J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 10, Number 5 · November 2010

Technology-Enabled and Universally Designed Assessment: Considering Access in Measuring the Achievement of Students with Disabilities— A Foundation for Research

Patricia Almond, Phoebe Winter,
Renée Cameto, Michael Russell, Edynn Sato,
Jody Clarke-Midura, Chloe Torres,
Geneva Haertel, Robert Dolan, Peter Beddow,
& Sheryl Lazarus

www.jtla.org

A publication of the Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College

JTLA

Volume 10, Number 5

Technology-Enabled and Universally Designed Assessment: Considering Access in Measuring the Achievement of Students with Disabilities—A Foundation for Research

Patricia Almond, Phoebe Winter, Renée Cameto, Michael Russell, Edynn Sato, Jody Clarke-Midura, Chloe Torres, Geneva Haertel, Robert Dolan, Peter Beddow, & Sheryl Lazarus

Editor: Michael Russell
russelmh@bc.edu
Technology and Assessment Study Collaborative
Lynch School of Education, Boston College
Chestnut Hill, MA 02467

Copy Editor: Jennifer Higgins
Design: Thomas Hoffmann
Layout: Aimee Levy

JTLA is a free online journal, published by the Technology and Assessment Study Collaborative, Caroline A. & Peter S. Lynch School of Education, Boston College.

Preferred citation:

Almond, P., Winter, P., Cameto, R., Russell, M., Sato, E., Clarke-Midura, J., Torres, C., Haertel, G., Dolan, R., Beddow, P., & Lazarus, S. (2010). Technology-Enabled and Universally Designed Assessment: Considering Access in Measuring the Achievement of Students with Disabilities—A Foundation for Research. *Journal of Technology, Learning, and Assessment, 10*(5). Retrieved [date] from http://www.jtla.org.

J·T·L·A

Abstract:

This paper represents one outcome from the *Invitational Research Symposium on Technology-Enabled and Universally Designed Assessments*, which examined technology-enabled assessments (TEA) and universal design (UD) as they relate to students with disabilities (SWD). It was developed to stimulate research into TEAs designed to make tests appropriate for the full range of the student population through enhanced accessibility. Four themes are explored: (a) a construct-centered approach to developing accessible assessments; (b) how technology and UD can provide access to targeted knowledge, skills, and abilities by embedding access and interactive features directly into systems that deliver TEAs; (c) the possibility of incorporating scaffolding directly into innovative assessment items; and (d) the importance of investigating the validity of inferences from TEAs that incorporate accessibility features designed to maximize validity. The article conveys the issues arising through the symposium and offers insights to researchers who conduct studies on the design, development, and validation of technology-enabled and universally designed assessments that include SWD. The paper proposes a focused research agenda and makes it clear that a principled program of research is needed to properly develop and use technology-enabled and universally designed educational assessments that encourage the inclusion of SWD. As research progresses, TEAs need to improve how they assess students' understanding of complex academic content and how they provide equitable access to all students including SWD.

J·T·L·A

Technology-Enabled and Universally Designed Assessment: Considering Access in Measuring the Achievement of Students with Disabilities—A Foundation for Research

Patricia Almond
University of Oregon
Phoebe Winter
Pacific Metrics Corporation
Renée Cameto
SRI International
Michael Russell
Boston College
Edynn Sato
WestEd
Jody Clarke-Midura
Harvard University
Chloe Torres
Measured Progress
Geneva Haertel
SRI International
Robert Dolan
Pearson
Peter Beddow
Vanderbilt University
Sheryl Lazarus
University of Minnesota

Overview

The Invitational Research Symposium on Technology-Enabled and Universally Designed Assessments was held in Arlington, Virginia, on July 23, 2009. Measured Progress and SRI International sponsored this meeting focused on the emerging and dynamic field of technology-enabled assessments (TEA) and the principles of universal design for assessment as they relate to students with disabilities. The symposium brought a group of researchers together from several areas of expertise including educational technology, cognitive psychology, students with disabilities, universal design for learning, and educational assessment. Among the participants were researchers who had completed or were engaged in research involving technology-enabled assessment, universal design for assessment, and/or students with disabilities, focused on two specific areas: cognition and access. The state of educational assessment and technology had recently been described in an article entitled *Beyond the Bubble: Technology and the Future of Student Assessment* (Tucker, 2009). Tucker drew attention to assessment challenges in the context of a cognitive model for assessment, and envisioned a future for assessment and technology that resolved many of the challenges. Tucker's article provided a foundation for the design of the symposium content, the meeting agenda, and motivated the plan to create a national research agenda that would capture the knowledge, expertise, and vision generated that day.

To launch the symposium deliberations, four cutting-edge research initiatives were conveyed to symposium participants. Chris Camacho with Children's Progress presented an adaptive and scaffolded assessment approach that provided prompts after incorrect responses and selected assessment items based on examinee responses to a previous question. Elizabeth (Boo) Murray, Center for Applied Special Technology, described an exemplar of universal design for learning, Strategic Reader that assesses maze and oral reading fluency via a web-based tool. Jody Clarke-Midura, Harvard Graduate School of Education, described immersive virtual performance assessments under development that are designed to assess knowledge and skills in science through items embedded within the context of virtual scenarios. Michael Russell of Nimble Assessment Systems demonstrated computer administered assessment tasks with embedded tools universally designed to facilitate access to content for students with special needs.

The presentations demonstrating assessment and technology innovations were followed by a large-group dialogue and discussion between the presenters and participants about the research initiatives including unique challenges and particular innovations. This discussion delved into the target areas of cognition and access and surfaced insights and questions arising from consideration of the future of TEAs. The debriefing session resulted in the large group dividing into two subgroups, one tackling issues regarding cognition and the other issues regarding access to assessment content. This article is based on the culmination of the symposium day plus the ongoing interactions among the subgroup members, who met to generate a research agenda regarding technology-enabled educational assessment and access for students with disabilities.

Twelve participants joined the symposium subgroup that addressed access to assessment content and students with disabilities. The twelve members of the access group included five researchers from universities, three from assessment publishers, three from research institutes, and one researcher from a national technology center. The subgroup members communicated via email and telephone conference calls over a nine-month period following the symposium. Eleven of the participants were contributing authors, writing components of this article. Three members of the symposium planning-team facilitated communication, assembled interim drafts, and assimilated revisions from subsequent reviews. Final edits were assembled by the facilitators and reviewed by contributing authors prior to submission for publication.

Introduction

The primary goal of this paper is to stimulate research into technology-enabled assessments (TEAs) that incorporate conditions designed to make tests appropriate for the full range of the student population through enhancing accessibility. We explore the concept of accessibility in TEAs, particularly as it applies to students with disabilities. In this context, four major themes related to access are explored. We begin with a description of a construct-centered approach to developing accessible assessments in which we emphasize the importance of preserving construct-related validity when developing methods for increasing access. The next two sections of the paper focus on how technology can be used to provide access to the targeted knowledge, skills, and abilities (KSAs) and the role that universal design plays in increasing accessibility—in the second section, we discuss embedding access and interactive features directly into systems that deliver TEAs and, in the third section, we examine the possibility of incorporating scaffolding directly into innovative items. The final theme addresses the importance of investigating the validity of inferences from TEAs that incorporate accessibility features designed to maximize validity.

This paper is aimed not only toward researchers but to policymakers as well. Recognition about how the opportunities for more accessible assessment can be afforded by technology-enabled assessment will bolsters support for needed research. Researchers and educators have a responsibility to explain the promises and issues inherent in TEAs. With their knowledge, they can formulate strategies for an ongoing program of research that does not hinder the use of TEAs but rather strengthens the way they are used. A program of research can allow TEAs to realize their potential to improve education and provide better data about academic achievement, particularly for students with disabilities.

Background

Between 2001 and 2010, the No Child Left Behind Act (NCLB) solidified the role of large-scale assessments in making summative judgments about student learning and school quality. Today, all 50 states and the District of Columbia administer annual tests to students across a wide range of grade levels. During this period, regulations stemming from NCLB and the Individuals with Disabilities Education Act (IDEA, 1997) and its amendments (2004) further solidified the importance of using assessment data from all students, including students with disabilities, in state and local accountability systems. Today, testing programs attempt to meet the needs of all students by providing a variety of test accommodations, and, in some cases, developing and administering alternate tests that are aligned with grade-level content standards. Students with the most significant cognitive disabilities participate in alternate assessments based on alternate achievement standards. Several states also have one or two additional alternate assessment options for some students with disabilities (alternate assessments based on modified achievement standards, alternate assessments based on grade-level achievement standards) (Lazarus & Thurlow, 2009).

According to Madaus, Russell, and Higgins (2009), large-scale summative assessments are composed primarily of multiple-choice items. Many state tests, however, also include short open-response items and/or extended writing items. A few states also employ extended problem solving or inquiry tasks that may require students to produce written responses, create tables or graphs, and/or to produce drawings or diagrams.

States are increasingly transitioning their assessment programs to computer-based administration (Tucker, 2009). Today, computers are used to administer either fixed-form tests that present items to students in a predetermined linear manner or adaptive tests that tailor the sequence of items presented to each student based on his/her response to prior items. Looking to the future, there is increasing evidence that states will continue to adopt technological solutions to enhance the efficiency and quality of their testing programs. In fact, in its Race to the Top Assessment Program, the U.S. Department of Education has launched a major initiative that could provide funding to develop and implement technology-enabled assessments that are more accessible for students with disabilities (U.S. Department of Education, April 2010). As an example, the SMARTER Balanced Assessment Consortium (2010), a consortium of 31 states, cites the use of computer technology within their assessment planning in several ways, for example, computer adaptive testing, computer-based simulations, and responses scored by computer.

As they design technology-enabled assessments, it is likely that test developers will experiment with innovative item types that require students to interact with more information and to demonstrate deeper levels of knowledge and understanding by manipulating information presented on a computer screen. For example, students may be required to manipulate digital representations of a microscope to locate and identify microscopic objects (e.g., an amoeba, a cell wall, mitochondria). Other items may require students to rearrange objects, such as line segments, to create shapes with specific characteristics (e.g., pentagon, perpendicular line, line of best fit). Students may also be presented with extended tasks that require them to conduct simulated experiments, such as separating a mixture into its separate compounds or determining the acidity of a substance. A task might involve searching for, selecting, and synthesizing information from a number of resources to support an interpretation of an historical event. Still other tasks may require students to engage in role-playing activities within virtual worlds to solve complex problems, such as assuming the role of a biologist who is trying to determine why a kelp forest is shrinking.

Access

As innovative item types are integrated into large-scale testing programs, it is vital that the needs of all students are considered during development so that TEAs are as accessible as possible to all students. Tucker (2009) notes, "New assessment models must not erode efforts to promote high expectations for all students" (p. 2). The goal underlying the design and use of TEAs is to obtain more valid inferences about the KSAs of students. For many students with disabilities and special needs, the validity of inferences is dependent on the accessibility of the items and tasks administered to them and with which they are required to interact.

Accessibility is a desired characteristic of testing by which students with various physical, cognitive, sensory, or linguistic barriers are provided the opportunity to demonstrate the KSA intended for measurement—the targeted KSA (Winter, Kopriva, Chen, & Emick, 2006; Ketterlin-Geller, 2008; Beddow, Elliott, & Keller, 2009). As such, accessibility is a prerequisite to validity, the degree to which a test score interpretation is justifiable for a particular purpose and supported by evidence and theory (AERA, APA, & NCME, 1999; Messick, 1989). Tests that require students to possess KSAs that are orthogonal to the intended constructs introduce construct-irrelevant variance into scores and hence compromise validity. A current example of reducing barriers and increasing accessibility is the provision of a read-aloud assessment administration for some students with learning disabilities who use recorded voice and/or text-to¬speech accommodations in the classroom. The challenge is that accessibility is not a static property of tests, but rather represents an interaction among test features and person characteristics that either permit or inhibit student responses to the targeted measurement content (Dolan & Rose, 2000; Winter et al., 2006). To continue the earlier example, a student may not benefit from a standard read-aloud version of a test if it is distracting to the student; in this case, another testing condition or accommodation such as a student-controlled text reader might be more beneficial.

Accessible Testing Through Universal Design

To address the challenge of designing and delivering tests that are accessible to and accurate for a wide range of students, the principles of universal design (UD) (Mace, 1991; Mace, Hardie & Place, 1996; see sidebar) have been applied to the design, construction, and delivery of tests. The core tenet of UD is to create flexible solutions that avoid post hoc adaptation by considering from the start the diverse ways in which individuals will interact with their environment. Rose and Meyer (2000, 2002) created a pedagogical application of universal design, which is referred to as Universal Design for Learning (UDL). The influence of UD on assessment is found in the contribution of Dolan and Hall (2001, 2007) who proposed that tests be designed to minimize potential sources of construct-irrelevant variance by supporting the ways that diverse students interact with the assessment process. Thompson, Johnstone, and Thurlow (2002, p. 1) adapted Mace's original elements from architecture to derive seven elements of accessible and fair tests: "(1) inclusive assessment population; (2) precisely defined constructs; (3) accessible, nonbiased items; (4) items amenable to accommodations; (5) simple, clear, and intuitive instructions and procedures; (6) maximum readability and comprehensibility; and (7) maximum legibility." Ketterlin-Geller (2005, p. 5) provides a more generic definition of universal design for testing: an "integrated system with a broad spectrum of possible supports" that permits inclusive, fair, and accurate testing of diverse students.

Glossary: Universal Design

Universal Design
The universal design concept was founded in the 1980s within the field of architecture by Ron Mace (1991; 1996) at North Carolina State University. The goal of universal design is to build structures and products that are inherently accessible by considering individuals' diverse needs from the outset (e.g., mobility and communication needs), thus reducing the need for retrofitting. As such, universal design seeks to minimize assumptions about how individuals will interact with what is being built. Television captioning provides a good example of universal design in practice. While originally intended for people with hearing impairments, who otherwise needed to retrofit their televisions by purchasing expensive decoder boxes to access the captions, captioning became standard and ubiquitous through legislation that called for building the feature into all televisions. This universal design feature now benefits not only those with hearing impairments, but far more individuals in health clubs, bars, and airports as well as individuals working on their language skills and couples who go to sleep at different times. Further, as a built-in feature, access to television captioning costs a few cents rather than several hundred dollars.
Universal Design for Learning
The educational framework of universal design for learning extends universal design from a physical space to a pedagogical space (Rose & Meyer, 2000, 2002), utilizing recent discoveries and advances in the cognitive sciences and digital technologies. It guides the design of curricula, materials, and assessments that are more accessible to most students, including those with disabilities and who are English learners. Flexibility is accomplished by accounting for individual differences in how students recognize, strategize, and engage in learning situations by providing the following: • Alternative formats for presenting information (multiple or transformable accessible media) • Alternative means for action and expression (writing, drawing, speaking, switch, use of graphic organizers, etc.) • Alternative means for engagement (background knowledge, options, challenge, and support) By providing these alternatives in flexible and customizable ways, UDL seeks to minimize learning barriers and maximize learning opportunities.

These applications of universal design share common elements. First, they propose a solution to accurate testing of the full student population, in particular students with disabilities and/or students who are English language learners. Second, they propose that test accessibility is best accomplished by considering the needs of all students *from the beginning* rather than trying to retrofit assessments later. Accommodations have been the typical solution to including students with disabilities in general assessment programs; when students receive appropriate accommodations, they are able to more meaningfully participate in the assessment (Thurlow, Thompson, & Lazarus, 2006) and thus provide a better indication of their knowledge, skills and abilities. However, no matter how well designed, administered, and utilized they are, accommodations are post hoc retrofits. In addition, some accommodations may affect students' performances differently on different items by introducing sources of construct-irrelevant variance or providing unintended construct-relevant information. As a result, they may fail to provide students adequate support and can compromise validity (Dolan & Hall, 2007). Finally, optimal testing provides choice in testing conditions to test administrators and/or students that takes into account the diverse ways in which construct-irrelevant challenges and disabilities are manifest and can be supported in ways that "one-size-fits-all" solutions rarely do (Rose & Meyer, 2000).

Recent research has suggested that applying universal design principles during test development and delivery can indeed improve testing of students with disabilities. For traditional multiple-choice and short-response items, research has demonstrated that principles of UD can be used to make a more inclusive testing environment that reduces the need for accommodations (Dolan, Hall, Banerjee, Chun, & Strangman, 2005; Lazarus, Thurlow, Lail, & Christensen, 2009; Russell, Hoffmann, & Higgins, 2009), and reduces construct-irrelevant variance (Johnstone, Bottsford-Miller, & Thompson, 2006).

The goal of UD is not to create a single assessment condition that is accessible for all students (Rose & Meyer, 2000). Instead, a universally designed assessment will anticipate the variety of accessibility needs of potential students and build in methods that allow all students to access, engage with, and respond to test content in the most accessible manner possible. There are two important steps to developing a universally designed assessment. First, test content must be developed in a way that anticipates the different representational needs of students and representational forms that meet those needs without violating the test construct. Second, the system employed to administer test items to students must be designed to flexibly alter the presentation, interaction, and response to items, and tailor access to alternate representations based on each student's individual need. When successfully executed, a universally designed assessment shifts the consideration of student/test interactions from determining post hoc changes required and providing test accommodations to a priori design and administration decisions and development of alternate representations during the item and test development stage (e.g., Kettler, Elliott, & Beddow, 2009; Ketterlin-Geller, 2008; Thompson, Johnstone, & Thurlow, 2002).

UD can also guide the development and delivery of "innovative items," those that use digital technologies to test students on greater depths of knowledge and skill than traditional items. For the reasons already stated, it is imperative that such items be accessible, accurate, and fair for a diverse range of students. To the extent that such items involve novel interfaces and tasks, they potentially introduce new forms of construct-irrelevant variance. A framework and guidelines for applying UD principles to the creation of innovative TEA has recently been proposed (Dolan et al., 2006; Dolan, Rose, Burling, Harms, & Way, 2007) with emphasis on interactions between students and test features as a function of individual differences in perceptual, linguistic, motoric, cognitive, executive, and affective processing during item presentation, strategic interaction, and response action.

Access and Interaction between Student Characteristics and Test Features

The term *access* is widely used in discussions of education policy and practice, particularly with regard to special student populations (e.g., students with disabilities, English learners). Recent research that examines the interaction between student characteristics and test or item features has appeared (e.g., Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000; Helwig, Rozek-Tedesco, & Tindal, 2002; Ketterlin-Geller, Yovanoff, & Tindal, 2007; Sato, Rabinowitz, Gallagher, & Huang, 2010), as has research evaluating the effects of test access on student performance (e.g., Abedi, Courtney, & Leon, 2003; Dolan, Hall, Banerjee, Chun, & Strangman, 2005; Rivera & Stansfield, 2001; Tindal & Ketterlin-Geller, 2004). The definition of access and the enumeration of elements that characterize effective access are still emerging (Sato et al., in press).

Each student has unique characteristics. When student characteristics cause construct-irrelevant item features to interfere with the student's opportunity to demonstrate knowledge, skills, or ability, access decreases and in turn, the degree to which the response reflects the student's achievement decreases. Item types intended to measure the same construct might require different processing by examinees, depending on the specific processing requirements of the item in terms of its complexity and structure (Messick, 1994; Pearson & Garavaglia, 2003; Russell, Goldberg, & O'Conner, 2003; Thissen, Wainer, & Wang, 1994). According to Pearson and Garavaglia (2003), while two items may be *psychometrically* equivalent, they may not be psychologically equivalent—the items may require students to access the content in different ways, subsequently affecting their processing. As a result, the items may measure either skills or knowledge that differ from the intended content (construct irrelevance) or may provide processing challenges that interfere with the student's ability to fully demonstrate what he or she knows and can do (underestimation). Therefore, a degree of flexibility in test design and delivery may be necessary to best ensure that students with disabilities and others have access to tested content, enhance comparability in scores (Ketterlin-Geller, 2008), and create assessments that support more valid inferences across subgroups of students.

Glossary: Access

Access
For the purposes of this paper, access in assessment depends on the interaction between construct-irrelevant item features and person characteristics that either permits or inhibits student response to the targeted measurement content (Winter, Kopriva, Chen, & Emick, 2006).

Test developers and users have often resisted the idea of flexibility in presentation and administration of test items and tasks, since standardization in these areas has been used as a basis for making common inferences across students, situations, test forms, and other conditions that vary in testing (see AERA, APA, & NCME, 1999). However, recent consideration has been given to the idea of purposeful flexibility in test design and delivery that can support comparable and valid score-based interpretations across student populations and testing conditions (Marion & Pellegrino, 2006; Sato et al. 2010). In order to best ensure effective flexibility, the interaction between student characteristics and the features of the test itself need to be understood (Ketterlin-Geller, 2008; Marion & Pellegrino, 2006; National Research Council, 2001).

Research revealing the nature of student-item interactions is emerging. For example, researchers have extended the assessment triangle (National Research Council, 2001) to develop frameworks that support understanding of how students represent knowledge in a domain and the types of observations that demonstrate learning, including the interaction between student characteristics and assessment techniques (Ketterlin-Geller [2008] and Marion and Pellegrino [2006] offer more detail on this subject). Frameworks for understanding interactions between categories of test taker characteristics and features of the test itself, including both targeted and ancillary interactions that affect construct-irrelevant variance in test scores, also have been developed (Mislevy & Haertel, 2006; Beddow, in press; Dolan, Burling, Harms, & Way, 2007). Additionally, systematic error has been examined (e.g., Kopriva, Wiley, & Winter, 2004); that is, the systematic ways in which target skills may be contaminated, misunderstood, or distorted by specific task factors such that the error influences task performance and affects the accurate measurement of students' knowledge and skills and the validity of interpretation of results.

In short, the concept of increasing accessibility to assessments for students with disabilities and others has been the focus of a number of recent efforts. Frameworks such as the ones mentioned can provide a start for organizing the findings around accessibility, identifying the principles that can be applied to assessments now and highlighting the issues that should be researched as accessibility is addressed in the design and development of TEAs. For the purposes of this white paper, we began with the assumption that students received adequate instruction in order to focus our discussion around accessibility. Students' opportunity to learn (OTL) academic knowledge, skills, and abilities as well as methods for demonstrating these KSAs is prerequisite to assessing achievement. Strategies for evaluating and improving OTL for students with disabilities are beyond the scope of this discussion.

A Construct-Centered Approach for Designing Accessible Assessments

All good educational assessment design begins with the definition of the targeted KSAs that the test intends to measure (Mislevy & Riconscente, 2006). There may be ancillary or non-targeted KSAs required for successful performance on particular tasks in an assessment, but they are not the focus of measurement. Ancillary KSAs can pose access barriers to some students. For example, one type of ancillary KSA requires that students be able to perceive the material being assessed (e.g., perceiving text on a page or computer screen). Some ancillary KSAs are prerequisites to the targeted KSAs. For example, some computation may be required in an assessment task that targets using the appropriate procedures for estimating the probability of a particular event. Both targeted and ancillary KSAs must be considered when designing an assessment, to uphold the validity of interpretations derived from test scores—the set of targeted KSAs or intended constructs must be held central (Kopriva et al., 2004; Abedi et al., 2003; Zhang et al., 2009, 2010). To continue the example, the required computation may be kept simple so that it is more likely that the task is indeed measuring skill in the targeted area— determining the procedure for estimating probability rather than computation; in a constructed-response item, the scoring rubric may ignore the accuracy of the computation and focus on characteristics of the procedure, or students may be provided with a calculator or other tool to assist with the computational aspects of a task. Messick (1994) explained clearly the necessity for such a grounded approach based upon thorough understanding and explication of constructs:

> A construct-centered approach would begin by asking what complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics (p. 17).

Glossary: KSAs

What are KSAs?
KSAs refer to the knowledge, skills and abilities being assessed. • Focal (or targeted) KSAs. The primary KSAs being assessed. • Additional (ancillary or non-focal) KSAs. Additional KSAs that may be required for successful performance on an assessment item or task but are not the primary focus. Some additional KSAs can be supported by the principles of Universal Design (UD) and testing accommodations.
What is the origin of the term KSA?
Mislevy and Riconscente (2006, p. 62) attribute the phrase knowledge, skills, and abilities to industrial psychologists who use KSAs to refer to the targets of the inferences they draw. They borrow the term and apply it more broadly for assessment to "the nature of the targets of inference and the kinds of information that will inform them… ."

Identifying targeted and ancillary KSAs is a two-step process. The first step is to prepare a clear statement of a test's targeted KSAs that are associated, for example, with content standards of interest in a subject matter domain. Demonstrating achievement on each standard likely requires multiple skills, some of which are targeted and some of which are ancillary. The identification of the ancillary KSAs is the second step in the design process. By identifying targeted versus ancillary KSAs that may be involved in measuring each standard, assessment designers are able to isolate potential sources of construct-irrelevant variance. They are then able to design supports that assist students in overcoming barriers presented by the ancillary KSAs. The validity of the inference made about a student's performance based on such a process is likely to be improved.

While there are some approaches for identifying item/task-level KSAs, few focus explicitly on identifying ancillary KSAs particularly as they apply to the design of assessments for diverse learners. Important exceptions are Kopriva (2008) in the context of English language learners, Solano-Flores & Li (2006) and Stansfield (2003) in the areas of translation and item templates, and Abedi et al. (e.g., Abedi, Courtney, & Leon, 2003; Abedi, Courtney, Mirocha, Leon, & Goldberg, 2005) and Sato et al. (in press) in linguistic modification. In the area of students with disabilities, DeBarger, Haertel, Villalba, and Colker (2009) and Cameto, Haertel, DeBarger, and Morrison (2010) have identified ancillary as well as focal KSAs. This approach has also been implemented in some alignment procedures such as those for alternate assessments (Flowers, Wakeman, Browder, & Karvonen, 2007) and in Achieve's alignment review process (Rothman, Slattery, Vranek, & Resnick, 2002).

Evidence-Centered Design as a Construct-Centered Approach

Evidence-centered assessment design (ECD) is an approach to creating educational assessments in terms of evidentiary arguments built upon intended constructs, with explicit attention paid to the potential influence of unintended constructs (Mislevy, Steinberg, & Almond, 2003). ECD accomplishes this in two ways. The first is by incorporating an overarching conception of assessment as an argument from imperfect evidence. This argument makes explicit the claims (the inferences that one intends to make based on scores) and the nature of the evidence that supports those claims (Hansen & Mislevy, 2008; Mislevy & Haertel, 2006). The second is by distinguishing the activities and structures involved in the assessment enterprise, in order to exemplify an assessment argument in operational processes. By making the underlying evidentiary argument more explicit, the framework makes operational elements more amenable to examination, sharing, and refinement. Making the argument more explicit also helps designers meet diverse assessment needs caused by changing technological, social, and legal environments (Hansen & Mislevy, 2008; Zhang et al., 2009).

The ECD process involves five layers of activities. The layers focus in turn on the identification of the substantive domain to be assessed; the assessment argument; the structure of assessment elements such as tasks, rubrics, and psychometric models; the implementation of these elements; and the way they function in an operational assessment, as described below.

1. Domain Analysis involves determining the specific content to be included in the assessment. Use of state content standards and the pending common core standards are examples of domain analyses.

2. In Domain Modeling, a high-level description of the overall components of the assessment is created and documented. Design Patterns used by Mislevy and Haertel (2006) exemplify this layer.

3. The Conceptual Assessment Framework is developed. In this layer, the KSAs to be assessed, the evidence that needs to be collected, and the features of the tasks that will elicit the evidence are specified in minute detail. Ancillary KSAs that may be required to respond correctly to an assessment task but are not the intended target of the assessment are also specified (for example, reading skills in a mathematics examination). By identifying these ancillary KSAs, construct-irrelevant variance can be minimized in item and task development—potential barriers created by the ancillary KSAs can be removed or their effects reduced through the provision of appropriate access features.

4. Implementation involves the development of the assessment items or tasks using the specifications created in the conceptual assessment framework just described. In addition, scoring rubrics are created and the scoring process is specified.

5. In Delivery, the processes for the assessment administration and reporting are created.

Combining ECD and UD for Accessible, Construct-Centered, Technology-Enabled Assessments

Neither ECD nor UD alone, as independent approaches to assessment design, can assure accessible assessments. ECD's strength is its explication of evidentiary arguments and processes for maintaining a focus on the intended target of measurement. However, the interactions between a student and an assessment are complex, especially when considering the diverse ways in which students approach learning, engage in instruction, and express what they know and can do. For an ECD approach to succeed with a range of students, the interactions between test takers and test items must be taken into consideration. While the principles of UD help identify the barriers that can limit the performance of students with diverse learning needs and ways to overcome these barriers, UD-based generalizations about student abilities and challenges alone also are inadequate. Design decisions must be made carefully at the construct and item levels with attention to both the diversity of learners, as represented in UD, and the need for evidence-centered design, as represented in ECD. This is especially true for technology-enabled assessments, where the opportunities for student and test interaction can be greater and are less understood.

As mentioned in the introduction, recently UD principles have been applied to the development of flexible, new media-based learning environments that support a wide range of learners, including those with disabilities and those who are English language learners (Rose & Meyer, 2002; Dolan et al., 2006). This elaborated, UD-based framework includes (1) test delivery considerations, (2) item content and delivery considerations, and (3) component content and delivery considerations. Alone, application of this three-tiered approach can help develop technology-enabled assessments that are likely to increase accessibility for a range of students. Adding an explicit construct-centered and validity argument-based approach toward assessment, such as that provided by ECD, to these efforts will strengthen their effects of accessibility.

ECD and UD frameworks are being combined to support the design of science assessment tasks as part of an IES-funded project titled Principled Science Assessment Designs for Students with Disabilities that is being conducted by SRI International, the University of Maryland, and the Center for Applied Special Technology (CAST) (DeBarger et al., 2009). In this project, the web-based PADI assessment design system is being augmented to explicate the types of processing (i.e., perceptual, linguistic, cognitive, motoric, executive, and affective) students engage in while interacting with test items and tasks. Test developers specify the targeted and ancillary KSAs and then select ways to increase accessibility on the items/tasks by supporting students' performance on the ancillary KSAs. For example, if vocabulary is an ancillary KSA, the assessment designer could select task features from a list of features that support linguistic processing. These support features, if implemented during the assessment, would provide students with the vocabulary words needed to overcome barriers presented by language and symbols that might inhibit their successful performance on the targeted KSAs. For example, strategies to support lack of technical vocabulary might include embedded support for key terms with a technical glossary, hyperlinks, or footnotes to definitions.

Extending the application of UD and ECD frameworks to assessment design and task development for students with significant disabilities, Cameto and colleagues (2010) are conducting research under two Enhanced Assessment Grants funded by the U.S. Department of Education. These two projects, conducted in close collaboration with consortia of states, are implementing several layers of the ECD process, including domain analysis, domain modeling, specification of the conceptual assessment framework, the authoring of exemplar assessment items and tasks, and the design of the assessment delivery system. The principles of UD are implemented during the domain modeling and task authoring layers of the ECD process. The next steps for the projects combining UD and ECD involve validation studies of the items and tasks developed using the combined UD and ECD frameworks. (You can view http://padi-se.sri.com/ for more details.)

Embedded Features Designed to Facilitate Access

Two strategies for changing tests and test items have been used to increase the access students with disabilities have to achievement testing: testing accommodations and universal design for assessment (Tucker, 2009). Testing accommodations involve changes to the standard materials and procedures employed to measure a given construct and are intended to decrease the effect that ancillary constructs have on a student's test performance. Accommodations are intended to ensure access for individual students and address particular student needs (e.g., Hollenbeck, 2002). Based on a comprehensive review of state testing program policies regarding accommodations, Thurlow et al. (2006) identified five categories of test accommodations: (1) presentation (e.g., large print booklets, Braille, signing), (2) equipment and/or materials (e.g., magnifying glass, noise buffer), (3) response methods (e.g., scribes, keyboard, pointing devices), (4) schedule and timing (e.g., extended time, breaks, multiple test sessions), and (5) setting (e.g., separate room, individual administration, carrel). Original conceptions of accommodations have often derived from paper-and-pencil testing formats.

As explained in the introduction, the application of the principles of universal design during test design and development produces tests and administration procedures that provide flexibility and access for all students, including students with disabilities. There are at least four components of accessibility through universal design that can be applied to technology-enabled assessments: (1) flexibility in the way test content is presented, (2) flexibility in the way students engage with test content, (3) system compatibility with a variety of assistive technologies (e.g., touch screen, single switch devices, alternate keyboards such as Intellikeys, speech-to-text software), and (4) availability of alternate representations by presenting students with alternate versions of text-based content. Reading aloud content, translating text-based content into sign language or Braille, tactile representations of graphical images, symbolic representations of text-based information, narrative representations of chemical compounds (e.g., "sodium chloride" instead of "NaCl"), and translating to a different language are all forms of alternate representations. Going forward in this paper, we refer to TEAs that incorporate universal design as technology-enabled and universally designed assessment (TE/UDA).

Mislevy, Wilson, Erickan, and Chudowsky (2001) indicate that alternate presentations change the form in which test content is presented to a student. At some point, altered presentations change the construct being assessed and before that point, they simply provide different versions of the same test content. In some instances, altered presentation in TE/UDAs can involve changing contrast, sizing, spacing, and so on. In the area of accommodations, changes have been viewed as residing along a continuum (Tindal, 1998) from accommodations to modifications. Within the context of digital technologies, the distinction between accommodated (same construct) and modified (different construct) becomes a multidimensional one. Within TE/UDA, there are numerous alterations, for example, changing the sensitivity of a mouse, making glossary definitions available with the click of a mouse, highlighting salient content, and masking potentially distracting content. Applying universal design concepts to educational assessment using digital technologies opens additional possibilities to increase usability for a diverse range of students, providing solutions not readily available through traditional paper-and-pencil testing approaches (Bennett, 1999; Burk, 1999; Dolan & Rose, 2000).

Technology-enabled testing can offer tools embedded into the assessment platform and reduce the need for after-the-fact accommodations. For example, instead of separately printed test booklets with enlarged print for students with reduced vision, a magnification tool can be embedded into the assessment delivery system; rather than relying on a teacher to read aloud test content to groups of students, students can independently have the computer provide read-aloud content on an individual, as-needed basis. In fact, a variety of digital technology features that promote access to the test content or student response can be embedded into the same testing program. Some of these features may be designed for and offered to all students, while others may be available only to students who need them because of their disabilities. For example, it is conceivable that a read-aloud tool on a science test may be available to all students, while a tool that translates words into sign language is not. In any case, applying universal design to TE/UDA offers the potential to allow all students to benefit from a testing environment that adapts to meet individual needs.

Building on universal design principles, TEA delivery, item writing, test design, and test development procedures could include alternate representational forms of item content and allow for alternate representational forms for student responses. Additional considerations for TE/UDA design, development, and delivery include the ability of access features that function analogously across operating systems and that function with all elements of the test, including the test items, directions, reference and formula sheets, calculators, and other digital technologies (e.g., protractors, rulers, magnifiers). Additional considerations involve the interaction between multiple features used simultaneously, the automatic recording of students' use of particular features, as well as the option to support students' offline research and analysis.

Embedded Digital Technology Features for Accessibility

Recent innovations and technological advances have permitted the individualization of specific access strategies without the demand for human resources that has been required for testing accommodations. Many computer based assessment applications such as the Kansas Computerized Assessment (http://www.cete.us/kap/), NimbleTools (http://nimbletools.com), TestNav (http://www.pearsonassessments.com/TestNav), and iTest (http://measuredprogress.org/) integrate common accessibility features into test delivery systems for general student use and provide additional access features that can be tailored to the test taking experience based on each student's individual needs and the unique interface offered by the particular technology-enabled assessment.

Once integrated into an assessment platform, these systems offer educators the opportunity to create individual student accessibility profiles prior to the administration of an assessment. An accessibility profile specifies the presentation and interaction tool options, alternate representations, and alternate response methods required for a particular student. The test delivery system can employ student accessibility profiles to tailor the test administration, that is, access features and representational forms available to a particular student during assessment. Alternatively, access tools integrated into the system could be selected by the student as needed.

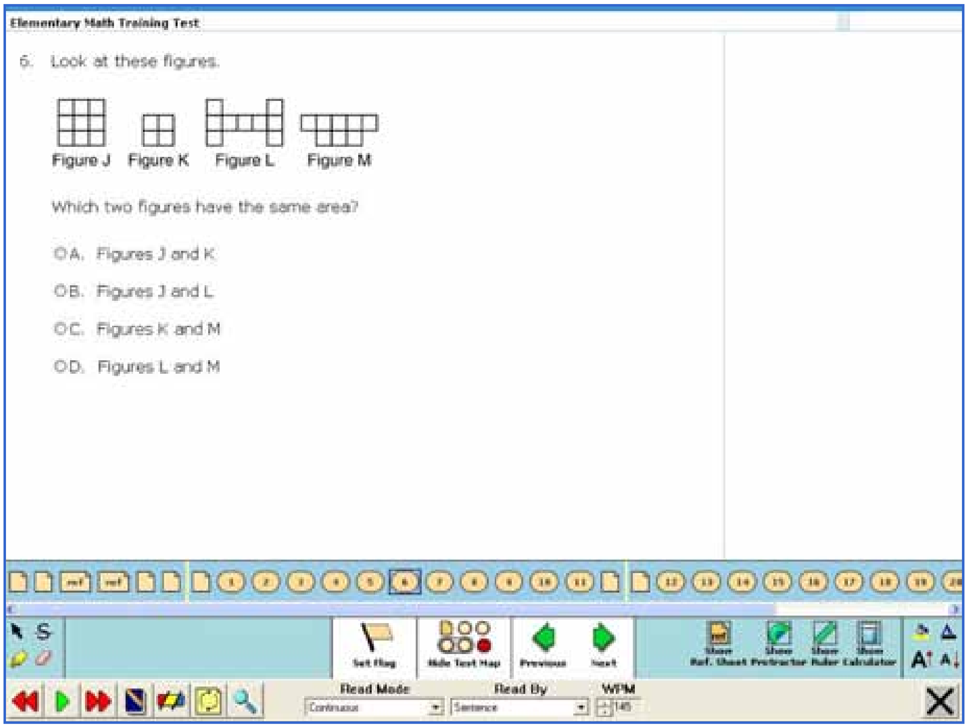
In applying principles of universal design, some vendors, including Measured Progress, NimbleTools, and Pearson, are embedding access features that address task comprehension, interaction, and response into their assessment delivery systems. These features include magnification, high contrast, altered color contrast, masking, and alternate representations of content such as verbal representations of text, tables, formulas, scientific notation, and graphics; Braille; signed English; American Sign Language; and other languages such as Spanish. Access features provide students with interaction and response options such as alternate keyboards, single-switch devices, writing supports, and speech recognition. Based on each user's individual needs, assessments with a variety of access features can tailor the availability of the features to ensure that each student has access he or she requires without being distracted by tools that are not needed. Tailoring the assessment environment to meet students' access needs allows students to more accurately demonstrate their KSAs, leading to more valid inferences about their achievement.

Examples of Technology-Enabled and Universally Designed Assessments

It is possible to increase accessibility within the framework of an existing assessment program using UD assessment software. For example, in Kentucky, assessment regulations required that a student who used assistive technology regularly in the classroom must be provided the option to use the same software during the assessment if the assistive technology usage was documented in the student's IEP. Because the state testing program was paper-based, online testing was offered as an accommodation for students who met the regulation's criteria. A subset of the existing paper test forms were delivered online via Measured Progress' *iTest* system,[1] which allowed students to use multiple assistive technology devices that were already in place in Kentucky schools on the assessment. This targeted approach was intended to drive the transition from paper-based to online assessment by building awareness and evaluating/updating the technology infrastructure while increasing access for the students who would benefit the most from online testing. This example, for practical purposes, is at the lower end of the spectrum of TE/UDA possibilities and describes a logical entry point for states that need to start with an existing item bank designed for paper-based assessment. Ideally, as states move forward, TEA platforms will integrate universally designed test content with universally designed assessment software during the design and development phase.

Measured Progress's *iTest* system provides a combination of built-in accessibility tools, such as high-and altered-color contrast and variable screen layouts, along with the ability to interact with third-party assistive technology products that can be either embedded or externally accessed. While built-in functionality and embedded tool bars allow all students to take advantage of a common set of accessibility tools, allowing access to external tools gives students with specific accessibility needs the option to utilize the same assistive technology software with their preexisting user profiles that are used in the classroom on a regular basis. Figure 1 presents a basic multiple-choice item, but shows a combination of *iTest* tools plus the Kurzweil[2] toolbar. The figure illustrates an example of an online test that interacts with third-party software/assistive technology. In the figure, the Kurzweil tool bar is the grey bar at the bottom of the screen. Above that is the *iTest* tool bar. On the right, there are tools for font size increase/decrease, for high contrast color palette options, and on-screen math tools. On the left, there are a strikethrough and highlighter tools.

Figure 1: *iTest* with Access Tools and the Kurzweil Toolbar



Source: Screen shot taken from Measured Progress's iTest system

Universal design can also be incorporated into innovative item formats. Figure 2 shows a science item that was designed and developed using UD guidelines (Dolan et al., 2006). The guidelines were used to minimize the introduction of sources of construct-irrelevant variance and to inform the user interface design in relation to the various modes of student interaction. In addition to having test-based accessibility supports such as magnification and read-aloud tools, the item is accessible through the keyboard (i.e., does not require pointer control) and can even be made single-switch compatible. Accomplishing this level of accessibility post hoc is difficult. For TE/UDA, accessibility and the provision of access features must be considered during assessment and item design.

Figure 2: Screen Shot Showing Universal Design with an Innovative Test Item

Source: Screen shot provided by Pearson (Assessment & Information)

The Benefits of Built-in Features

There is a need for a systematic approach to identifying individual access needs so that resulting interactions with testing features can be understood (Ketterlin-Geller, 2008). To date, testing accommodations have been selected individually for each student with a disability by the Individualized Education Program (IEP) team. Evidence indicates that when implemented with integrity, accommodations may result in differential score improvements for students with disabilities (Kettler & Elliott, in press); though other studies have found contradictory and inconclusive results (Thurlow, Lazarus & Christensen, in press). However, there is also evidence to suggest that accommodations are not always implemented as prescribed or may be implemented with poor fidelity. When accommodations are given to students without their input, some students dislike certain testing accommodations and may refuse to use them during testing or resent having them provided during testing (Elliott & Thurlow, 2005). To the extent that accommodations are implemented with poor fidelity or that students dislike using accommodations, the prescribed accommodations will be ineffective (Ketterlin-Geller, Alonzo, Braun-Monegan, & Tindal, 2007).

Recent research, however, suggests that many students are more willing to have their access needs addressed when access features are integrated into technology-enabled universally designed assessments. Recent implementations of TE/UDA found that nearly three times as many high school students opted to employ a "test accommodation" provided within a technology-enabled universally designed assessment compared to a paper-based version of the same test (Russell, Hoffmann, & Higgins, 2009). These implementations also found that the influence of ancillary constructs (e.g., reading and mathematics skills) was decreased when students performed the test using the TE/UDA as compared to the paper-based version. Further research is needed, however, to examine the similarities and differences between testing accommodations employed during more traditionally paper-based assessments and technology-enabled and universally designed assessments that have access features integrated into the test delivery platform.

Research is needed on several fronts, corresponding with the context of the old (accommodations) and the new (TE/UDA) approaches for increasing access when testing students with disabilities. Specifically, evidence is needed on the degree to which specific access strategies compromise or improve the inferences that can be made based on test scores. Research is also needed on the methods for selecting access features for individual students with disabilities who have individual needs. As with accommodations, no student or group should be given an unfair advantage over other students. Both test validity and equity need to be considered in the research (Sireci, Scarpati, & Li, 2005).

Using Scaffolding in Technology-Enabled Assessment

"Scaffolding"[3] has been used as a term to describe providing supports for student learning for a number of years; a recent definition is "explicit and sequentially organized support and guidance about possible strategies" (NRC, 2001, p. 278). Most writing about scaffolding uses a concept akin to Vygotsky's (1978) zone of proximal development (ZPD) in discussing how to develop, structure, and use scaffolds in instruction. The ZPD is the zone between a students' actual developmental level and the student's potential developmental level; scaffolds modify the levels of learning opportunities so that they lie between the two extremes and help students progress toward targeted levels of learning. In instruction, scaffolding is used to help students gain access to content or concepts by providing supports geared to their current learning and/or cognitive capabilities.

Bruner and colleagues define scaffolding as "constructing simple, noiseless opportunities for a child to grasp the sense and reference of various signs" (Bruner, 1983, p. 336; Wood, Bruner, & Ross, 1976, cited by Bruner). According to Quintana, Krajcik, and Soloway (2002), scaffolds help students "do cognitive tasks that are just out of their current developmental and intellectual capability ... [and] guide and support learners, but in a way that learners still need to think about the work they are doing" (p. 3). In instruction, scaffolds are meant to be temporary supports—once the student has reached the targeted level of learning, the specific instructional scaffolds are no longer necessary (new scaffolds may be used to bridge the gap to the *next* level of learning, of course). The transient nature of scaffolding differentiates it from learning strategies that are meant to be more generalized, such as goal setting before reading or using organizational tools for note taking.

Although scaffolding is a common instructional technique used with students with and without disabilities and a natural part of formative assessment (Shepard, 2005; Chin & Teou, 2009), it has not been used widely in formal summative assessment settings. Part of the reason for this is practical—scaffolding for paper-and-pencil tests is not straightforward; part is because scaffolding changes the nature of what is being measured, often making the item or task less difficult in terms of the targeted construct, and we have not had an incentive to devise ways of scoring scaffolded tests that take into consideration the difference; and part is because large-scale summative assessments are based on a static model of knowledge and skills and have not yet been much influenced by the increased understanding of how students learn. Scaffolding, if used appropriately, might allow us to better measure students' knowledge and skills by providing supports to students that allow them to respond to a task at the appropriate entry level.

Technology allows us to build assessment tasks that provide students with the opportunity to use or not use construct-relevant supports when they encounter an item, in a manner similar to the use of hints in online homework systems or intelligent tutoring systems that present content and/or scaffolds adaptively as a function of ongoing evaluation of a student's KSA compared against a model of the constructs to be learned (Woolf, 2008). Appropriately used, scaffolds allow students who would otherwise get the item wrong to demonstrate what they do know about the item/task content. We do not yet know how to use scaffolds or other supports in large-scale assessments. A number of ways of providing support can be imagined, as follows:

- A student selects an incorrect option on a multiple-choice item; that option is removed and the student selects a response from the remaining options.
- A student provides an incorrect response; the student is given an item that cues the student to an appropriate strategy for approaching the construct of the initial item and is then presented with an item parallel to the initial item.
- A student has the option to ask for a demonstration or item starter after being shown an item—for example, in a drag-and-drop fill-in-the-blank item, the student may request that one blank be filled in. In an item asking the student to perform a task, the student can request a demonstration.

This use of scaffolds provides students who have partial understanding with the opportunity to respond to an item more fully. Scaffolds can allow us to maintain the same expectations for performance for all students while providing the opportunity for responses from students along the full range of the achievement continuum. Scoring rules will need to be carefully developed for scaffolded items to take into account whether and how a scaffold is used in a response.

For this article, we differentiate scaffolds in assessment from universal design features and other features students may use to obtain access to test content and provide appropriate responses. A scaffold is purposely designed to affect the knowledge, skills, and abilities (KSAs) required to respond to a task in a defined way,[4] while accommodations and access features are not expected to affect the targeted KSAs. If a student uses scaffolds when interacting with an assessment task, we make an inference about the student's construct-related KSAs different from that for a student who does not use scaffolding; if a student uses an accommodation or tool provided with the test in responding to an item, we make the same inference about the student's targeted KSAs as we would for a student who did not use the tool and provided an equivalent response. In short, accommodations and access features are designed to reduce or eliminate construct-irrelevant variance in test scores by removing barriers to responding to the task presented; scaffolds are designed to reduce construct-irrelevant variance in test scores by changing the task so that it accesses the students' construct-relevant KSAs better than the task would without scaffolding. The examples above show how the item/task difficulty might change with the use of scaffolding while the targeted content assessed by the item/task stays the same.

Developing summative assessment tasks that incorporate scaffolding could allow us to do a better job of measuring the KSAs of students whose performance is at the lower end of the achievement spectrum, including some students with disabilities and low-performing students without identified disabilities. Scaffolding could provide us with the same types of benefits attributed to the use of traditional computer-adaptive tests (CATs). One reason CATs are appealing is that they measure reliably across a wide range of achievement by presenting different items to students based on how they perform on previous items—if a student gets an item wrong, the student is presented with an item that will be easier for the student.[5] Scaffolds can be based on relationships among concepts using techniques such as concept mapping so that appropriate scaffolds are supplied based on the nature of a student's incorrect response. Scaffolded assessments, then, may allow for an approach other than the linear, difficulty-based approach of traditional CATs to provide opportunity to perform for students at lower achievement levels and thus may be able to provide more valid and instructionally relevant results.

Scaffolding in Educational Technology

Advances in educational technology have allowed for implementing instructional techniques in nontraditional platforms. Scaffolding is commonly built into instructional software, and researchers have begun to systematize how scaffolds should be developed and used in instructional technology. Quintana and colleagues (2004), for example, present a framework for incorporating scaffolding into instructional software. In a monograph describing research on adaptive technologies, Shute and Zapata-Rivera (2007) discuss how such technologies can provide support for learning and propose a framework for organizing these technologies. These frameworks, while developed for instruction, can provide a foundation for the principled use of scaffolds as part of TE/UDAs.

Scaffolding is also being used in technology-enabled short-and medium-cycle formative assessments.[6] For example, Feng, Heffernan, and Koedinger (2009) describe the ASSISTment system, which incorporates instructional assistance into its assessments. Children's Progress has developed a computer-based formative assessment and instructional system for use in grades pre-K through 3 that is based on identifying a child's zone of proximal development so that appropriate scaffolds can be presented if the student responds incorrectly to a task (Camacho, 2009). By nature, the scaffold affects the KSAs required to respond to the question, and thus must be taken into account in the overall results. The use of scaffolding is reflected in how each item is scored (correct without scaffold = 1 point, correct with scaffold = 0.5, incorrect with scaffold = 0) and then in the selection of the next item, resulting in a guided path tailored to the student's cognitive support needs (Children's Progress Academic Assessment [CPAA] Technical Report, 2009). Because the scaffold is provided only when the student responds incorrectly, the scaffold is automatically removed over time, as the student becomes proficient. The CPAA blends scaffolding, which is cognitively based, with computer adaptive testing (CAT), which is difficulty based. Clarke (2009) described an immersive virtual performance assessment (IVPA) that is under development. The IVPA is a 3-D virtual environment, based on an authentic ecosystem. Students take on the identity of a scientist and engage in inquiry practices and problem solving. While performing their tasks, students can use scaffolds such as asking virtual scientists for help. The IVPA is built around how students learn and demonstrate inquiry skills and is an example of using cognitive science in developing scaffolded assessments. Note that none of these systems change the expectations for performance through their uses of scaffolds. Scaffolds allow the students to demonstrate the degree to which they are meeting those expectations.

Investigating the Validity of Inferences Made from Technology-Enabled Universally Designed Assessments

All the issues discussed in this paper relate to improving the validity of inferences made from test scores by taking advantage of the opportunities TEAs afford to improve access to test items. Psychometric considerations influencing the valid use of assessment results (e.g., reliability, freedom from bias) that apply to paper-and-pencil assessments extend to technology-enabled assessments. However, such considerations and the approaches used to address them are likely to manifest themselves in different ways in TEAs, especially those designed to improve accessibility for students with disabilities. Currently, the effects of accommodations on student scores are typically studied in terms of whether the accommodation alters the targeted construct such that inferences about students using the accommodation are not comparable to those about students who do not use the accommodation (Abedi, Courtney, & Leon, 2003; Abedi, Hofstetter, & Lord, 2004; Sireci, Scarpati, & Li, 2005; Thompson, Blount, & Thurlow, 2002). Currently, the prevailing method for checking on the validity of accommodations relies on the interaction, or "differential boost" hypothesis. This hypothesis asserts that both groups may benefit from the accommodation but that students with disabilities must receive larger performance gains (Sireci, et al., 2005). It is unclear how this test applies in the context of TE/UDA where universal access to "accommodations" features is a possibility. Claims to validity are strongest when threats to validity have been removed or reduced. For students with disabilities, access is arguably the most relevant threat to validity because a lack of appropriate access can contribute to construct-irrelevant variance, misrepresentation of students' abilities, and construct underrepresentation (Abedi, Courtney, & Leon, 2003; Abedi, Hofstetter, & Lord, 2004; Bielinski, Sheinker, & Ysseldyke, 2003; Elliot et al., 1999; Helwig, Rozek-Tedesco, Heath, & Tindal, 1999; Kopriva, Samuelson, Wiley, & Winter, 2003; Sireci, Li, & Scarpati, 2003; Thurlow & Wiener, 2000). That is, inadequate access could result in the measurement of abilities that are not related to the intended test content (construct irrelevance). Inaccessibility could allow the student's disability to interfere with that student's ability to fully demonstrate what he or she knows and can do, and subsequently the test results could misrepresent or underestimate the student's targeted KSAs. Inadequate access also could affect the intended construct in that the assessment no longer sufficiently measures the targeted domain (construct underrepresentation) (Sato et al., 2010). Therefore, providing students with disabilities with appropriate access is critical to ensuring the validity of the assessment and inferences drawn from assessment results.

TE/UDAs offer functionality and flexibility to embed a broader range of tools in the assessment tasks presented to students than we can offer with after-the-fact accommodations to paper-and-pencil tests. The differences in how access features and accommodations are offered, and the impact that potential within-item variability across students could have on validity (construct and/or consequential), need to be considered during the design of the items and assessment. That is, there needs to be up-front (rather than after-the-fact) consideration of the potential for differential use of embedded access features across students or student groups for a given item or item set, and of how such differences might affect how we score the assessment and interpret its results. The impact of these potential differences needs to be examined during the development and implementation of the assessment in terms of the psychometric considerations that influence the validity of assessment results. Thus, through the process of TE/UDA design, development, and implementation, systematic examination and verification of the technical quality of the assessment and its items need to occur.

Research Questions

In this article, we explored the concept of accessibility in technology-enabled assessment through four themes:

- Employing a construct-centered approach to assessment
- Embedding accessibility features into assessment systems as a means of applying the principals of universal design
- Incorporating scaffolding directly into innovative assessment items
- Investigating the validity of inferences from TEAs that incorporate accessibility features

Although, the intended audience for this article consists primarily of researchers, the authors understand that policymakers play a key role in establishing research priorities. We believe that researchers and educators recognize the need to understand the promises and issues inherent in TE/UDAs and that they can assist policymakers in setting goals for research into these tests. The proposed research questions are concerned with increasing the validity of inferences that can be drawn from TE/UDAs about the knowledge, skills, and abilities of students with disabilities and all students. We hope that this white paper and the research questions raised here will trigger an ongoing program of research that strengthens the possibilities and allows technology-enabled and universally designed assessments (TE/UDA) to provide better data about academic achievement, particularly for students who, because of access needs, have not been able to demonstrate the full extent of their KSAs. These questions illustrate some of the areas where research is most needed. The questions and areas have been separated, but, undoubtedly, the research will address multiple questions or aspects of multiple questions within a single study.

There is a need to compile and adapt existing item/task development procedures and research so that they can be used to guide the development of access-based TE/UDAs in a consistent, scalable, and cost-effective manner. Questions to be addressed include the following:

- What standards, guidelines, and procedures are available to guide the incorporation of access features into TE/UDAs so that the effects of ancillary KSAs are minimized?
- Are there procedures or existing research that will assist in the simultaneous development of items/tasks that measure the same targeted KSAs but that allow for different approaches to the item/task, provide different types of representations of the tested content, or eliminate specific barriers to access?
- Which access features can be integrated into the design of a test, built in to the test development phase, and which must be applied on a case-by-case basis, to preserve the validity of test score inferences?
- How do student access needs change over time, and how can these changes be considered in the design of access features and selection procedures?

Evidence-centered design and universal design were discussed as processes that can be used to define targeted and ancillary constructs in a principled manner. These procedures can inform how and when access features and scaffolds can be included with items/tasks and assessments. Other systematic processes to identify targeted and ancillary constructs need to be developed or derived from existing research. Additional areas for research include the following:

- How do we validate processes for defining targeted and ancillary constructs?
- How do we determine whether the use of scaffolded items creates a test that measures a multidimensional construct or multiple constructs? Are ancillary KSAs added when scaffolding is used?
- Are there consistent variables within certain content areas and contexts that are the sources of ancillary KSA requirements?
- What is the relationship between ancillary KSAs and disability-specific barriers in items/tasks?

Features intended to increase accessibility are already being incorporated into technology-enhanced assessments. Research on the effects and efficacy of specific access features is needed. Examples follow:

- What are the most appropriate and efficient ways to individualize read-aloud features to meet specific student needs for reading assistance (e.g., chunking text)? Can synthetic speech be used in lieu of human voice-recorded audio without sacrificing the validity of test score inferences? What is the effect of different voices? Similarly, what is the effect of different voices (such as those with accents similar to the student's accent) and/or giving students a choice of voices?

- What kinds of tools can be built into tests that are appropriate for students with various physical disabilities (e.g., motor impairment that precludes the use of a mouse, vision impairment, deaf and hard of hearing)? What additional features should be made available for students with low-incidence access needs?

The idea of "scaffolding" or branching is emerging as a possibility in technology enhanced assessments—these techniques might allow us to obtain a better measure of students' KSAs through providing partial credit if students use hints or other construct-related supports. Research on the design and impact of scaffolds or similar supports is needed. The research will first need to address a broad range of questions such as the following:

- Under what conditions are scaffolds and access features distinctly different or on a continuum? When does the reduction of cognitive load common to both techniques transition from having no effect on the targeted construct to affecting the targeted construct?
- What types of models can be used to design scaffolded items and tasks, and how do they work in a summative assessment? Are models equally suitable for less-structured (e.g., literature, history) and well-structured (e.g., algebra, physics) domains?
- How can we systematically categorize scaffolding in assessment (e.g., effects on item difficulty, effects on level of complexity)?
- How do we design scaffolds to incorporate appropriate pathways to supports and subsequent items/tasks in scaffolded tests? How do we incorporate knowledge about students' changing needs for scaffolding in how we develop scaffolded items/tasks?
- How do we determine the effects of scaffolds on the difficulty of items/tasks?
- What scoring models can provide a basis for appropriate inferences about students' KSAs based on TE/UDAs that include scaffolded items? Can partial credit models be used to appropriately score scaffolded items/tasks or are there models other than those currently used in educational measurement that are more appropriate for scaffolded items?

We do not know the best ways to incorporate the use of access features, alternative representations and pathways, and scaffolds into technology-enhanced assessments. For example, should students determine whether to use the tools, should teachers select the appropriate tools for students, or should there be a mixture of ways to provide the tools? It is critical that students and teachers know how to take advantage of these features in a way that supports valid inferences. Some areas of research are as follows:

- How can students be taught to choose access strategies and features so that they are afforded the greatest opportunity to demonstrate their knowledge, skills, and abilities?
- If teachers select features for students, can selection procedures used for students with disabilities be generalized for use with students without disabilities?
- How can students be taught to decide whether to use scaffolding or to attempt an item/task without scaffolding? Can we detect overuse or unnecessary use of construct-related supports?

While all the research topics discussed above have a bearing on the validity of inferences from TE/UDAs, research focusing directly on validity and related technical characteristics is needed. Issues of validity, reliability, and fairness may be different in TE/UDAs that include universally available access features from those that arise from the use of accommodations in paper-and-pencil testing, and the use of scaffolding or branching items adds another dimension to these issues. Research questions are as follows:

- What is a framework that can be used to support the validity of inferences from access-based TEAs? What specific components should be included in that framework and what questions should be answered? How can that framework be applied in test design, development, implementation, and interpretation?
- What research designs other than the "differential effect" model for evaluating accommodations can be developed for investigating access-based TEAs? These designs should be sensitive to the effects of ancillary requirements on students, whether they have a disability or not.
- How do universally accessible tools affect the constructs assessed? Are the constructs comparable for students who do and do not use the features? Is this equivalence mediated by disability status?
- To what extent do access-based TEAs provide opportunity to perform for students with disabilities, facilitating access as intended?

- What are the effects of specific access features, representation options, and scaffolding approaches on student test scores? Do these effects differ as a function of student or group characteristics, or disability status?

Conclusions

At this writing, states are being encouraged by the U.S. Department of Education to develop and implement innovative assessments through various competitive grant programs, including the Race to the Top competition, Enhanced Assessment Grants, and the Investing in Innovation Fund Grants. The administration, in its *Blueprint for Reform* (U.S. Department of Education, March 2010), calls for "new assessment systems [that] will better capture higher-order skills, provide more accurate measures of student growth, and better inform classroom instruction to respond to academic needs" (p. 4). The goals for reauthorization of the Elementary and Secondary Education Act (U.S. Department of Education, March 2010) continue to emphasize the importance of educating all students well and accounting for all students' learning. The possibility of improving what we assess and how we assess it with technology is real.

As evidenced by the discussion of the state of the art in technology-enabled assessment and the issues that should be addressed as we take advantage of technology in our assessment programs, a principled program of research is needed to properly develop and use technology-enabled universally designed assessments. As the research progresses, we can incorporate what we have learned to build TE/UDAs, with the understanding that they will continue to improve in how they assess students and how they provide access to students, particularly students with disabilities.

The program of research will need to address the four major themes in this paper—the use of a construct-centered approach in designing accessible tests, the incorporation of access features into the test delivery system, the use of scaffolding in designing items and tests, and the need for a well-structured validation framework—as it examines increasing access to assessments that are likely to measure learning in ways that are not possible through paper-and-pencil tests.[7] Research designs will need to be carefully thought out first, through the generation of clear and targeted research questions, so that the effects of providing access features and the effects of changing what is being measured can be estimated. At some point, it may be that access features are so thoroughly understood and commonly used that building accessible assessments is second nature; until then, there will be a need to show whether the tools corrupt or clarify the targeted measure.

As part of this carefully considered effort, researchers will need to consider alternatives to the prevalent paradigm in accommodations research, one that tests the "interaction hypothesis": if the accommodation improves the scores for students with disabilities more than it improves scores for students without disabilities, then there is support that the accommodation is a "valid" one. The interaction hypothesis as a way to study the effects of accommodations has been losing support in the field.[8] One reason for this is that there can be a number of interpretations of the same result from studies relying on demonstrating an interaction effect. For example, it could be that the accommodation under study appropriately removes a construct-irrelevant access barrier for some students with disabilities and removes the same barrier for some students without disabilities. The results might show that both groups, on average, increased their scores equally on the accommodated version of the test, with the inference, according to the interaction hypothesis, that the accommodation is not valid – for an accommodation to produce valid scores, the interaction hypothesis requires the accommodation to increase the scores of students with disabilities more than the scores of students without disabilities. In this case, however, the accommodation is providing for better (more valid) scores for students in both groups, rather than reducing the validity of scores for students with disabilities (by providing an unfair advantage).

The interaction hypothesis may not adequately consider the interaction between student characteristics and the features of the test itself. Discussions of the four themes in this white paper call for researchers to tease out the relative importance of the interaction between the needs of students with disabilities and the cognitive demands of the test and item features within the context of TE/UDA approaches. When individual test takers encounter test items, the research needs to consider under what conditions target skills are properly conveyed in this interaction and, in particular, when communication about targeted information becomes systematically contaminated, misunderstood, or distorted. Research needs to examine error, which occurs in regular and predictable ways when individuals with specific characteristics interact with specific task factors. The error, which influences task performance but is not part of what one intends to measure, can influence results and interfere with accurately measuring targeted knowledge and skills and in interpreting scores. Based on the major themes in this paper, it is possible to envision circumstances in which the following occurs:

- Test results do not reflect student knowledge, skills, and abilities but rather ancillary KSAs that interfere with measuring targeted KSAs.
- Access features embedded into a TE/UDA testing platform are available to a student with disabilities who needs them but are not used by the student.

- Scaffolding alters the targeted construct and therefore the inferences that can be made.

Research comparing scores of students with and without disabilities designed around the interaction hypothesis may not detect subtleties of the student/test interaction. For example, there may be no interaction effect found in the following circumstances in which student scores are compared with and without a read-aloud accommodation:

- Some students with disabilities do not need the accommodation to access the test so their scores do not improve with the read-aloud provision.
- Some students without disabilities who are poor readers benefit from the accommodation so their scores improve with the read-aloud provision.

In this case, the read-aloud accommodation appropriately affected student performance, but it is not apparent from the mean scores of the groups. In other cases, there may be appear to be an interaction but it does not reflect a real effect, as in the following example:

- Students with disabilities use the accommodation and their test scores improve.
- Students without disabilities pay no attention to the accommodation provided and their test scores do not change.

In this case, mean scores show that the accommodation affected scores of students with disabilities and did not affect scores of students without disabilities, but these means differ because of poor implementation of the accommodation (students without disabilities did not use it when provided).

The National Research Council report (2004) puts it this way: "For the most part, existing research has investigated the effects of accommodations on test performance but is not informative about the *validity of inferences* [emphasis added] based on scores from accommodated administrations" (p. 101). TE/UDAs have potential to present students with opportunities to fully demonstrate what they know and can do, so that results of their performance are valid and can be used to effectively guide instruction that supports students' development of deeper levels of knowledge and understanding, and greater complexity of skills. We can no longer delay the systematic examination of tools such as TE/UDAs that hold great promise to support our efforts to fully include students and effectively facilitate their learning and achievement.

Endnotes

1. Measured Progress was the testing vendor for Kentucky. More information is available at http://www.measuredprogress.org/.
2. Kurzweil (http://www.kurzweiledu.com/) offers a comprehensive reading, writing and learning software solution for any struggling reader, including individuals with learning difficulties commonly used in K–12 schools.
3. The authors recognize that "scaffolds" and "scaffolding" are used in the cognitive psychology and instructional literature to refer to a variety of concepts and techniques, sometimes narrowly defined and sometimes broadly. We use the term in discussing assessment because it has already been applied in an assessment context (e.g., Hall, Strangman, & Meyer, 2009), and we hope to link to the current applications of what is being called scaffolding in testing situations.
4. In current state assessment terminology, such changes are referred to as "modifications" to the test.
5. We recognize that this is a gross oversimplification of how CAT models actually work.
6. Wiliam and Thompson (2007, cited in Wiliam, 2007) typify short-cycle formative assessments as those used within and between lessons and medium-cycle formative assessments as those used within and between units.
7. See the white paper by Bechard, Sheinker, Abell, Barton, Blackorby, Burling, Camacho, Cameto, Haertel, Hansen, Johnstone, Kingston, Murray, Parker, Redfield, Rodriquez, & Tucker (2010) for a full discussion about how what is being measured is changing, including a discussion of learning progressions.
8. For detailed information, see Sireci, Li, and Scarpati (2003) (specifically p. 48 and pp. 60-63) and the National Research Council (2004) (specifically pp. 87–88, p. 96, and p. 101).

References

Abedi, J., Courtney, M., & Leon, S. (2003). *Research-supported accommodation for English language learners in NAEP.* Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles.

Abedi, J., Courtney, M., Mirocha, J., Leon, S., & Goldberg, J. (2005). *Language Accommodations for English Language Learners in Large-Scale Assessments: Bilingual Dictionaries and Linguistic Modification—CSE Report 666*. National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles.

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

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

Bechard, S., Sheinker, J., Abell, R., Barton, K., Blackorby, J., Burling, K., Camacho, C., Cameto, R., Haertel, G., Hansen, E., Johnstone, C., Kingston, N., Murray, E., Parker, C., Redfield, D., Rodriquez, J., and Tucker, B. (2010). *Measuring Cognition of Students with Disabilities Using Technology-Enabled Assessments: Recommendations for a Research Agenda*. Dover, NH: Measured Progress and Menlo Park, CA: SRI International.

Beddow, P.A. (in press). Beyond universal design: Accessibility theory to advance testing for all students. In M. Russell (Ed.), *Assessing students in the margins: Challenges, strategies, and techniques*. In M. Russell (Ed.), Assessing Students in the Margins. Information Age Publishing, Charlotte, NC.

Beddow, P.A., Elliott, S.N., & Kettler, R.J. (2009). *TAMI Accessibility Rating Matrix*. Nashville, TN: Vanderbilt University. Available at http://peabody.vanderbilt.edu/tami.xml.

Bennett, R.E. (1999). Computer-based testing for examinees with disabilities: On the road to generalized accommodation. In S.J. Messick (Ed.), *Assessment in higher education: Issues of access, quality, student development, and public policy*. (pp. 181–191). Mahwah, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

Bielinski, J., Sheinker, A., and Ysseldyke, J. (2003, April). *Varied opinions on how to report accommodated test scores* (NCEO Synthesis Report 49). Minneapolis: National Center on Educational Outcomes.

Bruner, J. (1983). *Child's talk*. New York, NY: Norton.

Burk, M. (1999). *Computerized test accommodations: A new approach for inclusion and success for students with disabilities*. Washington, DC: A.U. Software, Inc.

Camacho, C. (2009, July). *Scaffolded assessments*. Presentation at the Invitational Research Symposium on Universally Designed Technology Enabled Assessments, Arlington, VA.

Cameto, R., Haertel, G., DeBarger, A., & Morrison, K. (2010). *Applying Evidence-Centered Design to Alternate Assessments in Mathematics for Students with Significant Cognitive Disabilities (Alternate Assessment Design-Mathematics Technical Report #1: Project Overview)*. Menlo Park, CA: SRI International.

Children's Progress, Inc. (2009). *Children's progress: Insights for educators. Success for students* (1999-2010) (Children's Progress Academic Assessment [CPAA] Technical Report). Retrieved from http://www.childrensprogress.com/resources/research.shtml, August 24, 2009.

Chin, C., & Teou, L. (2009). Using concept cartoons in formative assessment: Scaffolding students' argumentation. International *Journal of Science Education, 31*, 1307–1332.

Chudowsky, N., & Pellegrino, J. (2003). Large-scale assessments that support learning: What will it take? *Theory into Practice, 42*, 75–83.

Clarke, J. (2009). *Immersive interfaces for assessing student learning*. Presentation at the Invitational Research Symposium on Universally Designed Technology Enabled Assessments, July 23, 2009, Arlington, VA.

DeBarger, A., Haertel, G., Villalba, S., Hamel, L., & Colker, A. (2009). *Integration of evidence-centered design and universal design principles using PADI, an online assessment design system*. A paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, CA.

Dolan, R.P., Burling, K.S., Harms, M., Beck, R., Hanna, E., Jude, J., et al. (2006). *Universal design for computer-based testing guidelines*. Retrieved May 4, 2009, from http://www.pearsonedmeasurement.com/cast/index.html.

Dolan, R.P., Rose, D.H., Burling, K.S., Harms, M., & Way, W. (2007, April). *The universal design for computer-based testing framework: A structure for developing guidelines for constructing innovative computer-administered tests*. Paper presented at the National Council on Measurement in Education Annual Meeting, Chicago, IL.

Dolan, R.P., Hall, T.E., Banerjee, M., Chun, E., & Strangman, N. (2005). Applying principles of universal design to test delivery: The effect of computer-based read-aloud on test performance of high school students with learning disabilities. *Journal of Technology, Learning & Assessment, 3*(7), 1–32.

Dolan, R.P., & Hall, T.E. (2001). Universal Design for Learning: Implications for large-scale assessment. *IDA Perspectives, 27*(4), 22–25.

Dolan, R.P., & Hall, T.E. (2007). Developing accessible tests with universal design and digital technologies: Ensuring we standardize the right things. In L. L. Cook & C. C. Cahalan (Eds.), *Large-scale assessment and accommodations: What works* (pp. 95–111). Arlington, VA: Council for Exception Children.

Dolan, R.P., Hall, T.E., Banerjee, M., Chun, E., & Strangman, N. (2005). Applying principles of universal design to test delivery: The effect of computer-based read-aloud on test performance of high school students with learning disabilities. *Journal of Technology, Learning, and Assessment, 3*(7). Available from http://www.jtla.org.

Dolan, R.P., & Rose, D.H. (2000). Accurate assessment through Universal Design for Learning. *Journal of Special Education Technology, 15*(4).

Dolan, R.P., Rose, D.H., Burling, K.S., Harms, M., & Way, W. (2007, April). *The universal design for computer-based testing framework: A structure for developing guidelines for constructing innovative computer-administered tests*. Paper presented at the National Council on Measurement in Education Annual Meeting, Chicago, IL.

Elliott, J.L. & Thurlow, M.L. (2006). *Improving Test Performance of Students with Disabilities…on District and State Assessments* (2nd ed.).Thousand Oaks CA: Corwin Press.

Elliott, S.N., Kratochwill, T.R., McKevitt, B, Schulte, A.G., Marquart, A., & Mroch, A. (June, 1999). *Experimental analysis of the effects of testing accommodations on the scores of students with and without disabilities: Mid-project results*. A paper presented at the CCSSO Large-Scale Assessment Conference, Snowbird, Utah, June, 1999.Fuchs, Eaton, Hamlett, & Karns, 2000.

Feng, M., Heffernan, N., & Koedinger, K. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. *The Journal of User Modeling and User-Adapted Interaction, 19*, 243–266.

Flowers, C., Wakeman, S., Browder, D. & Karvonen, M. (2007). *Links for academic learning: An alignment protocol for alternate assessments based on alternate achievement standards*. Charlotte, North Carolina: University of North Carolina at Charlotte.

Fuchs, L.S., Fuchs, D., Eaton, S. B., Hamlett, C.L., & Karns, K.M. (2000). Supplementing teacher judgment of mathematics test accommodations with objective data sources. *School Psychology Review, 29*(1), 65–85.

Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The effect of computer-based tests on racial/ethnic and gender groups. *Journal of Educational Measurement, 39*(2), 133–147.

Hall, T., Strangman, N., & Meyer, A. (November, 2009). *Differentiated instruction and implications for UDL implementation*. Retrieved February 17, 2010, from http://www.cast.org/publications/ncac/ncac_diffinstructudl.html.

Hansen, E.G. & Mislevy, R.J. (2008). *Design Patterns for Improving Accessibility for Test Takers With Disabilities*. Princeton, NJ, ETS Research Report No. RR-08-49.

Helwig, R., Rozek-Tedesco, M.A., & Tindal, G. (2002). An oral versus a standard administration of a large-scale math test. *Journal of Special Education, 36*(1), 39–47.

Helwig, R., Rozek-Tedesco, M.A., Heath, B., & Tindal, G. (1999). Reading as an access to math problem solving on multiple choice tests. *Journal of Educational Research, 93*, 113–125.

Madaus, G., Russell, M., & Higgins, J. (2009). *The Paradoxes of High-Stakes Testing: How The Affect Our Children, Their Parents, Teachers, Principals, Schools, and Society*. Charlotte, NC: Information Age Publishing.

Hollenbeck, K. (2002). Determining when test alterations are valid accommodations or modifications for large-scale assessment. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 395–427). Mahwah, NJ: Erlbaum.

Individuals with Disabilities Act. (1997). Public Law 105-117. Washington, DC: U. S. Government Printing Office.

Individuals with Disabilities Act. (2004). Public Law 108-446. Washington, DC: U. S. Government Printing Office.

International Association for the Evaluation of Educational Achievement. *TIMSS Mathematics Items for the Middle School Years: Released Set for Population 2 (Seventh and Eighth Grade)* (1995). TIMSS International Study Center, Boston College, Chestnut Hill, MA. Downloaded from http://timss.bc.edu/timss1995i/TIMSSPDF/BMItems.pdf on April 12 2010.

Johnstone, C.J., Bottsford-Miller, N.A., & Thompson, S.J. (2006). *Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners* (Technical Report 44). Minneapolis MN: University of Minnesota, National Center on Educational Outcomes.

Kane, M.T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (Volume 4, pp. 17–64). Westport, CT: American Council on Education and Praeger.

Ketterlin-Geller, L.R. (2005). Knowing what all students know: Procedures for developing universal design for assessment. *Journal of Technology, Learning, and Assessment, 4*(2), 1–23.

Ketterlin-Geller, L.R. (2008). Testing students with special needs: A model for understanding the interaction between assessment and student characteristics in a universally designed environment. *Educational Measurement: Issues and Practice, 27*, 3–16.

Ketterlin-Geller, L.R., Alonzo, J., Braun-Monegan, J., & Tindal, G. (2007). Recommendations for accommodations: Implications of (in) consistency. *Remedial and Special Education, 28*(4), 194–206.

Ketterlin-Geller, L.R., Yovanoff, P., & Tindal, G. (2007). Developing a new paradigm for accommodations research. *Exceptional Children, 73*(3), 331–347.

Kettler, R.J., & Elliott, S.N. (in press). Assessment accommodations for children with special needs. In B. McGaw, E. Baker, & P. Peterson (Eds.), *International encyclopedia of education: 3rd edition*. Oxford, UK: Elsevier.

Kettler, R.J., Elliott, S.N., & Beddow, P.A. (2009). Alternate assessments based on modified achievement standards: Introduction to the federal policy and related implementation issues. *Peabody Journal of Education, 84*, 529–551.

Kopriva, R., Samuelson, K., Wiley, D., & Winter, P. (2003, April). *Evidentiary logic in the assessment of diverse learners*. Paper presented at the annual conference of the National Council on Measurement in Education, Chicago, IL.

Kopriva, R., Wiley, D., & Winter, P. (2004, April). *Rethinking the role of individual differences in educational assessment*. Paper presented at the Annual Meeting of the National Council of Measurement in Education, San Diego, CA.

Lazarus, S.S., & Thurlow, M.L. (2009). The changing landscape of alternate assessments based on modified academic achievement standards (AA-MAS): An analysis of early adopters of AA-MASs. *Peabody Journal of Education, 84*(4).

Lazarus, S.S., Thurlow, M.L., Lail, K.E., & Christensen, L. (2009). A longitudinal analysis of state accommodations policies: Twelve years of change 1993-2005. *Journal of Special Education, 43*(2), 67–80.

Mace, R.L. (1991). *Definitions: Accessible, adaptable, and universal design* (Fact Sheet). Raleigh, NC: Center for Universal Design, NCSU.

Mace, R.L., Hardie, G.J., & Place, J.P. (1996). *Accessible environments: Toward universal design*. Raleigh, NC: Center for Universal Design.

Madaus, G., Russell, M., & Higgins, J. (2009). *The paradoxes of high-stakes testing: How they affect students, their parents, teachers, principals, schools, and society*. Charlotte, NC: Information Age Publishing.

Marion, S.F., & Pellegrino, J.W. (Winter 2006). A validity framework for evaluating the technical quality of alternate assessments. *Educational Measurement: Issues and Practice, 25*(4), 47–57.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103.). Washington, DC: American Council on Education.

Messick, S. (1994). *Alternative modes of assessment, uniform standards of validity* (No. ETS RR-94-60). Princeton, NJ: Educational Testing Service.

Messick, S. (1996). “Validity in performance assessments”, in G.W. Phillips (Eds), National Center for Education Statistics (NCES 96–802), Washington DC, pp. 1–18.

Mislevy, R.J., Almond, R.G., & Lukas, J.F. (2003). *A brief introduction to evidence-centered design*. Retrieved October 23, 2006, from http://www.education.umd.edu/EDMS/mislevy/papers/BriefIntroECD.pdf.

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

Mislevy, R.J., & Riconscente, M.M. (2006). Evidence centered assessment design. In Downing, S.M. & Haladyna, T.H. (Eds.), *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003). On the structure of educational assessments. Measurement: Interdisciplinary *Research and Perspectives, 1*(1), 3–62.

Mislevy, R.J., Wilson, M. R., Erickan, E., & Chudowsky, N. (September, 2001). Psychometric principles in student assessment. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International Handbook of Educational Evaluation*. Dordrecht, the Netherlands: Kluwer Academic Press.

Murray, E., Gravel, J., Rose, D., Haertel, G., DeBarger, A., Cheng, B., Blackorby, J., Javitz, H., Colker, A., Rucctinger, L., Snow, E., Mislevy, R., Zhang, T., & Hansen. E. (2010). *Designing large-scale assessment tasks for students with high-incidence disabilities: The use of evidence-centered design, design patterns, and universal design for learning*. Paper presented at the Annual Meeting of the American Educational Research Association, Denver, Colorado.

National Research Council. (2004). *Keeping Score for All: The Effects of Inclusion and Accommodation Policies on Large-Scale Educational Assessments*. Committee on Participation of English Language Learners and Students with Disabilities in NAEP and Other Large-Scale Assessments. Judith A. Koenig and Lyle F. Bachman, Editors. Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council (2001). Knowing what students know: The science and design of educational assessment. Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds). Washington, DC: National Academies Press.

Quintana, C., Krajcik, J., & Soloway, E. (2002). *Scaffolding design guidelines for learner-centered software environments*. Paper presented at the 2002 annual meeting of the American Educational Research Association, New Orleans, LA.

Quintana, C., Reiser, B.J., Davis, E.A., Krajcik, J., Fretz, E., Duncan, R. G., Kyza, E., Edelson, D., & Soloway, E. (2004). A scaffolding design framework for software to support scientific inquiry. *Journal of the Learning Sciences, 13*, 337–386.

Rabinowitz, S.N., & Brandt, T. (2001). *Computer-based assessment: Can it deliver on its promise?* San Francisco, CA: WestEd.

Rabinowitz, S., & Sato, E. (in press). *Expanding notions of validity: Incremental validity and assessments for special student populations* [working title]. San Francisco: WestEd.

Rabinowitz, S., & Sato, E. (2005). *A technical review of high-stakes assessments for English language learners*. San Francisco: WestEd.

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

Rose, D.H., & Meyer, A. (2002). *Teaching every student in the digital age: Universal design for learning*. Alexandria, VA: ASCD Press.

Rose, D. & Meyer, A. (2000). Universal Design for Learning. Journal of *Special Education Technology, 15*(1), 67–70.

Rothman, R., Slattery, J.B., Vranek, J.L., & Resnick, L.B. (2002). *Benchmarking and alignment of standards and testing*. National Center for Research on Evaluation, Standards, and Student Testing. Retrieved June 2, 2005, from http://www.cresst.org/.

Russell, M., Goldberg, A., & O'Conner, K. (2003). Computer-based testing and validity: A look back and into the future. *Assessment in Education, 10*(3), p. 279–293.

Russell, M., Hoffmann, T., & Higgins, J. (2009). NimbleTools: A universally designed test delivery system. *Teaching Exceptional Children, 42*(2), p6–12.

Sato, E., Moughamian, A., Lagunoff, R., Rayyes, N., Rivera, M., & Francis, D. (in press). *Access and opportunity to learn for English language learner students and students with disabilities*. San Francisco: WestEd.

Sato, E., Rabinowitz, S., Gallagher, C., & Huang, C.W. (2010). *Accommodations for English language learners*. San Francisco, CA: WestEd.Shepard, L. A. (2005). Linking formative assessment to scaffolding. Educational Leadership, November 2005, 66–70.

Shepard, L.A. (2005). Linking formative assessment to scaffolding. *Educational Leadership, 63*(3) November 2005, 66–70.

Shute, V., & Zapata-Rivera, D. (2007). *Adaptive technologies* (Educational Testing Service Research Report RR-07-05). Princeton, NJ: Educational Testing Service.

Sireci, S.G., Li, S., & Scarpati, S. (2003). *The effects of test accommodation on test performance: A review of the literature* (Research Rep. No. 485). Amherst, MA: Center for Educational Assessment.

Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: an analysis of the interaction hypothesis. *Review of Educational Research, 75*(4), 457–490.

SMARTER Balanced Assessment Consortum (2010). *Executive Summary*. Retrieved from http://www.k12.wa.us/SMARTER/pubdocs/Exec_Smarter.pdf on July 11, 2010.

Solano-Flores, G. & Li, M. (2006). The use of generalizability (G) theory in the testing of linguistic minorities. *Educational Measurement: Issues and Practice, 25*(1), 13-22.

Stansfield, C.W. (2003). Test translation and adaptation in public education in the USA. *Language Testing*, 20: 189-207.

Thissen, D., Wainer, H., & Wang, X. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests? *Journal of Educational Measurement, 31*, 2.

Thompson, S.J., Johnstone, C.J. & Thurlow, M.L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S., Blount, A., & Thurlow, M. (2002). *A summary of research on the effects of test accommodations: 1999 through 2001* (Technical Report 34). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S.L., Thurlow, M.L., Quenemoen, R., & Lehr, C.A. (2002). *Access to computer-based testing for students with disabilities* (Synthesis Report 45). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M.L., Lazarus, S.S., & Christensen, L.L., (in press). Accommodations and modifications for assessment. In Lloyd, J. and Landrum, T. (Ed.), *Effective Practices for Special Education*. Arlington VA: Council for Exceptional Children/Pearson.

Thurlow, M.L., Thompson, S.J., & Lazarus, S.S. (2006). Considerations for the administration of tests to special needs students: Accommodations, modifications, and more. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*. (653–673). Mahwah NJ: Lawrence Erlbaum, Inc.

Thurlow, M.L., & Wiener, D. (2000). *Non-approved accommodations: Recommendations for use and reporting* (Policy Directions No. 11). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Tindal, G. (1998). *Models for understanding the task comparability in accommodated testing*. Washington, DC: Council of Chief State School Officers. Retrieved from the National Center on Educational Outcomes website at http://www.cehd.umn.edu/nceo/OnlinePubs/Accomm/TaskComparability.htm, December 30, 2009.

Tindal, G., & Ketterlin-Geller, L. (2004). *Research on mathematics test accommodations relevant to NAEP testing*. Washington, DC: National Assessment Governing Board.

Tucker, B. (2009). *Beyond the bubble: Technology and the future of student assessment*. Washington, DC: Education Sector.

U.S. Department of Education (March 2010). *Blueprint for reform*. Washington, DC: Author.

U.S. Department of Education NCLB and the Individuals with Disabilities Education Act (1990, 1997, and 2004).

Vygotsky, L.S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wiliam, D. (June, 2007). *What does research say the benefits of formative assessment are?* (NCTM Research Briefs and Clips). Available at http://www.nctm.org/news/content.aspx?id=11466.

Wiliam, D., & Thompson, M. (2007). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Winter, P.C. (2010). Comparability and test variations. In P.C. Winter (Ed.), *Evaluating the comparability of results from educational achievement test variations*. Washington, DC: Council of Chief State School Officers.

Winter, P. C., Kopriva, R. J., Chen, C-S., & Emick, J. E. (2006). Exploring individual and item factors that affect assessment validity for diverse learners: Results from a large-scale cognitive lab. *Learning and Individual Differences 16*(4) (2006) 267–276. Available at http://www.sciencedirect.com/.

Wood, D., Bruner, J., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*, 89–100.

Woolf, B.P. (2008). *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Amsterdam: Morgan Kaufmann/Elsevier.

Zhang, T., Mislevy, R.J., Haertel, G., Javitz, H., Murray, E., & Gravel, J. (2010). *A design pattern for a spelling assessment for students with disabilities*. Technical Report #2. Menlo Park, CA: SRI International.

Zhang, T., Haertel, G., Javitz, H., Mislevy, R., Murray, E., & Wasson, J. (2009). *A design pattern for a spelling bee assessment for students with disabilities*. A paper presented at the annual conference of the American Psychological Association, Montreal, Canada.

Author Biographies

Patricia Almond is a Research Associate at Center for Advanced Technology in Education at the University of Oregon. She can be reached at pjalmond@uoregon.edu. Dr. Almond is currently engaged in a systematic program of research investigating the influence of electronic text supports on access to content and achievement in mathematics for students with print disabilities in grades 4–9. She is a member of a team of researchers at SRI International that is applying evidence-centered design to academic content, mathematics and English language arts, in the design and development of alternate assessments. Dr. Almond is a nationally recognized expert in the areas of assessment and students with disabilities (SWD) particularly related to assessment accommodations, alternate assessments for students with significant cognitive disabilities, and statewide assessment systems. Dr. Almond has worked with numerous states on projects involving SWD, testing accommodations, and/or alternate assessment, including work with the states of Washington, Alaska, and Wyoming on their design, implementation, and validation of alternate assessments.

Phoebe Winter is Vice President for Education Policy at Pacific Metrics Corporation and can be reached at pwinter@pacificmetrics.com. Dr. Winter is a nationally recognized expert in education policy and educational assessment research and development. She consults with states on issues related to assessment and accountability programs and has worked with the U.S. Department of Education to develop and review guidelines for technical aspects of state assessment programs. As the research director at the Center for the Study of Assessment Validity and Evaluation at the University of Maryland, Dr. Winter coordinated research on assessing English learners. Earlier in her career she was a project director with the Council of Chief State School Officers, and worked on assessment systems at the Virginia Department of Education and the South Carolina Department of Education. She has written a number of books and articles that focus on policy and technical issues in state assessment. Dr. Winter holds a Ph.D. from Columbia University in psychology, with an emphasis on measurement, evaluation, and applied statistics.

Renée Cameto is Principal Scientist at SRI International and can be reached at renee.cameto@sri.com. Dr. Cameto has worked on assessing outcomes for students with disabilities for 25 years. She leads two Enhanced Assessment Grants projects to develop banks of tasks in mathematics and reading/English language arts for students with significant cognitive disabilities (SWSCD). Using evidence-centered design principles, tasks for each construct are created that reflect a range of cognitive complexity and depth of knowledge. She co-directs the IES-funded National Longitudinal Transition Study-2 (NLTS2), a 10-year study following a nationally representative sample of students with disabilities, and lead the assessments of both NLTS2 and Special Education Elementary Longitudinal Study sample members. Dr. Cameto directed the IES-funded National Study on Alternate Assessments from 2005 to 2010, publishing the only comprehensive national review of alternate assessments based on alternate achievement standards (AA-AAS) and a report on the implementation of AA-AAS by teachers. She cosponsors, with Measured Progress, research symposia on assessment of students with disabilities.

Michael Russell is an Associate Professor in the Lynch School of Education at Boston College and can be reached at russelmh@bc.edu.

Edynn Sato is the Research and English Language Learner Assessment Director at WestEdand can be reached at esato@wested.org.

Jody Clarke-Midura is a Research Associate at Harvard Graduate School of Education and can be reached at jec294@mail.harvard.edu.

Chloe Torres is a Program Manager at Measured Progress and can be reached at torres.chloe@measuredprogress.org.

Geneva Haertel is the Director of Assessment Research and Design, Center for Technology in Learning at SRI International and can be reached at geneva.haertel@sri.com.

Robert P. Dolan is a Senior Research Scientist at Pearson and can be reached at bob.dolan@pearson.com.

Peter Beddow is an Assistant Professor at Vanderbilt University and can be reached at peterbeddow@gmail.com.

Sheryl Lazarus is a Research Associate at the National Center on Educational Outcomes at the University of Minnesota and can be reached at laza0019@umn.edu.

JTLA

The Journal of Technology, Learning, and Assessment

www.jtla.org

Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College