seminal research on IVEs was discussed both in the U.S. Department of Education's National Educational Technology Plan (2010) and in the National Research Council's report on games and simulations in science education (2011).

Even though paper-and-pencil tests captured some of students' learning, we found that students' performance on the multiple-choice pre-post-tests typically used as measures in this type of research did not necessarily reflect learning that we saw via interviews, observations, summative essays, and analyses of log file data that capture students' activity as they interact with the environment (Clarke, 2006; Ketelhut & Dede, 2006; Ketelhut et al., 2007). We built rich case studies of student learning in which we triangulated and compared different data sources, both qualitative and quantitative, in order to illustrate and understand students' inquiry learning (Clarke, 2006; Clarke & Dede, 2005, 2007; Ketelhut et al., 2007). A finding from our experience was that paper-and-pencil item-based assessments, even after extensive refinement, do not fully capture students' learning of inquiry skills.

Further, our immersive curricular environments and similar interactive, immersive media enable the collection of very rich datastreams about individual learners that provide better ways to assess inquiry processes (Clarke, 2009; Ketelhut et al., 2007). We believe that these streams of "active" behavioral data on student performances can be utilized in the development of virtual assessments. While research on game-like simulations for fostering student learning is starting to proliferate, studying the potential of this medium for summative assessments of student learning in a standardized fashion is still in its infancy.

The U.S. Department of Education's 2010 National Educational Technology Plan (NETP) identifies increased global economic competition as a fundamental challenge facing the U.S. over the next decade. The NETP therefore calls for immediate action to ensure that today's U.S. students are learning 21st century skills that foster innovation and economic prosperity.

One recommendation in this Plan focuses specifically on developing sophisticated forms of technology-based assessment: *"2.3 Conduct research and development that explores how embedded assessment technologies, such as simulations, collaboration environments, virtual worlds, games, and cognitive tutors, can be used to engage and motivate learners while assessing complex skills"* (U.S. Department of Education (USDOE), 2010, page 19). In research on assessment, IVEs enable investigators to measure authentic, situated performances that reflect the inquiry processes used by scientists (Donovan & Bransford, 2005).

Our current studies center on whether IVEs can provide reliable, practical, affordable summative assessments in accountability settings that are more valid than paper-and-pencil, item-based tests in measuring science inquiry and similar sophisticated STEM skills. Immersive virtual performance assessments use the IVE interface to offer a new model for how we measure higher-order skills such as problem-solving, causal reasoning, and inquiry learning. Our research is synthesizing over two decades of studies on performance assessments, measurement theory, cognition, technology, and videogame design. In the following section, we provide a case study of our work developing and studying summative virtual performance assessments that measure science inquiry processes as part of a national or state high stakes testing program. This research exemplifies the new types of assessments Partnership for the Assessment of Readiness for College and Careers (PARCC) and SMARTER Balanced Assessment Consortium (SBAC) should explore and promote.

The Virtual Performance Assessment (VPA) Project

With funding from the Institute of Education Sciences (IES), the VPA project at the Harvard Graduate School of Education is developing and studying the feasibility of immersive virtual performance assessments to assess scientific inquiry of middle school students as a standardized component of an accountability program (see http://vpa.gse.harvard.edu). The goal is to provide states with reliable and valid technology-based performance assessments linked to state and National Science Education Standards (NSES) academic standards for science content and inquiry processes, extending capabilities to conduct rigorous studies that provide empirical data on student academic achievement in middle school science.

In order to ensure that we were measuring what we intended to measure (inquiry), we used the Evidence Centered Design (ECD) framework (Mislevey & Haertel, 2006; Mislevy & Rahman, 2009) to design our assessments. ECD formalizes the procedures generally done by expert assessment developers. Using the ECD approach allowed us to articulate every aspect of the assessment from the knowledge, skills, and abilities (KSAs) that they are measuring to the types of evidence that will allow one to make claims about what students know. In addition, we are using the Principled Assessment Designs for Inquiry (PADI) system, which is software for creating assessments for science inquiry based on the ECD framework. Design templates allow assessment developers to create multiple forms of the same assessment. Using these frameworks, we have reframed science inquiry constructs (theorizing, questioning and hypothesizing, investigating, analyzing and synthesizing) into specific KSAs aligned with current national standards. Through the process of articulating the exact details of what is being measured and how it is being measured, it is easy to link the KSAs to evidence of student learning (see Appendix A for a Table describing an extended framework for VPA design and development). Linking KSAs like this provides a measure of validity that research has found often lacking in performance assessments (e.g. Linn, Baker et al. 1991).

Description of the Assessments

Traditional assessments often focus on individual test items and rely on student affirmation as a response that indicates knowledge. In our VPAs, we base the evaluation of student performance on measurements captured as in-world interactions. These interactions allow us to assess what students know and do not know about science inquiry and problem solving. As a part of the inqui-

VPA in Action: An Illustration

To demonstrate how the performance assessment works at the classroom level, the following is a brief description of a VPA.

It is May, and students in Ms. Jones' eighth grade science class are participating in a virtual performance assessment pilot program. As part of their state accountability program, all students in eighth grade must demonstrate proficiency in integrated science practices. These assessments are summative assessments that are meant to sample the domain of inquiry. Students take the virtual assessment in a block period. The assessment lasts about fifty minutes.

Ms. Jones logs into the VPA teacher's portal and creates accounts for her students, selecting the initial assessment she wants them to take. When class starts, the students sit at individual computers and login to begin their simulated experience.

Arielle sits at her computer and logs into the student portal. She opens the assessment and is immediately allowed to choose what her avatar looks like. She selects an avatar and enters the world.

Figure 3. A VPA avatar selection screen.

The camera slowly provides an aerial view of the world to orient Arielle to the problem space. Arielle sees that there is a village and what appear to be farms with ponds. The camera then focuses in on a multi-colored frog with six legs. Arielle wonders, "What could be causing this frog to have six legs?" The assessment begins. A scientist and farmers who have just discovered this mutated frog greet Arielle. The farmers all offer competing hypotheses for why the frog is mutated. The scientist turns to Arielle's avatar and tells her that she must conduct an investigation and come up with her own theory, backed up with evidence. He asks her if she thinks any of the hypotheses provided are plausible.

Figure 4. Characters presenting competing hypotheses in a VPA.

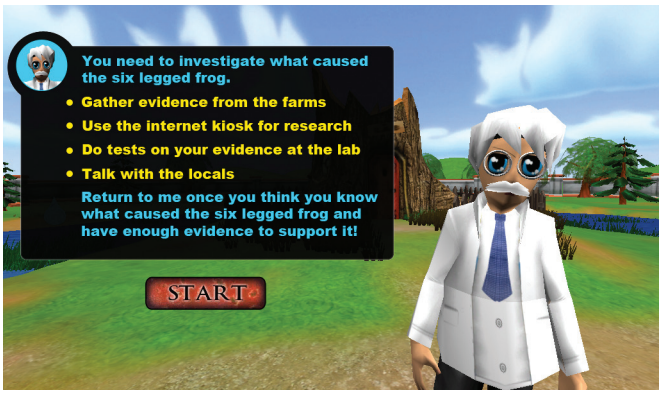


Figure 5. Setting up the problem.

Arielle must come up with a claim and support it with evidence and reasoning. In order to make her claim she must first gather data. She sets out to explore the farms.

Just as scientists collect data, Arielle has options of different kinds of data she can collect. There is a lab she can visit to run tests. For example, she can collect water, tadpoles, and frogs from each of the four farms. She can then bring them to the lab to conduct water tests, blood tests, and genetic tests. In addition, prior research studies are available, and residents provide data about their points of view.

Figure 6. A VPA backpack containing a limited number of items.

Arielle must make a choice about what data she thinks is the most important or that she wants to investigate first. One aspect of doing inquiry is knowing what data will justify a claim. We want to examine students' data gathering strategies and see how they correlate to the claims they make. Thus, we limit the amount of data that students can carry in their backpack. If students were allowed to pick up every piece of data in the world, then it would be difficult to make inferences about their knowledge of what data is important evidence in the investigation. If students were asked to evaluate a piece of data every time they collected it, then the task would become boring.

Thus, the design requires students to make choices through actions. Students can carry only eight pieces of data at a time. They can go to the lab at any time to run tests on the data (e.g. water tests, blood tests, genetic tests). Any piece of data discarded from the backpack will go back into the world and can be picked back up at any time (given there is space in the backpack). This is not a design that is related to game play, nor is it meant to model how many samples to collect. It is a structural feature that applies constraints in order to force students to be more thoughtful about the data they collect in this assessment.

Also, the availability of prior research studies helps to "level the playing field" between students who start with stronger content knowledge and those who begin with weaker. This is important in ensuring that the assessment is measuring inquiry process skills rather than content knowledge.

Arielle collects eight pieces of data from two farms. She realizes that she cannot carry any more and decides to go to the lab to run some tests. She arrives at the lab and examines the water samples. Her tests show that the lab water and water from one of the farms contains pesticides. However, one of the farms has clean water. She runs genetic tests on the two frogs she collects and sees that they are the same. She notes that both of the frogs have high counts of white blood cells. She decides that she needs to learn more about what these tests tell her. She goes to the research kiosk and looks up information on blood tests and pesticides. Arielle started with examining data and then moved to research that would allow her to reason from the data.

At the computer on Arielle's left, Maria is tackling the assessment differently. Initially, she decides to examine prior research studies. She examines the research that is available on frogs and tadpoles. She reads about viruses and genetic mutations in frogs and decides to gather data in order to determine which is the cause. She goes to each of the four farms and collects a tadpole and a frog to run tests on. Back at the lab, she finds that all of the frogs have similar genetic make-up. However, two of the tadpoles have small tails. She notes that a frog from the same farm also has a virus in its blood. She looks up the virus in the research documents and believes she has found evidence. She speaks to the scientist and builds a claim for why the frog is mutated, including evidence-based reasoning from the research she conducted.

After class, Ms. Jones reviews the reporting tool to see the diagnoses the assessment provides about what each student knows and does not know about gathering data about a claim, making a claim, and then supporting it with evidence and reasoning. The tool presents data at both the individual and class level; and Ms. Jones finds that, while the majority of students are strong in providing evidence, they are weak in reasoning from evidence. Also, some students collect data related to hypotheses they have already discarded, or collect all possible data without seeming to have a hypothesis. This shows weaknesses in their inquiry skills.

This particular VPA is designed as a summative assessment to sample from the domain of science inquiry. Overall, VPAs can serve as one component of a comprehensive science assessment (e.g. an inquiry portion of a summative assessment for science).

ry progression embedded within the VPAs, students are required to make a series of choices as they participate in an ongoing narrative. The focus of the VPAs is not the attainment of a single right answer, but rather on the result of a series of choices that students make. The series of interactions produces rich observations that enable us to make a fine distinction of students' understanding of the various facets of inquiry discussed earlier.

Important Aspects of the VPA Model

The skills we are measuring in this particular VPA focus on gathering data about a claim, making a claim, and supporting it with evidence and reasoning—skills that we argue are difficult to capture in multiple-choice and open-response tests. By setting up the assessment in a virtual environment, we can follow students' trajectories of data gathering. We then can correlate these to the claims they build and the evidence and reasoning they use to support those assertions. Each challenge in our assessments relies on students collecting data and providing evidence to support a claim, and students' scores are based on the evidence and reasoning they provide for a given claim.

In our prior research in immersive virtual environments mentioned above, we did not find students' experience playing video or computer games to be a predictor of their performance in the curriculum. Thus, we hypothesize that videogame and computer game experience will not be a predictor of student performance on our VPAs; we are conducting research in order to test this hypothesis. If we are correct in our assumption, then the only strategies students can learn in order to do better on our VPAs involve applying inquiry practices—the domain processes on which instruction should focus. In contrast, students can learn test-taking skills that aid them in correctly selecting multiple-choice answers even though these skills do not provide additional knowledge of the domain; this distorts what is measured.

We support the concept of multiple modes of assessments that triangulate to form claims about what a student does or does not know. Our early work in virtual performance assessments has centered on measuring students' ability to reason from evidence, as a demonstration of concept for a much broader set of VPAs measuring a wide variety of knowledge, skills, and abilities important in authentic practices. These can then complement more conventional forms of testing.

Facing a mandate to implement digital assessments by 2014, states are already moving towards computer-based testing because of the types of sophisticated intellectual and psychosocial performances they can measure. Whether these future assessments are merely digitized versions of paper-and-pencil tests or offer the increased features of VPAs, similar infrastructure, resource, and policy issues must be addressed. Fortunately, based on our experiences with scaling up MUVE-based curricula, we believe that the investments made to enable simple digital assessments will be sufficient to enable the use of VPAs. In other words, at scale, the expense involved in constructing and using VPAs is roughly comparable to other forms of high quality testing.

In summary, virtual performance assessments have numerous advantages over hands-on performance assessments:

1. Standardizing the administration of hands-on performance assessment is difficult, so extensive training is required. In contrast, VPAs ensure standardization by delivering instruction to students in an identical manner via the technology.
2. VPAs alleviate the need for developing, shipping, and providing schools with materials and kits for hands-on tasks. All that is needed are a computer and an internet connection.
3. Scoring will all be done by the technology, so no raters are necessary, reducing cost, training, and the possibility of human error.
4. VPAs potentially have fewer problems with task and occasion variability.

VPAs also have advantages over conventional paper-and-pencil tests and digitized versions of these:

1. Multiple-choice, short answer, and essay questions do not present a realistic context within which to elicit the processes of complex performances such as scientific inquiry.
2. VPAs mirror for students the types of inquiry processes to which teachers should orient instruction more accurately than do conventional measures of inquiry.

3. The use of test-taking strategies can distort the outcomes of conventional item-based measures, but prior studies suggest that this may not be the case with VPAs.
4. VPAs can seamlessly incorporate features to minimize the importance of prior content knowledge and can track the extent to which a student utilizes these.
5. VPAs provide a more detailed record of student actions than do conventional item-based tests.
6. At scale, VPAs are as cost-effective and practical as other forms of digital assessment.

Our research is working to establish whether VPAs' psychometric properties are sufficient to justify their use in high stakes testing, thereby realizing these advantages.

Implications for Policy

In this paper, we have discussed virtual performance assessments as a new model for re-conceptualizing the assessment of science inquiry. VPAs are based on over two decades of research on performance assessments and assessment design, cognition, technology, and video-game design. This type of assessment has considerable promise not only for measuring higher-order skills in science, but also for evaluating students' progress on other sophisticated intellectual and psychosocial performances.

We offer the following recommendations for the practical, scalable implementation of VPAs as part of comprehensive state assessment systems:

Include virtual performance assessments as part of comprehensive state assessment systems. Given the advantages of VPAs discussed above, the redesign of state assessment systems should move beyond using technology to digitize and automate conventional assessments.

Ensure technological capacity. When investing in infrastructure for digital assessments, schools and districts should purchase machines with modern videocards capable of displaying detailed graphics and animations. The ability to render visually rich environments, such as virtual worlds, is important not only for VPAs, but also for instruction in general. As schools increasingly use digital content to aid learning, the devices and networks they purchase should be capable of delivering the full features of this material.

Provide teachers and students with opportunities to use virtual performance assessments. In order for new assessments to be a valid measure of students' knowledge, the technology cannot be a barrier to them demonstrating this knowledge. Students must have experience with and comfort using computers in order for their knowledge and skills to be assessed on computers. Similarly, professional development for teachers should include opportunities to use virtual performance assessments. Teachers understand the format of paper-and-pencil, item-based tests and how to gear instruction to that type of measure. Being able to explore and use VPAs is important for teachers, as they then will know what types of learning experiences students need in order to perform well on these assessments. These experiences are essential in shifting what is taught towards the knowledge, skills and abilities that most matter in 21st century STEM learning.

Provide opportunities for key stakeholders to experience virtual performance assessments. Parents, school boards, and community members should have the opportunity to experience performing tasks in a virtual environment that provides feedback on their accomplishments. This will alleviate fears that students are being taught and assessed in ways that maximize "gameplay" rather than key educational objectives. Further, since VPAs simulate authentic practices in settings similar to the real world, this type of measure has a "face validity" that highlights why it is a valuable complement to traditional tests.

Provide professional development to teachers to foster the instruction that will lead to high performance on VPAs. Educational stakeholders should be informed of the curricular and professional development investments necessary to increase student outcomes on measures of sophisticated intellectual and psychosocial performances. Teachers will need to understand both the knowledge, skills and abilities that underlie that domain and instructional methods for helping stu-

dents attain that knowledge. Curriculum selection must include materials that foster deep experiential learning of core authentic practices in the domain. The focus of instruction must move from broad, shallow coverage to focus on fewer core topics. These are key educational improvements central to preparing students for further schooling, work, and citizenship in the 21st century that, if taught well, will be evident by student outcomes on VPAs.

Conclusion

This is not a time to be conservative about implementing new forms of assessment, such as VPAs. It is abundantly clear that digitized versions of paper-and-pencil, item-based tests are insufficient to assess vital STEM skills like science inquiry. In fact, substantial evidence shows that these tests are undercutting students' learning the key knowledge, skills and abilities required for careers that could help the United States to compete in the global, knowledge-based economy (NRC, 2006; PCAST, 2010; USDOE, 2010; NRC, 2011). Policymakers and state education leaders have a once-in-a-generation opportunity to fundamentally improve the way students' science-related knowledge, skills and abilities are assessed. We hope they have the courage and foresight to rapidly implement innovations that can foster students' mastery of science and ultimately, their success in college, career and life.

Acknowledgements

Our Virtual Performance Assessments research is funded by the Institute for Education Sciences, U. S. Department of Education, grant R305A080141. The investigators' findings, interpretations, and conclusions do not constitute an official position of the U. S. Department of Education. The authors gratefully acknowledge the contributions of Drs. Jillianne Code, Nick Zap, Geordie Dukas, and Michael Mayrath, as well as the assistance of numerous students at the Harvard Graduate School of Education.

References

Clarke, J., & Dede, C. (2005). *Making learning meaningful: An exploratory study of using multi-user environments (MUVEs) in middle school science.* Paper presented at the American Educational Research Association Conference, Montreal, Canada.

Clarke, J. (2006). *Making Learning Meaningful: An Exploratory Study of Multi-User Virtual Environments in Middle School Science.* Qualifying Paper submitted to the Harvard Graduate School of Education, Cambridge, MA.

Clarke, J., Dede, C., Ketelhut, D. J., & Nelson, B. (2006). A design-based research strategy to promote scalability for educational innovations. *Educational Technology,* 46, 3 (May-June), 27-36.

Clarke, J. & Dede, C. (2007). *MUVEs as a Powerful Way to Study Situated Learning.* The Proceedings of Conference for Computer Supported Collaborative Learning (CSCL). Mahwah, NJ: Lawrence Erlbaum Associates.

Clarke, J. (2009). *Exploring the compexity of inquiry learning in an open-ended problem space.* Unpublished Doctoral Dissertation, Harvard University, Cambridge, MA.

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

Darling-Hammond, L. (2010). *Performance counts: Assessment systems that support high quality learning.* Washington, DC: Council of Chief State School Officers.

Dede, C. (2009). Immersive interfaces for engagement and learning. *Science,* 323(5910), 66-69.

Donovan, S., & Bransford, J. (2005). *How students learn—mathematics in the classroom.* Washington, DC: National Research Council.

Fletcher, J. D. (2009). Education and training technology in the military. *Science* 323(5910), 72-75.

Ketelhut, D. J., & Dede, C. (2006). *Assessing inquiry learning.* Paper presented at the National Association of Research in Science Teaching, San Francisco, CA.

Ketelhut, D. J., Dede, C., Clarke, J., Nelson, B. & Bowman, C. (2007). Studying Situated Learning in a Multi-User Virtual Environment. In E. Baker & J. Dickieson & W. Wulfeck & H. O'Neil (Eds.) *Assessment of Problem Solving Using Simulations.* Mahwah, NJ: Lawrence Erlbaum Associates.

Ketelhut, D. J. (2007). The impact of student self-efficacy on scientific inquiry skills: An exploratory investigation in River City, a multi-user virtual environment. *The Journal of Science Education and Technology,* 16(1), 99–111.

Kneebone, R. (2005). Evaluating clinical simulations for learning procedural skills: a theory-based approach. *Academic Medicine* 80(6):549-53.

Kuhn, D., Black, J., Keselman, A., & Kaplan, D. (2000). The development of cognitive skills to support inquiry learning. *Cognition and Instruction,* 18(4), 495-523.

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

Mayo, M.J. (2009). Videogames: A route to large-scale STEM education? *Science* 323(5910), 79-82.

Mislevy, R., & Haertel, G. (2006). *Implications of Evidence-Centered Design for Educational Testing (Draft PADI Technical Report 17).* Menlo Park, CA: SRI International.

Mislevy, R., & Rahman, T. (2009). *Design pattern for assessing cause and effect reasoning in reading comprehension.* Menlo Park, CA: SRI Interantional.

National Research Council (2006). *Systems for state science assessment.* Washington, DC: The National Academies Press.

National Research Council (2011). *Learning science through games and simulations.* Washington, DC: National Academies Press.

Nelson, B. (2007). Exploring the use of individualized, reflective guidance in an educational multi-user virtual environment. *The Journal of Science Education and Technology,* 16(1) 83–97.

President's Council of Advisors on Science and Technology (PCAST) (2010). *Prepare and inspire: K-12 education in science, technology, engineering, and math (STEM) for America's future.* Washington, DC: Office of Science and Technology Policy, Executive Office of the President.

Quellmalz, E., P. Kreikemeier, et al. (2007). *A study of the alignment of the NAEP, TIMSS, and New Standards Science Assessments with the inquiry abilities in the National Science Education Standards.* Annual Meeting of the American Educational Research Association, Chicago, Il.

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

Stecher, B. M. & Klein, S. P. (1997). The cost of science performance assessments in large-scale testing programs. *Educational Evaluation and Policy Analysis,* 19(1), 1-14.

U.S. Department of Education (2010). *National Education Technology Plan 2010.* Washington, DC: U.S. Department of Education.

White, B. Y., Collins, A. & Frederiksen, J. R. (in press) The nature of scientific meta-knowledge. In M. S. Khine & I. Saleh (Eds.) *Dynamic modeling: Cognitive tool for scientific enquiry.* London: Springer.

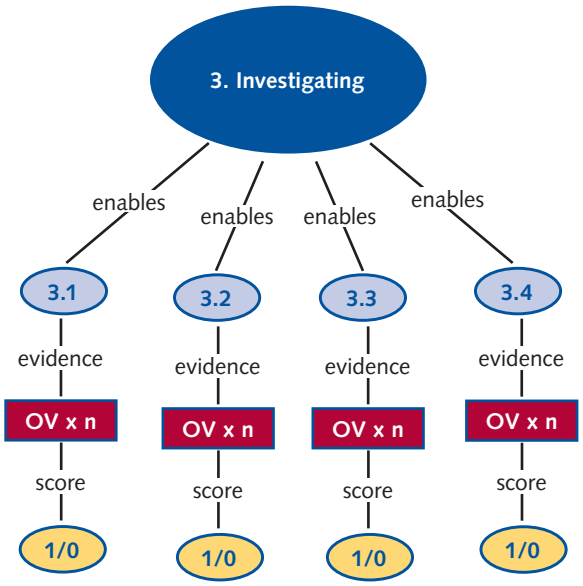
Appendix A

Table 1. Extended ECD Framework for VPA design and development.

Modified ECD framework	Description
I. Domain Analysis	Develop purpose for assessment. Compile research on the development of inquiry skills. Develop definition of competence. Develop knowledge, skills, and abilities (KSAs) we are measuring. Consult experts in the fields about our chosen definitions and definitions of inquiry and assessment objectives.
II. Domain Modeling	Use information from the domain analysis to establish relationships among proficiencies, tasks, and evidence. Explore different approaches and develop high-level sketches that are consistent with what students have learned about the domain so far. Develop narrative descriptions of proficiencies of inquiry, ways of getting observations that evidence proficiency, and ways of arranging situations in which students provide evidence of targeted proficiencies. Create graphic representations and schema to convey these complex relationships, and develop prototypes.
III. Conceptual Assessment Framework: • Cognitive Model • Student Model • Observation/Tasks Evidence • Interpretation	1. Cognitive Model: Identify set of theory or beliefs about how students represent knowledge and develop competence in a subject. 2. Student model: What complex of knowledge, skills, or other abilities should be assessed? 3. Observations/Tasks: Identify kinds of tasks or situations (interactions) that will prompt students to say, do, or create something that demonstrates important knowledge, skills, and competencies. 4. Evidence: Identify behaviors and performances that reveal knowledge and skill identified in the student model. Identify and summarize evidence. 5. Interpretation: Develop a method for interpreting observations and evidence.
IV. Compilation: • Task creation • Statistical Assembly • Assessment • Implementation	Develop purpose for assessment. Compile research on the development of inquiry skills. Develop definition of competence. Develop knowledge, skills, and abilities (KSAs) we are measuring. Consult experts in the fields about our chosen definitions and definitions of inquiry and assessment objectives.
IV. Four-Process Delivery Architecture: • Presentation • Response Scoring • Summary Scoring • Activity Selection	Develop architecture and processes for implementing assessments. Develop back-end architecture that will capture and score student data. Develop prototype. Pilot.
VI. Refinement	Refine assessment based on pilot data. Iterative cycle.

Figure 7. Representation for the larger construct, "Investigating," and how it is broken down into KSAs.

3.1: Student gathers data that help explain or provide evidence to justify the claim being made. 3.2: Student determines which data from a specific investigation can be used as evidence to address an explanation. 3.3: Student distinguishes credible data from non-credible data in terms of quality. 3.4: Student is able to gather data after the experiment that will provide the evidence needed to prove or disprove whether the causal relationship was true. OVxn: Observational variables associated with each KSA. Score for each step can be either zero or one.



Closing Remarks

The three papers in this report discuss powerful assessments that bear little resemblance to the typical statewide assessments of today. The assessment strategies that the authors describe begin to fulfill the promise of an assessment system that supports richer instruction and deeper learning for students, while providing fair and accurate information about the performance of students, teachers, and schools. In the near term, an aggressive timeline and limited funding for development have the potential to overwhelm the good intentions and lofty ambitions of the two assessment consortia. The purpose of this report is not to diminish the gravity of these challenges, but to say that they must—and we believe—can be overcome. This requires a long-term view, with a focus not just on our next steps, but on the road ahead.

To put it plainly, the time will never be better to make drastic changes to the way the vast majority of U.S. students are assessed. The assessment consortia do not have funding to administer the new assessments long-term. So, due to funding constraints and the fact that so many states will need to sign off on future changes, it will be quite arduous to make changes once the consortia have completed their work. The once-in-a-generation opportunity to increase the number of students who leave school ready for success in college and careers cannot be lost.

The three papers include many suggestions for how to move forward, but in our view these are the most urgent:

- **Devote sufficient time for thoughtful planning and pilot testing of computer adaptive testing.** Computer adaptive assessment provides the essential foundation for a system that can produce fair and accurate measurement of English learners' knowledge and of all students' knowledge and skills in science and other subjects. In our view, developing computer adaptive assessments is a necessary intermediate step toward a system that makes assessment more authentic by tightly linking assessment tasks and instructional activities—and ultimately embedding assessment in instruction. It will be vital to keep these goals in mind, even as we recognize technological and resource constraints.
- **Integrate the development of new assessments with assessments of English-language proficiency.** The next generation of ELP assessments will need to be based on ELP standards that sufficiently specify the target academic language competencies that English learners need to progress in and gain mastery of the Common Core State Standards. As Robert Linquanti makes clear, "Acknowledging and overcoming the challenges involved in fairly and accurately assessing ELs is integral and not peripheral to the task of developing an assessment system that serves all students well. Treating the assessment of ELs as a separate problem—or, worse yet, as one that can be left for later—calls into question the basic legitimacy of assessment systems that drive high stakes decisions about students, teachers, and schools."
- **Include virtual performance assessments as part of comprehensive state assessment systems.** Virtual performance assessments have considerable promise for measuring students' inquiry and problem-solving skills in science and in other subject areas, with authentic assessment closely tied to or even embedded in instruction. The simulation of authentic practices in settings similar to the real world opens the way to assessment of students' deeper learning and their mastery of 21st century skills.

As the authors of these papers make clear, the technologies required to accomplish these ambitious goals are already in use. It is critical that the work of the assessment consortia focus on building a system that provides fair and accurate information about all students' performance and supports the work of teachers and schools to prepare them for success as adults.

Glossary of Key Assessment Terms

Validity is "the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests" (page 9). In other words, validity refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences.

Reliability is "the consistency of measurements when the testing procedure is repeated on a population of individuals or groups" (page 25). In other words, reliability refers to the consistency of measurement, or the degree to which a test measures the same way each time it is used under the same condition with the same subjects.

A **construct** is a nonmaterial human dimension or attribute that a test seeks to measure.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.