Smarter Balanced Assessment Consortium:
Annotated Bibliography

Developed by Measured Progress / ETS Collaborative
Item and Task Specifications
and Guidelines Project

April 16, 2012

Acosta, B., Rivera, C. & Shafer Willner, L. (2008). *Best practices in the accommodation of English language learners: A Delphi study.* Report prepared for the U.S. Department of Education LEP Partnership. Arlington, VA: The George Washington University Center for Equity and Excellence in Education.

Charged by the U.S. Department of Education, The George Washington University Center for Equity and Excellence in Education (GW-CEEE) developed a Guide for state education agencies (SEAs) to use to improve state assessment policies for accommodating English language learners (ELLs). As a foundation for the Guide, GW-CEEE designed two studies, the Descriptive Study and the Best Practices Study. For the Descriptive Study GW-CEEE reviewed state assessment policies and examined the number and types of accommodations specified for ELLs (Shafer Willner, Rivera, & Acosta, 2008). The Best Practices Study involved the application of a Delphi technique to obtain consensus from an expert panel about which accommodations identified in the Descriptive Study were ELL-responsive. Members of the panel, which included experts knowledgeable about research, policy and practice in the areas of assessment, psychometrics, language testing, second language acquisition, and instruction of ELLs, relied on professional judgment to vet a list of ELL-responsive accommodations and then mapped these accommodations to English language proficiency (ELP) levels and to selected student background variables.

Component 1	Component 2	Component 3	Component 4	Component 5
				

ACT. (2011). *Fairness report for the ACT tests*. Iowa City, Iowa: Author.

The purpose of this report is to describe the procedures ACT followed when preparing the multiple-choice ACT® test forms that were administered in 2010–2011 to help ensure that these tests are as fair as possible to all examinees who take them. It is ACT's goal to accurately assess what students can do with what they know in the content areas covered by ACT's testing programs. If we were to allow factors other than the academic skills and knowledge in those content areas to intrude, we would provide a less accurate picture of what students know and can do and would risk subjecting students to situations in which their performance might be adversely affected by language or contexts that are perceived to be unfair.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Adkin, A. (2004). *Advantages of Uncontracted Braille*, *See/Hear* Spring 2004. http://www.tsbvi.edu/seehear/spring04/uncontracted.htm.

This brief article addresses issues pertaining to the selection of contracted or uncontracted braille for various educational purposes.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Advocates for Children of New York. (2010). *Students with Interrupted Formal Education: A Challenge for the New York City Public Schools.* New York, NY: Advocates for Children of New York.

New York City has nearly 150,000 English Language Learners (ELLs). Recent data show that only 39.7% of ELLs in the class of 2009 graduated from high school in four years and 19% had dropped out before completing four years of high school. To increase overall ELL graduation rates, New York City's Department of Education (DOE) has acknowledged that it must address the needs of certain subpopulations of high-needs ELLs. These ELLs with high needs include students who have interrupted formal education, ELLs with special education needs, and long-term ELLs. This paper focuses on one of these subpopulations—students with interrupted formal education or SIFE. It outlines some of these students' unique needs and recommends ways to strengthen New York City's efforts to meet those needs.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Ainley, M., Hidi, S., & Berndorff, D. (2002). Interest, learning and the psychological processes that mediate their relationship. *Journal of Educational Psychology*, *94*(3), 545-561.

The authors investigated how individual and situational interest factors contribute to topic interest and text learning. Traditional self-report measures were combined with novel interactive computerized methods of recording cognitive and affective reactions to science and popular culture texts, monitoring their development in real time. Australian and Canadian students read four expository texts. Both individual interest variables and specific text titles influenced topic interest. Examination of processes predictive of text learning suggested a chain reaction where topic interest was related to affective response, affect to persistence, and persistence to learning.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Allman, C. (2009). *Making Tests Accessible for Students with Visual Impairments: A Guide for Test Publishers, Test Developers, and State Assessment Personnel.* (4th edition.) Louisville, KY: American Printing House for the Blind. Available from http://www.aph.org

This document was created by the American Printing House for the Blind as a guide for making tests accessible in tactile, large print, and audio formats, and is intended to be used by organizations involved in creating, adapting, and administering tests to students with visual impairments.

Component 1	Component 2	Component 3	Component 4	Component 5
				

American Educational Research Association, National Council on Measurement in Education, and American Psychological Association. (1999). *The Standards for Educational and Psychological Testing.* Washington, DC: AERA.

The standards outlined in this book have been developed to provide criteria for the evaluation of tests, testing practices, and the effects of test use. The "Standards" provides a frame of reference to ensure that relevant issues are addressed. The first part of the book, "Test Construction, Evaluation, and Documentation," contains standards for validity, reliability, test development, scaling, norming, test administration, reporting, and supporting documentation for tests. The second section

addresses fairness in testing, and the third section considers specific testing applications. The chapters are: (1) "Validity"; (2) "Reliability and Errors of Measurement"; (3) "Test Development and Revision"; (4) "Scales, Norms, and Score Comparability"; (5) "Test Administration, Scoring, and Reporting"; (6) "Supporting Documentation for Tests"; (7) "Fairness in Testing and Test Use"; (8) "The Rights and Responsibilities of Test Takers"; (9) "Testing Individuals of Diverse Linguistic Backgrounds"; (10) "Testing Individuals with Disabilities"; (11) "The Responsibilities of Test Users"; (12) "Psychological Testing and Assessment"; (13) "Educational Testing and Assessment"; (14) "Testing in Employment and Credentialing"; and (15) "Testing in Program Evaluation and Public Policy." A glossary and an index are included.

Component 1	Component 2	Component 3	Component 4	Component 5
				

American Psychological Association. (2009). *Publication manual of the American Psychological Association*. Washington, DC: Author.

The *Publication Manual of the American Psychological Association* is the style manual of choice for writers, editors, students, and educators in the social and behavioral sciences. It provides invaluable guidance on all aspects of the writing process, from the ethics of authorship to the word choice that best reduces bias in language. Well-known for its authoritative and easy-to-use reference and citation system, the *Publication Manual* also offers guidance on choosing the headings, tables, figures, and tone that will result in strong, simple, and elegant scientific communication.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Amick, N., et. al. (1997). Guidelines for Design of Tactile Graphics. http://www.aph.org/edresearch/guides.htm.

This is a brief list of general guidelines for tactile graphic creation.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Anstrom, K., Butler, F., DiCerbo, P., Katz, A., Millet, J., & Rivera, C. (2010). *A Review of the literature on Academic English: Implications for K–12 English Language Learners.* Arlington, VA: The George Washington University Center for Equity and Excellence in Education.

A Review of the Literature on Academic English: Implications for K-12 English Language Learners reviews current literature to determine what is known and not known about the nature of academic English, instructional practices used to teach it, teacher preparation and training to improve instructional practice, and policies that support academic English. The report also raises critical challenges for the field in defining academic English and suggests areas for further inquiry.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Attali, Y., Freedman, M., Harrison, M., & Obetz, S., Powers, D. (2008). *Automated scoring of short-answer open-ended GRE® Subject Test items* (RR-08-20). Princeton, NJ: Educational Testing Service.

This report describes the development, administration, and scoring of open-ended variants of GRE® Subject Test items in biology and psychology. These questions were administered in a Web-based experiment to registered examinees of the respective Subject Tests. The questions required a short answer of 1-3 sentences, and responses were automatically scored by natural language processing methods, using the *c-rater*™ scoring engine, immediately after participants submitted their responses. Participants received immediate feedback on the correctness of their answers, and an opportunity to revise their answers. Subsequent human scoring of the responses allowed an evaluation of the quality of automated scoring. This report focuses on the success of the automated scoring process. A separate report describes the feedback and revision results.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Ayala, C. C., Schultz, S. E., Shavelson, R. J., & Yin, Y. (2002). Reasoning dimensions underlying science achievement: The case of performance assessment. *Educational Assessment*, *8*(2), 101-122.

This study sought to explore to what extent three reasoning skills in science (basic, spatial, and quantitative) were unique to the National Education Longitudinal Study of 1988 (NELS:88) by measuring these same skills on other assessments, including performance tasks. The study found some evidence of convergence between PAs and NEL:88 with a small sample of 35. The performance assessments used in this study to test the convergence with other multiple choice tests proved to measure multiple reasoning skills beyond the scope of the study. The authors made note that these science PAs effectively measured reasoning and other knowledge dimensions such as procedural, declarative and schematic. It was noted that a benefit of performance assessments are intended to measure multiple standards. The authors concluded that additional research is needed in this area with larger samples.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bailey, A., & Kelly, K. (2010). *The Use and Validity of Home Language Surveys in State English Language Proficiency Assessment Systems: A Review and Issues Perspective* [White Paper]. Retrieved from http://www.ode.state.or.us/opportunities/grants/nclb/title_iii/white-paper-2010.pdf.

This paper is a project deliverable for the U.S. Department of Education funded Enhanced Assessment Grant Evaluating the Validity of English Language Proficiency Assessments (EVEA; CFDA 84.368) that was awarded to the Office of the Superintendent for Public Instruction of the State of Washington. The project involves five states: Idaho, Indiana, Montana, Oregon, and Washington. These states currently do not belong to an existing English language proficiency assessment (ELPA) consortium; rather they have each worked with commercial test developers to create state-wide ELPAs that are aligned with their state English language development/proficiency standards. The main project goal is for each state to create a validity argument for its ELPA system. Project outcomes include building individual State Interpretive Arguments, as well as a more general

Common Interpretive Argument; designing a set of studies and instruments to support and pilot test these arguments; and making instruments publicly available at the close of the project for the wider education community to access. This paper is focused on the different Home Language Surveys (HLS) used across states as a means of initially identifying those students who may be eligible for language services. It grew out of conversations that took place at EVEA project meetings in January 2010, when a number of the project states recognized that the role of an HLS in their ELPA systems necessitated its further scrutiny as part of the validation process. Their main concern was a lack of evidence for the validity of an HLS as an initial identifying instrument.

Component 1	Component 2	Component 3	Component 4	Component 5
				✓

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, 37(2), 122-147.

This seminal book covers the theory of self-efficacy, which is defined as a personal judgment of how well an individual can execute courses of action which are necessary to deal with prospective situations. The concept of self-efficacy resonates in the test taking literature and provides an avenue for explaining a range of test taking behaviors such as explanations for omit rates for certain types of item formats.

Component 1	Component 2	Component 3	Component 4	Component 5
✓	✓	✓		

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.

The Test Accessibility and Modification Inventory (TAMI; Beddow, Kettler, & Elliott, 2008) and TAMI Accessibility Rating Matrix (ARM; Beddow, Elliott, & Kettler, 2009) are a set of evaluation tools designed to facilitate a comprehensive analysis of tests and test items for the purpose of enhancing access and meaningful responses from all students.

The TAMI was developed as part of the Consortium for Alternate Assessment Assessment Validity and Experimental Studies (CAAVES). The rating component of the TAMI, the ARM, was developed as part of the Consortium for Modified Alternate Assessment Development and Implementation (CMAADI). The ARM consists of 2 rubrics: the Item Analysis rubric and the Overall Analysis rubric. The ARM provides a systematic method for evaluating and modifying test items with a focus on improving their accessibility for all students.

The TAMI ARM has been used to conduct accessibility reviews of items from several states' large-scale assessments to help ensure the tests yield scores from which inferences are equally valid for all test-takers.

By using the TAMI and TAMI ARM, new and existing tests and test items can be improved to enhance testing practices for students.

Component 1	Component 2	Component 3	Component 4	Component 5
				✓

Bejar, I. I. (1991). A methodology for scoring open-ended architectural design problems. *Journal of Applied Psychology, 76*, 522-532.

Psychometric and architectural principles were integrated to create a general approach for scoring open-ended architectural site-design test problems. In this approach, solutions are examined and described in terms of design features, and those features are then mapped onto a scoring scale by means of scoring rules. This methodology was applied to two problems that had been administered as part of a national certification test. Because the test is not currently administered by computer, the paper-and-pencil solutions were first converted to machine-readable form. One problem dealt with the spatial arrangement of buildings in a country club, and the other called for regrading of a site by rearranging contours. In both instances, the results suggest that computer scoring is feasible.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 199-218). Mahwah, NJ: Earlbaum.

Discusses briefly a definition and rationale for generative testing with an emphasis on the prerequisites to implement generative testing. It is contended that these prerequisites involve a thorough construct analysis that culminates in precise specifications capable of supporting generative testing. Then, the author describes an application that led to an operational licensing exam. This chapter concludes with some future prospects and outlines some needed work to advance a generative approach. (PsycINFO Database Record (c) 2010 APA, all rights reserved).

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bejar, I. I. (2010). *R&D Connections — can speech technology improve assessment and learning? New capabilities may facilitate assessment innovations (*RDC-15). Princeton, NJ: Educational Testing Service.

This article explores speech technology's potential to help address education challenges such as the development of literacy, especially reading proficiency, and the acquisition of communicative competence in English. This is the 15th edition in the R&D Connections series.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bejar, I. I., Bennett, R. E., Lawless, R. R., Morley, M. E., Revuelta, J. & Wagner, M. E. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *Journal of Technology, Learning, and Assessment, 2*(3). Available from http://www.jtla.org.

The goal of this study was to assess the feasibility of an approach to adaptive testing using item models based on the quantitative section of the Graduate Record Examination (GRE) test. An item model is a means of generating items that are isomorphic, that is, equivalent in content and equivalent psychometrically. Item models, like items, are calibrated by fitting an IRT response model. The resulting set of parameter estimates is imputed to all the items generated by the model. An on-the-fly adaptive test tailors the test to examinees and presents instances of an item model rather than independently developed items. A simulation study was designed to explore the effect an on-the-fly test design would have on score precision and bias as a function of the level of item model isomorphicity. In addition, two types of experimental tests were administered – an experimental, on-the-fly, adaptive quantitative-reasoning test as well as an experimental quantitative-reasoning linear test consisting of items based on item models. Results of the simulation study showed that under different levels of isomorphicity, there was no bias, but precision of measurement was eroded at some level. However, the comparison of experimental, on-the-fly adaptive test scores with the GRE test scores closely matched the test-retest correlation observed under operational conditions. Analyses of item functioning on the experimental linear test forms suggested that a high level of isomorphicity across items within models was achieved. The current study provides a promising first step toward significant cost reduction and theoretical improvement in test creation methodology for educational assessment.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bejar, I. I., & Braun, H. (1994). On the synergy between assessment and instruction: Early lessons from computer-based simulations. ***Machine-Mediated Learning, 4*****, 5-25.**

This article argues that synergy between computer-based instruction and automated assessment is possible because of the common needs in assessment and instruction, outlines a framework for characterizing performance, and examines procedures developed as part of an ongoing project to develop fully automated scoring of architectural design for a licensing exam.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bejar, I. I., & Graf, E.A. (2010). Updating the duplex design for test-based accountability in the twenty-first century. ***Measurement: Interdisciplinary Research & Perspective, 8*****(2), 110-129.**

The duplex design by Bock and Mislevy for school-based testing is revisited and evaluated as a potential platform in test-based accountability assessments today. We conclude that the model could be useful in meeting the many competing demands of today's test-based accountability assessments, although many research questions will need to be answered in future studies.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bejar, I. I., & Yocom, P. (1986). *A generative approach to the development of hidden figure items* (RR-86-20-ONR). Princeton, NJ: Educational Testing Service.

This report explores an approach to item development and psychometric modeling which explicitly incorporates knowledge about the mental models used by examinees in the solution of items into a psychometric model that characterize performances on a test, as well as incorporating that knowledge into the item development process. The paper focuses on the hidden-figure item type. Although there is extensive literature on the correlates of performance for this type of item, little is known about the mental models that may explain performance on the item. The approach taken in this paper is to search for a complexity dimension that accounts for the difficulty of hidden figures. Although several complexity dimensions can be postulated, we chose one inspired by artificial intelligence research on vision. A computer-based system was developed to analyze as well as generate items based on this framework. To empirically determine the validity of the chosen framework two experiments were conducted. The results suggest that this approach to psychometric modeling is viable. The practical and theoretical implications of the research are discussed. (48pp.)

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bennett, R. E. (1999). *Graphical modeling: A new response type for measuring the qualitative component of mathematical reasoning*. (RR-99-21). Princeton, NJ: Educational Testing Service.

This study investigated the psychometric functioning of Graphical Modeling (GM), a new computer-delivered response type for assessing mathematical reasoning that asks candidates to respond to a problem situation by creating a graphical representation. GM problems can be like the single-best-answer items currently found on the GRE® General Test, or they can be more loosely defined, allowing for multiple correct responses. Two GM tests differing from one another in the manipulation of specific item features were randomly spiraled among study participants. Analyses were performed relating to internal consistency reliability, relations with external criteria, features that contribute to item difficulty, adverse gender impact, and examinee perceptions. Results showed that GM scores were very reliable and moderately related to the General Test's quantitative section, suggesting that the introduction of GM items on the General Test might help broaden the GRE quantitative construct. In exploratory analyses of difficulty, one of three manipulated item features, problem structure, had a significant effect. Our impact analyses detected no significant gender differences independent of those associated with the GRE quantitative section. Finally, while more participants preferred regular multiple-choice graphical reasoning questions to GM items, more also thought GM was the fairer indicator of their ability to undertake graduate study.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bennett, R. E. (2002). *Inexorable and Inevitable: The continuing story of technology and assessment* (RM-02-03). Princeton, NJ: Educational Testing Service.

In this paper, the author argues that the inexorable advance of technology will force fundamental changes in the format and content of assessment. Technology is infusing the workplace, leading to widespread requirements for workers skilled in the use of computers. Technology is also finding a key place in education. This is occurring not only because technology skill has become a workplace

requirement. It is also happening because technology provides information resources central to the pursuit of knowledge and because the medium allows for the delivery of instruction to individuals who couldn't otherwise obtain it. As technology becomes more central to schooling, assessing students in a medium different from the one in which they typically learn will become increasingly untenable. Education leaders in several states and numerous school districts are acting on that implication, implementing technology-based tests for low- and high-stakes decisions in elementary and secondary school and across all key content areas. While some of these examinations are already being administered statewide, others will take several years to bring to fully operational status. These groundbreaking efforts will undoubtedly encounter significant difficulties that may include cost, measurement, technological-dependability, and security issues. But most importantly, state efforts will need to go beyond the initial achievement of computerizing traditional multiple-choice tests to create assessments that facilitate learning and instruction in ways that paper measures cannot.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. ***Measurement, 8,*** **70-91.**

CBAL (Cognitively Based Assessment of, for, and as Learning) is a research initiative intended to create a model for an innovative K-12 assessment system that documents what students have achieved ("of learning"); helps identify how to plan instruction ("for learning"); and is considered by students and teachers to be a worthwhile educational experience in and of itself ("as learning"). Because CBAL intends to not only measure student achievement but also facilitate it, CBAL, like any similar assessment program, requires a theory of action. This paper describes the notion of theory of action, offers a preliminary version of such a theory for CBAL, and outlines a provisional research program for evaluating that theory.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). ***Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP*** **(RM-08-11). Princeton, NJ: Educational Testing Service.**

This report describes selected results from the 2001 Math Online (MOL) study, 1 of 3 field investigations sponsored by the National Center for Education Statistics to explore the use of new technology for the National Assessment of Educational Progress (NAEP). Of particular interest in the MOL study was the comparability of scores from paper- and computer-based tests. A nationally representative sample of 8th-grade students was administered a computer-based mathematics test and a test of computer facility, among other measures. In addition, a randomly parallel group of students was administered a paper-based test containing the same math items as the computer-based test. Results showed that the computer-based mathematics test was significantly harder statistically than the paper-based test. In addition, computer facility predicted online mathematics test performance after controlling for performance on a paper-based mathematics test, suggesting that degree of familiarity with computers may matter when taking a computer-based mathematics test in NAEP.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bennett, R. E., Braun, H. I., Frye, D. & Soloway, E. (1989). *Developing and evaluating a machine-scorable, constrained constructed-response item* (RR-89-30). Princeton, NJ: Educational Testing Service.

The use of constructed response items in large scale standardized testing has been hampered by the costs and difficulties associated with obtaining reliable scores. The advent of expert systems may signal the eventual removal of this impediment. This study investigated the accuracy with which expert systems could score a new, non-multiple choice item type. The item type presents a faulty solution to a computer programming problem and asks the student to correct the solution. This item type was administered to a sample of high school seniors enrolled in an Advanced Placement course in Computer Science who also took the Advanced Placement Computer Science (APCS) Test. Results indicated that the expert systems were able to produce scores for between 82% and 97% of the solutions encountered and to display high agreement with a human reader on which solutions were and were not correct. Diagnoses of the specific errors produced by students were less accurate. Correlations with scores on the objective and free-response sections of the APCS examination were moderate. Implications for additional research and for testing practice are offered. (48pp.)

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K-12 assessment: Integrating accountability testing, formative assessment, and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43-61). New York, NY: Springer.

This paper presents a brief overview of the status of K-12 accountability testing in the United States. Following that review, we describe an assessment-system model designed to overcome the problems associated with current approaches to accountability testing. In particular, we propose a model in which accountability assessment, formative assessment, and professional support are built on the same conceptual base and work synergistically with one another. We close with a brief discussion of the role of technology and a review of the challenges that must be met if the highly ambitious system we suggest is to be realized.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bennett, R. E., Jenkins, F., Persky, H., & Weiss, A. (2003). *Assessing complex problem-solving performances* (RM-03-03). Princeton, NJ: Educational Testing Service.

Computer-based simulations can give a more nuanced understanding of what students know and can do than traditional testing methods. These extended, integrated tasks, however, introduce particular problems, including producing an overwhelming amount of data, multidimensionality, and local dependence. In this paper, we describe an approach to understanding the data from complex performances based on Evidence-centered Design, a methodology for devising assessments and for using the evidence observed in complex student performances to make inferences about proficiency.

We use as an illustration the NAEP Problem-Solving in Technology-Rich Environments Study, which is being conducted to exemplify how nontraditional skills might be assessed in a sample-based national survey. The paper focuses on the inferential uses of ECD, especially how features are extracted from student performance, how these extractions are evaluated, and how the evaluations are accumulated to make evaluative judgments.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bennett, R. E., Novatkoski, I., & Rock, D. A. (1989). Differential item functioning on the SAT-M Braille edition. *Journal of Educational Measurement, 26*(1), 67–79.

This study attempted to pinpoint the causes of differential item difficulty for blind students taking the braille edition of the Scholastic Aptitude Test's Mathematical section (SAT-M). The study method involved reviewing the literature to identify factors that might cause differential item functioning for these examinees, forming item categories based on these factors, identifying categories that functioned differentially, and assessing the functioning of the items comprising deviant categories to determine if the differential effect was pervasive. Results showed an association between selected item categories and differential functioning, particularly for items that included figures in the stimulus, items for which spatial estimation was helpful in eliminating at least two of the options, and items that presented figures that were small or medium in size. The precise meaning of this association was unclear, however, because some items from the suspected categories functioned normally, factors other than the hypothesized ones might have caused the observed aberrant item behavior, and the differential difficulty might reflect real population differences in relevant content knowledge.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bennett, R. E. & Persky, H. (2001). *Problem solving in technology-rich environments (*RM-01-02). Princeton, NJ: Educational Testing Service.

This paper describes the Problem Solving in Technology-Rich Environments (TRE) study. The TRE study will produce a set of example modules to assess problem solving with technology, and use these to address research questions related to employing technology in the National Assessment of Educational Progress (NAEP). The TRE modules are built around electronic information search and simulation (the latter of which is the focus of this report). Among other things, the modules are designed to incorporate incidental learning as a goal of good assessment, capture the multidimensional nature of problem solving in technology environments, take advantage of the unique capabilities of the computer, and disentangle component skills to describe student characteristics more meaningfully. In operational NAEP assessments, many such modules might be randomly spiraled among groups of students to provide evidence of problem solving with technology generally. Alternatively, a few such modules might be combined with a traditional subject-matter survey as a means of adding depth to the picture of what students know and can do.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bennett, R.E., Persky, H., Weiss, A.R., and Jenkins, F. (2007). *Problem solving in technology-rich environments: A report From the NAEP Technology-Based Assessment Projec*t (NCES 2007–466). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

The Problem Solving in Technology-Rich Environments (TRE) study was designed to demonstrate and explore innovative use of computers for developing, administering, scoring, and analyzing the results of National Assessment of Educational Progress (NAEP) assessments. Two scenarios (Search and Simulation) were created for measuring problem solving with technology and administered to nationally representative samples of students. Resultant data were used to describe the measurement characteristics of the scenarios and student performance. The Search scenario required students to locate and synthesize information from a simulated World Wide Web environment. The Simulation scenario required students to experiment to solve problems of increasing complexity. TRE Search consisted of 11 observables and produced a total score and two subscores: scientific inquiry and computer skills. Findings of the Search scenario include: (1) Internal consistency of the three TRE Search scores (total, scientific inquiry, and computer skills) ranged from 0.65 to 0.74; (2) Search scores provided overlapping but not redundant information; (3) Scientific inquiry skill scale score was most related to the relevance of the World Wide Web pages visited or bookmarked, the quality of the constructed response to a question designed to motivate students to search for and synthesize information from the Web, and the degree of use of relevant search terms; (4) The computer skills scale score was related primarily to the use of hyperlinks, the use of the Back button, the number of searches needed to get relevant hits, and the use of bookmarking; and (5) Statistically significant differences in performance were found on one or more TRE Search scales for NAEP reporting groups categorized by race/ethnicity, parents' highest education level, student eligibility for free or reduced-price school lunch, and school location, but not for reporting groups categorized by gender. The TRE Simulation scenario consisted of 28 observables and produced a total score and three subscores: scientific exploration, scientific synthesis, and computer skills. Findings of the Simulation scenario include: (1) Internal consistency of the four scales ranged from 0.73 to 0.89; (2) Simulation scores provided overlapping but not redundant information; (3) Scientific exploration skill scale score was most related to which experiments students chose to solve the Simulation problems; (4) Scientific synthesis scale was primarily related to the degree of correctness and completeness of conclusions drawn for each problem; (5) Performance on the computer skills scale was related mainly to the number of characters in the written responses students gave for each of the Simulation problems; and (6) Statistically significant differences in performance were found on one or more TRE Simulation scales for NAEP reporting groups categorized by race/ethnicity, parents' highest education level, and student eligibility for free or reduced-price school lunch, but not for reporting groups categorized by gender or school location. It is noted that this report presents results that do not reach definitive conclusions at this point in time and techniques and inferences made from the data may be subject to future revision. Appendixes include: (A) Development Committee for the Problem Solving in Technology-Rich Environments (TRE) Study; (B) Sample Selection; (C) Technical Specifications for Participating Schools; (D) Prior Knowledge and Background Questions for Search and Simulation Scenarios; (E) TRE Simulation Glossary, Help and Tutorial Screens; (F) Bayesian Estimation in the Problem Solving in Technology-Rich Environments Study; (G) C-rater Rules for Scoring Students' Search Queries; (H) TRE Search and Simulation Scale Scores and Percentiles by Student Reporting Groups for Scales on Which Statistically Significant Group Differences Were Observed; (I) Summary Statistics for Prior Knowledge Measures and Mean Scale Scores for Background-Question Response Options; (J) Performance on Problem Solving in Technology-Rich Environments (TRE)

Observables; and (K) Understanding NAEP Reporting Groups. (Contains 116 figures and 44 tables.) [This report was written in collaboration with: Douglas Forer, Bruce Kaplan, Michael Wagner, and Lou Mang. The NAEP Problem Solving in Technology-Rich Environments (TRE) study was part of the Technology-Based Assessment (TBA) project, a collaborative effort led by the National Center for Education Statistics (NCES) and the National Assessment Governing Board, and carried out by Educational Testing Service (ETS) and Westat. The Problem Solving in TRE study is the last of three field investigations in the NAEP Technology-Based Assessment Project, which explores the use of new technology in administering NAEP. For previous investigations in this series, see ED485780.].

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bennett, R. E. & Rock, D. A. (1993). *Generalizability, validity, and examinee perceptions of a computer-delivered formulating hypotheses test* (RR-93-46). Princeton, NJ: Educational Testing Service.

Formulating-Hypotheses (F-H) items present a situation and ask the examinee to generate as many explanations for it as possible. This study examined the generalizability, validity, and examinee perceptions of a computer-delivered version of the task. Eight F-H questions were administered to 192 graduate students. Half of the items restricted examinees to 7 words per explanation, and half allowed up to 15 words. Generalizability results showed high interrater agreement, with tests of between two and four items scored by one judge achieving coefficients in the .80s. As in studies of paper-and-pencil versions, validity analyses found that although F-H was highly reliable, it was only weakly related to GRE General Test Scores, differing from that test primarily in relating more strongly to a measure of ideational fluency. Versions of F-H based on different response limits tapped somewhat different abilities, with items employing the 15-word constraint appearing more useful for graduate assessment. These items added to conventional measures in explaining school performance and creative expression. Finally, although the overwhelming majority of examinees found the F-H interface easy to use, some experienced difficulty, suggesting the possibility that computer familiarity constitutes a source of irrelevant variance in F-H scores.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bennett, R. E., & Rock, D. A. (1997). *Examining the validity of a computer-based generating explanations test in an operational setting* (RR-97-18). Princeton, NJ: Educational Testing Service.

Generating explanations (GE) is a computer-delivered item type that presents a situation and asks the examinee to pose as many plausible reasons for it as possible. Previous research suggests that GE measures a divergent thinking ability largely independent of the convergent skills tapped by the GRE General Test. This study was conducted to determine if prior GE validity results generalized to the GRE candidate population, how population groups performed, what effects partial-credit modeling might have for validity, and what problems were associated with operational administration. Validity results showed that earlier findings were generally supported: GE was found to be reliable but only marginally related to the General Test and to make significant (but small) independent contributions to the explanation of relevant criteria. With respect to population groups, GE produced smaller gender and ethnic group differences than did the General Test and showed the

same relations to outside criteria across groups, suggesting it was measuring similar skills in each population. Attempts to model GE responses on a partial-credit IRT scale succeeded but produced no improvement in relations with external criteria over those obtained by summing raw item scores. Finally, interviews conducted with examinees to detect potential delivery problems suggested that the directions needed to be shortened.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bennett, R. E., & Sebrechts, M. M. (1994). ***The accuracy of automatic qualitative analyses of constructed-response solutions to algebra word problems*** **(RR-94-04). Princeton, NJ: Educational Testing Service.**

This study evaluated expert system diagnoses of examinees' solutions to complex constructed-response algebra word problems. Problems were presented to three samples, each of which had taken the GRE General Test. One sample took the problems in paper-and-pencil form and the other two on computer. Responses were then diagnostically analyzed by an expert system, GIDE, and by four ETS mathematics test developers using a fine-grained categorization of error types. Results were highly consistent across the samples. Human judges agreed among themselves almost perfectly in describing responses as right or wrong but concurred at much lower levels (37% to 64% agreement) in categorizing the specific bugs they detected in incorrect solutions. The expert system agreed highly with the judges' right/wrong decisions (95% to 97% concurrence) and somewhat less closely (71% to 74%) with the bug categorizations that the judges, themselves, agreed on. Seven principal causes of machine-rater disagreement were detected, most of which could be remedied by making adjustments to GIDE, modifying the test presentation interface to constrain the form of examinee solutions, and working with test developers to specify rules for automatically dealing with special cases. These results suggest that highly accurate diagnostic analysis through knowledge-based understanding of complex responses may be difficult to achieve at the fine-grained level used by GIDE. The accuracy of qualitative judgments might be increased by using a smaller set of more general diagnostic categories and by integrating information from other sources, including performance on diverse item types.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bernard, M., Mills, M., Frank, T., & McKown, J. (2001). Which Fonts Do Children Prefer to Read Online? 3(1) Software Usability Research Library Wichita State University.

Children today are reading large amounts of text on computer screens, either in the classroom or for leisure. This study investigates preferences for different types and sizes of fonts for reading online. By examining four types of fonts at 12- and 14-point sizes, this study determines the font combination that is perceived as most readable on computer screens and most preferred by children.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Betrancourt, M. (2005). The animation and interactivity principles in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (pp. 287–296). New York, NY: Cambridge University Press.

Computer animation has tremendous potential to provide visualizations of dynamic phenomena that involve change over time. However, the research reviewed in the article showed that learners did not systematically take advantage of animated graphics in terms of comprehension of the underlying causal or functional model. This chapter reviewed the literature about the interface and content features that affect the potential benefits of animation over static graphics.

In the last decade, with the rapid progression of computing capacities and the progress of graphic design technologies, multimedia learning environments have evolved from sequential static text and picture frames to increasing sophisticated visualizations. Two characteristics appear to be popular among instruction designers and practitioners: the use of animated graphics as soon as depiction of dynamic system is involved, and the capability for learners to interact with the instructional material. Animation can provide benefits when it is interactive and the system reacts to the learner's input. Also, when learners have interactive control over their interaction with the animation, they find the material more enjoyable and easier to understand. Due to the cognitive load of processing animations, animation should only be used when truly needed, such as when the phenomenon changes over time, making static representations unacceptable, and when learners are novices in the domain and need assistance in forming mental models.

Again, like the articles covered in Interaction Design 2, The use of animation in interaction design must be leveraged on as an needed basis. Just like symbols and signs, too many of these visual sign can cause cognitve load issues for the elarners. Therefore, from a scaffolding perspective, it would be great to leverage animation interactions for novice users or for when the interaction design warrents proper use of it (i.e.: to provide feedback via an avatar to the learner).

Component 1	Component 2	Component 3	Component 4	Component 5
	✓			

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. Assessment in Education: Principles, Policy, and Practice, 5(1), 7–74.

This review of literature on classroom assessment utilized approximately 250 studies pertaining to classroom assessment constructs. The purpose of this review was to identify research on formative assessment constructs and make recommendations to classrooms and policy makers. Formative functions, such as feedback, self-assessment, peer assessment, and clear learning goals were featured as strategies that significantly increase student achievement. The research studies reported on in this review show conclusively formative assessment increases student learning (by as much as a .7 effect size). A call for more research on the intricacies of formative assessment is recommended by the authors in their concluding remarks. Although this review centers on formative assessment, the authors reveal the importance of students being involved in the assessment process. This includes understanding the learning goal, reflecting on their current level of achievement in reference to the learning goal, and using feedback to improve; all of which can be useful in setting up and implementing formative functions of pre-performance task activities.

Component 1	Component 2	Component 3	Component 4	Component 5
		✓		

Bolt, S. E., & Ysseldyke, J. (2008). Accommodating students with disabilities in large-scale testing: A comparison of differential item functioning (DIF) identified across disability types. *Journal of Psychoeducational Assessment, 26*, 121-138.

Many students with disabilities are provided accommodations to enable their participation in statewide assessment programs; however, there is concern that accommodations may invalidate test results. For test administrations to be considered valid for all student groups, there must be comparable measurement across groups. This can ensure that decisions based on test results are made in a fair manner for all students. In this study, measurement comparability for two groups of accommodated students with disabilities (i.e., accommodated students with physical disabilities and accommodated students with mental disabilities) was examined using differential item functioning (DIF) analysis, in which item-level characteristics of the test for these groups were systematically compared with those for a reference group of nonaccommodated students without disabilities. A relatively large number of DIF items were identified for both accommodated disability groups, suggesting that more attention to the testing needs of students with disabilities is warranted. Suggestions for future research are provided.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bond, L. (1993). Comments on the O'Neill & McPeek paper. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 277-280). Hillsdale, NJ: Lawrence Erlbaum Associates.

This brief review discusses the difficulty in attributing causes for DIF in subpopulations. Notes that it is difficult to say why an item is behaving differently for one group than for another.

Component 1	Component 2	Component 3	Component 4	Component 5
				

BANA. (1997). *Braille Formats: Principles of Print to Braille Transcription*, 1997 (http://brl.org/formats).

The electronic version of *Braille Formats*, available from the American Printing House for the Blind, serves as a reference for braille transcribers on the applicable transcription and contraction rules.

Component 1	Component 2	Component 3	Component 4	Component 5
				

BANA (Braille Authority of North America). (2011). *Guidelines and Standards For Tactile Graphics*, Web Version, 2011. http://www.brailleauthority.org/tg/web-manual/index.html.

This document provides a comprehensive collection of tactile graphics recommendations, including criteria for including tactile graphics, as well as principles for design, editing, planning, and technical production details. A section on standardized tests is included.

Component 1	Component 2	Component 3	Component 4	Component 5

				

Breland, H. M. (1999). *Exploration of an automated editing task as a GRE writing measure* (RR-99-09). Princeton, NJ: Educational Testing Service.

Two editing tasks were developed and programmed for the computer to explore the possibility that such tasks might be useful as measures of writing skill. An informal data collection was then conducted with 52 prospective graduate students. These students completed the editing tasks with no time limit, as well as a writing experience questionnaire. Scores obtained on the two editing tasks were correlated with variables developed from the questionnaire. The total score for the two editing tasks correlated .52 with student self-assessments of their writing ability, .46 with grade-point average (GPA) based on courses requiring at least some writing, and .30 with writing accomplishments. The correlation with GPA, however, was only .14. The reliability of the total editing score was estimated at .84.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bridgeman, B., Cline, F., & Levin, J. (2008). *Effects of Calculator Availability on GRE Quantitative Questions,* (ETS Research Report Series RR-08–31). Princeton, NJ: Educational Testing Service.

In order to estimate the likely effects on item difficulty when a calculator becomes available on the quantitative section of the Graduate Record Examinations® (GRE®-Q), 168 items (in six 28-item forms) were administered either with or without access to an on-screen four-function calculator. The forms were administered as a special research section at the end of operational tests, with student volunteers randomly assigned to the calculator or no-calculator groups. Usable data were obtained from 13,159 participants. Test development specialists were asked to rate which items they thought would become easier with a calculator. In general, the specialists were successful in identifying the items with relatively large calculator effects, though even these effects were quite small. An increase of only about four points in the percent correct should suffice for the items identified as likely to show calculator effects with no adjustment needed for the majority of the items. Introduction of a calculator should have little or no effect on gender and ethnic differences.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bridgeman, B., Harvey, A., & Braswell, J. (1995). Effects of calculator use on scores on a test of mathematical reasoning. *Journal of Educational Measurement, 32*(4), 323–340.

Half of a sample of 11,457 college-bound juniors used a calculator on Scholastic Aptitude Test mathematics questions, while half did not. Both genders and three ethnic groups benefited about equally from calculator use. Students who routinely used calculators were relatively advantaged, but effects on individual test items varied.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bridgeman, B., Lennon, M. L. & Jackenthal, A. (2001). *The effects of screen size, screen resolution, and display rate on computer-based test performance* (RR-01-23). Princeton, NJ: Educational Testing Service.

Computer-based tests administered in established commercial testing centers typically have used monitors of uniform size running at a set resolution. Web-based delivery of tests promises to greatly expand access, but at the price of less standardization in equipment. The current study evaluated the effects of variations in screen size, screen resolution, and presentation delay on verbal and mathematics scores in a sample of 357 college-bound high school juniors. The students were randomly assigned to one of six experimental conditions-three screen display conditions crossed with two presentation rate conditions. The three display conditions were: a 17-inch \monitor set to a resolution of 1024 x 768, a 17-inch monitor set to a resolution of 640 x 480, and a 15-inch monitor set to a resolution of 640 x 480. Items were presented either with no delay or with a five-second delay between questions (to emulate a slow Internet connection). No significant effects on math scores were found. Verbal scores were higher, by about a quarter of a standard deviation (28 points on the SAT scale), with the high-resolution display.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education,* 16, 191–205.

Studied the effects of variations in screen size, resolution, and presentation delay on verbal and mathematics scores on a computerized test for 357 high school juniors. No significant differences were found for mathematics scores, but verbal scores were higher with the larger resolution display.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Bridgeman, B., & Rock, D. A. (1992). *Development and evaluation of computer-administered analytical questions for the GRE General test* (RR-92-49). Princeton, NJ: Educational Testing Service.

Three new computer-administered item types for the analytical scale of the Graduate Record Examination General Test were developed and evaluated. One item type was a free-response version of the current analytical reasoning item type. The second item type was a somewhat constrained free-response version of the pattern identification (or number series) item type in which the student had to state the rule that generated the series. The third item type used the computer to administer yes/no analysis of explanations questions with a limited branching strategy. The computer tests were administered at four ETS regional offices to a sample of students who had previously taken the GRE General Test. Scores from the regular GRE administration and the special computer administration were matched for a sample of 349 students. A number of test administration design issues were identified, including the need to provide adequate practice exercises, design of an interface comfortable for computer-literate students, and problems with item- level timing. The pattern identification items were too difficult (or the practice was inadequate), but the other items appeared

to function well. There was no evidence that the open-ended analytical reasoning items were measuring anything beyond what is measured by the current multiple-choice version of these items.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Burns, E. (1990). Multiple-choice computer diagnostic test designs. Educational Technology, 30(1), 49-53.

The discussion in this article focuses on factors that should be considered when designing multiple-choice diagnostic tests for use with microcomputer systems. The article describes characteristics of item content, examines ways to present feedback and results, and discusses database management functions. Finally, the benefits for students with learning problems and physically disabled students are emphasized.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Burstein, J. C., Kaplan, R. M., Rohen-Wolff, S., Zuckerman, D. I., & Lu, C. (1999). *A review of computer-based speech technology for the TOEFL® 2000 Test* (RM-99-05). Princeton, NJ: Educational Testing Service.

Computer-based speech technology, the capability of a computer system to accept and process spoken language, is considered a potentially super-enabling technology for computer users. Once a computer can adequately "understand" spoken language, the accessibility of computers increases by many orders of magnitude. As part of our on-going effort to examine enabling and important technologies, we have undertaken this study to review the state of the art in computer-based speech technology in the context of the Test of English as a Foreign Language™ (TOEFL®) testing program. Our goal in this study is to assess the readiness of various computer-based speech technologies for this testing program. This paper focuses on desktop applications for speech recognition and speech synthesis. This study investigated several commerically available speech-based technologies. The systems we evaluated are based on desktop computer technology. Due to the length of this study, this evaluation was conducted on a small, but representative sample of state-of-the-art desktop systems. Systems were chosen because they were among the top-ranked personal computer-based, speech technology applications. The systems evaluated were: Dragon Dictate by Dragon Systems, Inc ., Kurzweil VOICE for Windows by Kurzweil Applied Intelligence, Inc., and a text-to-speech synthesis system, DECTalk PC 4.2 by Digital Equipment Corporation.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Buzick, H. M., & Laitusis, C. C. (2010). *A summary of models and standards-based applications for grade-to-grade growth on statewide assessments and implications for students with disabilities,* (ETS Research Report Series RR-10–14). Princeton, NJ: Educational Testing Service.

Recently growth-based approaches to accountability have received considerable attention because they have the potential to reward schools and teachers for improving student performance over time

by measuring the progress of students at all levels of the performance spectrum (including those who have not yet reached proficiency on state accountability assessments). While the use of growth in accountability holds promise for students with disabilities, measuring changes over time in their academic performance is complex. This paper summarizes models and approaches that use individual student test scores from multiple years for 3 different purposes: determination of adequate yearly progress under the federal accountability system, research on individual growth trajectories, and evaluation of the contribution of teachers and schools to student learning. Practical challenges in measuring and modeling growth for students with disabilities are then discussed. Finally, 3 areas in need of research on the measurement of growth from large-scale annual accountability assessments are identified and described: testing accommodations, test difficulty, and understanding the longitudinal characteristics of the population of students with disabilities.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Buzick, H. M., & Laitusis, C. C. (2010). Using growth for accountability: Measurement challenges for students with disabilities and recommendations for research. *Educational Researcher, 39*, 537–544.

Growth-based approaches to federal accountability are receiving considerable attention because they have the potential to reward schools and teachers for improving student performance over time by measuring student progress at all levels of the performance spectrum, including progress by students who have not reached proficiency on state accountability assessments. The use of growth in accountability holds promise for students with disabilities, but measuring changes over time in academic performance with large-scale annual assessments is complex. The authors discuss practical challenges in measuring and modeling growth for students with disabilities. In addition, they identify and describe three areas in need of research on the measurement of growth: the impact of testing accommodations, the impact of test difficulty, and the longitudinal characteristics of the population of students with disabilities.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Buzick, H. M., & Laitusis, C. C. (2011). *An Overview of Applications That Use Student Academic Growth for K–12 Accountability and Implications for Students with Disabilities*. ETS Research Spotlight. Princeton, NJ: Educational Testing Service.

In their work, Buzick and Laitusis examined the challenges associated with measuring year-to-year growth for students with disabilities.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Buzick, H. M., & Stone, E. (2011). Recommendations for conducting differential item functioning analyses for students with disabilities based on previous DIF studies. ETS Research Report. Princeton, NJ.

The purpose of this study is to help ensure that strategies for differential item functioning (DIF) detection for students with disabilities are appropriate and lead to meaningful results. We surveyed existing DIF studies for students with disabilities and describe them in terms of study design, statistical approach, sample characteristics, and DIF results. Based on descriptive and graphical summaries of previous DIF studies, we make recommendations for future studies of DIF for students with disabilities. Differential Item Functioning (DIF) Studies Surveyed are appended.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Cahalan, C., Mandinach, E.B., & Camara, W. J. (2002). *Predictive Validity of SAT I: Reasoning Test for Test Takers with Learning Disabilities and Extended Time Accommodations*. (Research Report 2002–5). New York, NY: College Entrance Examination Board.

The predictive validity of the SAT I: Reasoning Test was examined for students who took the test with an extended time accommodation for a learning disability. The sample included college students with learning disabilities who took the SAT I between 1995 and 1998 with extended time accommodations. First year grade point average (FGPA) was used as a measure of student performance. Although positive, the adjusted correlation between FGPA and SAT scores was lower for test takers with a learning disability than has been shown in prior research on test takers without disabilities. In addition, the SAT scores obtained with an extended time accommodation appear to overpredict FGPA for male test takers with a learning disability and accurately predict FGPA for female test takers with a learning disability. When the same students were examined using both SAT I test scores and self-reported high school grade point average (HSGPA) to predict FGPA, the scores and grades of male test takers did not under- or overpredict while the scores of female test takers underpredicted FGPA. Due to the relatively small sample size, additional research is required to examine group differences (e.g., type of learning disability, severity of disability) and the impact of differential support received from college disability service offices during the first year of college.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Cahalan-Laitusis, C., King, T. C., Cline, F., & Bridgeman, B. (2006). *Observational Timing Study on the SAT Reasoning Test for Test Takers with Learning Disability and/or AD/HD.* (Research Report 2006–4). New York, NY: College Entrance Examination Board.

The purpose of this study is to provide information on the actual time used by students with disabilities on the new SAT®. This study observed students with learning disabilities (LD) and/or attention deficit/hyperactivity disorder (AD/HD) as they took the SAT items under strict time limits and recorded the amount of times taken for each item. The study is a replication of Study 2 in Bridgeman, Cahalan, and Cline (2003), which observed students without disabilities completing the same test times that are included in this study. Comparisons of the results from this study to the results of Bridgeman et al. are made and recommendations on appropriate extended-time limits for most students with disabilities are provided.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Camilli, G. (2006). Test fairness. In R.L. Brennan (Ed.), *Educational measurement* (pp. 221-256). Washington, DC: American Council on Education / Praeger.

This chapter provides a conception framework for understanding both test fairness and the assumptions and modes of inferences that underlie corresponding statistical analyses. Recent methodological developments are examined as weill as item sensitivity review, fairness in classroom assessment, and historical themes regarding fairness in college admissions. (p221)

Component 1	Component 2	Component 3	Component 4	Component 5
				

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.

Utah, Idaho, and Florida have formed a consortium with SRI International to improve their AA-AAS using ECD to design and develop assessment tasks that are linked to state extended content standards in mathematics. In this report, we describe

- Project goals and activities
- The development of assessments for accountability purposes for students with significant cognitive disabilities
- ECD and UDL frameworks and describe how they are applied through a co-design process
- Our plan to produce a series of technical reports, including procedural guidelines, design documents, and associated sample assessment tasks
- Our dissemination plan including the project website, www.alternateassessmentdesign.sri.com

Component 1	Component 2	Component 3	Component 4	Component 5
				

Candell, G. L., & Ercikan, K. (1992). *Assessing the reliability of the Maryland School Performance Assessment Program using generalizability theory.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Chaparro, B. S., Shaikh, A. D., Chaparroa, A. & Merkle, E. C. (2010). Comparing the Legibility of Six Clear Type Typefaces to Verdana and Times New Roman. *Information Design Journal 18*(1).

This study compares the on-screen legibility of six ClearType typefaces to that of two existing typefaces widely used for business documents, email, and websites. Participants were presented with individual letters, digits, and symbols from each typeface for brief durations and asked to verbally identify the character.

Percent correct identification for each character was calculated and graphical sunflower plots were used to highlight the characters misidentified. Results show that the legibility was higher for the ClearType typefaces Consolas and Cambria as well as the non-ClearType typeface Verdana than for Times New Roman, especially for digits and symbols.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Christensen, L. L., Lazarus, S. S., Crone, M., & Thurlow, M. L. (2008). *2007 state policies on assessment participation and accommodations for students with disabilities* (Synthesis Report 69). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

This document presents an update of a 2006 report from NCEO tracking and analyzing state policies on assessment participation and accommodations since 1992. The purpose of the current analysis is to update information on these policies that was last reported by NCEO in 2006 (based on 2005 data). In this analysis, policies from all 50 states, plus 8 of the unique states, were reviewed. Two unique states, the Bureau of Indian Education and the U.S. Virgin Islands, were not included in the analysis. The current analysis of states' 2007 participation and accommodation policies found that state policies on participation and accommodation continue to evolve, and that they have become more detailed and specific than in previous years. The study found that state policies focus more on accommodations that allow for valid scores. There is a greater differentiation among accommodations for different groups of students (students with Individualized Education Programs, students with 504 Plans, English language learners). All regular states and some unique states have Web sites where users can access their policies. The "read aloud questions" and "sign interpret questions" accommodations continue to be controversial. More states have policies that prohibit certain accommodations than they did in 2005. More states have guidelines for the use of accommodations requiring a third party/access assistant (scribe, reader, sign language interpreter). This document is a descriptive analysis of the written policies that states have for the participation of students with disabilities in assessments and the use of accommodations during their assessments. Appended are: (A) State Documents Used in Analysis of Participation and Accommodation Policies; and (B) Participation and Accommodation Guidelines by State.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Cizek, G. (1994). The Effect of Altering the Position of Options in a Multiple-Choice Examination. *Educational and Psychological Measurement, 54*(1), 8-20.

Performance of a common set of test items on an examination in which the order of options for one test form was experimentally manipulated. Results for 759 medical specialty board examinees find that reordering item options results in significant but unpredictable effects on item difficulty.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Clarke, J. (2009). Studying the Potential of Virtual Performance Assessments for Measuring Student Achievement in Science. Paper presented at the American Educational Research Association (AERA), San Diego: April 13-17.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Cleary, T. A. (1968). Test Bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement*, 5, 115-124.

Summary of research study focused potential bias in Scholastic Aptitude Test (SAT) based on race.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Cole, N. S. (1973). Bias in selection. *Journal of Educational Measurement.* 10, 237-255.

It is the purpose of this paper to identify the values and beliefs about fairness which are the bases for several definitions of bias and to provide actual procedures for the practitioner to follow to alleviate bias according to the definition he chooses.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Cole, N. S., & Zieky, M. J. (2001). The new faces of fairness. *Journal of Educational Measurement*. *38*(4), 369-382.

For most of the 20th century, measurement professionals paid little interest to item and test fairness. A confluence of events in the late 1960s and early 1970s led to an intense interest in fairness issues among measurement professionals. In spite of more than 30 years of effort, there is still no generally accepted definition of fairness with respect to testing and no measure that can prove or disprove the fairness of a test. To advance the fairness of tests, measurement professionals must pay more attention to reducing group differences at the design stage of test development, to providing all examinees an opportunity to demonstrate their knowledge and skills, to deterring test misuse, and to accommodating differences among individuals.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Common Core State Standards Initiative (CCSSI). (June, 2010). Retrieved August 1, 2010 from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf.

In the document, Common Core Standards for Mathematics, K-12 mathematics standards are presented. Within each grade level (K-8), an overview of critical areas and recommendations for how

to use instructional time is given. Each grade level also provides "Domains," "Standards," and "Clusters" – the broad to specific trajectory of mathematical content. For high school, rather than grade level standards, content standards have been identified (e.g. Geometry, Modeling, and Algebra).

In addition to the sections on, "Standards for Mathematical Content," the document provides an Introduction and, "Standards for Mathematical Practice," and a Glossary of terms. In the Introduction, it is noted that aims of the mathematics standards is clarity, focus, coherence and being respectful of sequencing standards based on learning theory. A subsection on how to read the document is also provided.

The "Standards for Mathematical Content" section provides guidelines for student "processes and proficiencies," adapted from NCTM and National Research Council's report, Adding it Up. For example, two of the eight process standards defined include, "Make sense of problems and persevere in solving them," and "Look for and make use of structure." A brief summary of how content and practice ought to intersect is provided.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Cook, L. L., Eignor, D. R., Sawaki, Y., Steinberg, J., & Cline, F. (2010). Using Factor Analysis to Investigate Accommodations Used by Students with Disabilities on an English-Language Arts Assessment. ***Applied Measurement in Education, 23***(2), 187–208.

This study compared the underlying factors measured by a state standards-based grade 4 English-Language Arts (ELA) assessment given to several groups of students. The focus of the research was to gather evidence regarding whether or not the tests measured the same construct or constructs for students without disabilities who took the test under standard conditions, students with learning disabilities who took the test under standard conditions, students with learning disabilities who took the test with accommodations as specified in their Individualized Educational Program (IEP) or 504 plan, and students with learning disabilities who took the test with a read-aloud accommodation/modification. The ELA assessment contained both reading and writing portions. A total of 75 multiple-choice items were analyzed. A series of nested hypotheses were tested to determine if the ELA measured the same factors for students with disabilities who took the assessment with and without accommodations and students without disabilities who took the test without accommodations. The results of these analyses, although not conclusive, indicated that the assessment had a similar factor structure for all groups included in the study.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Cormier, D. C., Altman, J. R., Shyyan, V., & Thurlow, M. L. (2010). ***A summary of the research on the effects of test accommodations: 2007-2008*** **(Technical Report 56). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.**

The use of accommodations for both instruction and assessment continues to be of great importance for students with disabilities. The purpose of this report is to provide an update on the state of the research on testing accommodations, as well as to identify promising areas of research

likely to contribute to understanding of current and emerging issues. The research is summarized to facilitate a discussion of trends in current research and to provide a better understanding of the implications related to accommodations use in the development of future policy directions, implementation of current and new accommodations, and the reliable and valid interpretation when used in testing situations. Many of the 40 research studies reviewed sought to study the effects of accommodations on scores or to compare accommodated scores to non-accommodated versions of a similar testing instrument. The most researched content areas were mathematics and reading. Most studies used a large sample of more than 300 participants, who often were K-12 students; students often were from multiple grade levels. Research samples most often included students with learning disabilities compared to other disability classifications. Presentation accommodations were studied by more than half of all the research studies published in 2007-2008. Findings from these studies were mixed for most specific accommodations, such as read-aloud and extended time, as well as for studies in which accommodations were aggregated. There was some consensus on the equivalence of computer-based tests and paper-and-pencil test formats. Appendices include: (1) Research Purposes; (2) Research Characteristics; (3) Assessment/Instrument Characteristics; (4) Participant and Sample Characteristics; (5) Accommodations Studied; (6) Research Findings; and (7) Limitations and Future Research.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Croker, L., (1997). Assessing Content Representativeness of Performance Assessment Exercises. Applied Measurement in Education 10(1), 83-95.

Collection of validation evidence for certification presents new challenges for performance assessments when expert judgments of content are used. In particular, the complexity of the exercises, the newness of the format, the restricted number of tasks, maintaining security for memorable tasks, and the need for scoring rubrics create a range of methodological concerns. Although this study is focused on establishing validity for certification procedures for highly accomplished teachers, the issues raised around performance assessment are relevant for the current context.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Crosson, A. C., Lesaux, N. K., & Martiniello, M. (2008). Factors that Influence Comprehension of Connectives among Language Minority Children from Spanish-Speaking Backgrounds. *Applied Psycholinguistics, 29*(4), 603–625.

This study explores factors influencing the degree to which language minority (LM) children from Spanish-dominant homes understand how connectives, such as "in contrast" and "because", signal relationships between text propositions. Standardized tasks of vocabulary, listening comprehension, word reading, and a researcher-designed text cohesion task were administered to 90 fourth-grade LM students. Understanding of connectives was influenced by vocabulary knowledge and listening comprehension. The degree of challenge that specific connectives posed to LM students was predicted by the difficulty that connectives presented as vocabulary items and also by the type of semantic relationship between clauses they signaled. The findings point to factors that may present

sources of difficulty underlying reading comprehension, in particular the critical role of oral language competencies.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Dana, T. M., & Tippins, D. J. (1993). Considering alternative assessment for middle level learners. *Middle School Journal*, 25(2), 3-5.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Darling-Hammond, L. (1997). *The right to learn: A blueprint for creating schools that work*. San Francisco: Jossey-Bass.

This book outlines the policies and practices needed to create learner-centered, intellectually rigorous, and exciting schools and classroom. Each chapter explores important components of such systems. Chapters include: The Right to Learn, The Limits of the Education Bureaucracy, What Matters for Teaching, Teaching and Learning for Understanding, Structuring Learner-Centered Schools, Staffing Schools for Teaching and Learning, Creating Standards without Standardization, Ensuring Access to Knowledge, Building a Democratic Profession of Teaching, and the Conclusion: An Agenda for Re-Creating Public Education. Throughout the book, Darling-Hammond relies on teacher interview data and research to illustrate best practices and policy implementation that allow for learner-centered contexts.

From pages 114 to 120, Darling-Hammond specifically explores authentic assessment as an important practice for students to demonstrate what they know and are able to do. She argues that performance assessment is "critical to the development of competence" (p. 115). Real-world assessment experience, being inseparable from curriculum and instruction, and motivating to students are three key criteria for high quality performance assessment. She also recommends evaluation of performance assessment should be multi-dimensional; measuring multiple domains, and made clear to students. She cites Vermont's state-wide portfolio system as one of the earliest examples that positively influenced the community, students, and instruction.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Darling-Hammond, L., & Pecheone, R. (2010). *Developing an internationally comparable balanced assessment system that supports high-quality learning*. Presented at the National Conference on Next Generation K–12 Assessment Systems, Center for K–12 Assessment & Performance Management with the Education Commission of the States (ECS) and the Council of Great City Schools (CGCS), Washington, DC.

This paper summarizes assessment results and practices in the U.S. compared to other high achieving countries, an assessment system design that promotes high-quality assessment practices, ways to achieve such lofty goals, and how this assessment system would work. First, the authors

outline the decline of U.S. student performance on the international test, PISA. They argue the U.S. assessment structure is lacking because it relies on multiple choice tests and shallow curriculum. Second, a theory of action is presented that recommends, "An integrated system of curriculum and assessment…will support higher-quality, more coherent instruction." (p. 12) Third, recommendations are put forth to achieve this coherent assessment system. Governmental Roles – at the federal, state, and local district levels are discussed and examples of levels of participation are suggested. Finally, the authors recommend action items and tasks for how to develop such an assessment system (e.g. development of assessments, core standards, and utilizing state consortia).

In conclusion, this paper responds to several "Guiding Questions" that incorporate many examples from other countries' assessment systems (e.g. assessment tasks and blueprints). This section is organized by assessment-related categories such as technical quality, reporting, and accessibility. An overview of a timeline and cost to develop such a system is also provided.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Darlington, R. B. (1971). Another look at "culture fairness." *Journal of Educational Measurement*, 8, 71-82.

Four definitions of cultural fairness" are critically examined. Suggestions for dealing with conflicts between the two goals of maximizing a test's validity and minimizing its culture-group discrimination, are presented. Terms in which this judgment should be made, and methods of using its results are described.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Data Recognition Corporation. (2003). *Fairness in testing: Guidelines for training bias, fairness and sensitivity issues*. Maple Grove, MI: Author.

The most important part of the development of any new test is to ensure balanced treatment and control of potential bias, stereotyping, and insensitivity in the items or in the test-related materials. Data Recognition Corporation (DRC) understands that the presence of any type of bias in a test is undesirable not only from a civil rights point of view, but also from a measurement point of view. Issues of bias, fairness, and sensitivity in testing can have a direct impact on test scores. Our test developers are committed to the development of items and tests that are fair for all students. At every stage of the item and test development process, we employ procedures that are designed to ensure that our items and tests meet Standard 7.4 of the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999).

Component 1	Component 2	Component 3	Component 4	Component 5
				

Dolan, R. P., Burling, K. S., Harms, M., Beck, R., Hanna, E., Jude, J., Murray, E. A., Rose, D. H. & Way, W. (2009). *Universal Design for Computer-Based Testing Guidelines.* Iowa City, IA: Pearson.

The Universal Design for Computer-Based Testing (UD-CBT) guidelines is a systematized representation of the multi-dimensional UD-CBT framework (Harms et al., 2006; Burling et al., 2006) to support test item development and analysis of item designs. These guidelines are organized according to three tiers: test delivery considerations, item content and delivery considerations, and component content and delivery considerations. The component content and delivery considerations tier is further sub-organized according to the various categories of processing students apply during testing; the former two tiers consider the processing categories implicitly. These processing categories, which will be defined in greater detail shortly, were developed from the principles of Universal Design for Learning (UDL, Rose and Meyer, 2002). They provide a logical framework to organize the guidelines within the component categories, and they facilitate the identification of those guidelines relevant to particular student populations, most notably but not exclusively students with disabilities.

Component 1	Component 2	Component 3	Component 4	Component 5
				

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.**

Standards-based reform efforts are highly dependent on accurate assessment of all students, including those with disabilities. The accuracy of current large-scale assessments is undermined by construct-irrelevant factors including access barriers, a particular problem for students with disabilities. Testing accommodations such as the read-aloud have led to improvement, but research findings suggest the need for a more flexible, individualized approach to accommodations. The current pilot study applies principles of Universal Design for Learning to the creation of a prototype computer-based test delivery tool that provides students with a flexible, customizable testing environment with the option for read-aloud of test content. Two contrasting methods were used to deliver two equivalent forms of a National Assessment of Educational Progress United States history and civics test to ten high school students with learning disabilities. In a counterbalanced design, students were administered one form via traditional paper-and-pencil (PPT) and the other via a computer-based system with optional text-to-speech (CBT-TTS). Test scores were calculated, and student surveys, structured interviews, field observations, and usage tracking were conducted to derive information about student preferences and patterns of use. Results indicate a significant increase in scores on the CBT-TTS versus PPT administration for questions with reading passages greater than 100 words in length. Qualitative findings also support the effectiveness of CBT-TTS, which students generally preferred over PPT. The results of this pilot study provide preliminary support for the potential benefits and usability of digital technologies in creating universally designed assessments that more fairly and accurately test students with disabilities.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Dorans, N. (1989). Two new approaches to assessing differential item functioning: standardization and the Mantel-Haenszel method. ***Applied Measurement in Education, 2,*** **3, 217-233.**

The standardization and Mantel-Haenszel approaches to the assessment of differential item functioning (DIF) are described and compared. For rightwrong scoring of items, these two

approaches, which emphasize the importance of comparing comparable groups of examinees, use the same data base for analysis, namely, a 2 (Group) x 2 (Item Score: Correct or Incorrect) x S (Score Level) contingency table for each item studied. The two procedures differ with respect to how they operate on these basic data tables to compare the performance of the two groups of examinees. Whereas the operations employed by Mantel-Haenszel are motivated by statistical power considerations, the operations employed by standardization are motivated by data interpretation considerations. These differences in operation culminate in different measures of DIF effect-size that are very highly related indicators of degree of departure from the null hypothesis of no DIF.

Component 1	Component 2	Component 3	Component 4	Component 5
✓	✓		✓	

Dorans, N., & Holland, P. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. Holland & H. Wainer (Eds.), *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaum.

At the Educational Testing Service, the Mantel-Haenszel procedure is used for differential item functioning (DIF) detection, and the standardization procedure is used to describe DIF. This report describes these procedures. First, an important distinction is made between DIF and impact, pointing to the need to compare the comparable. Then, these two contingency table DIF procedures are described in some detail, first in terms of their own origins as DIF procedures, and then from a common framework that points out similarities and differences. The relationship between the Mantel-Haenszel procedure and item response theory models, in general, and the Rasch model, in particular, is discussed. The utility of the standardization approach for assessing differential distractor functioning is described. Several issues in applied DIF analyses are discussed, including inclusion of the studied item in the matching variable and refinement of the matching variable. Future research topics dealing with the matching variable, the studied variable, and the group variable are also discussed.

Component 1	Component 2	Component 3	Component 4	Component 5
✓	✓		✓	

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*(4), 289-303.

This article discusses the technical quality of performance assessments, making the point that individual performance tasks reflect limited reliability and generalizability of results. It argues that for such assessments to have lasting effects on instruction and learning, their technical properties must be understood and appreciated by developer and practitioner alike. It provides evidence of how increasing the number of tasks to four or five increases reliability dramatically. Of course, with careful attention to content alignment, better representation of the content domain and therefore content validity and generalizability would be enhanced considerably, too.

Component 1	Component 2	Component 3	Component 4	Component 5
		✓		

Eisner, E. W. (1999). The uses and limits of performance assessment. Phi Delta Kappan, 80(9), 658-660.

In this conceptual article, Eisner puts performance assessment, its strengths and limitations, into the broader context of education in the society. He emphasizes the positive attributes of performance assessment but is weary of the U.S. education system embracing such assessment practices. He argues for an educational system that embraces students' individualism and an assessment system that helps influence teaching and learning. One of the positive attributes of performance assessment Eisner describes is tasks having multiple approaches so students can show what they know in multiple ways. An additional benefit of performance assessment is students are judged by human scorers on what they know and are able to do. These judgments allow for multiple "right answers" and interpretations of the given content.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Edman, P. (1992). *Tactile Graphics.* American Foundation for the Blind Press.

A handbook on how to create tactile graphics, covering theory, techniques, materials, and instructions for creating tactile materials.

Component 1	Component 2	Component 3	Component 4	Component 5
				

ETS. (2002). *The ETS Standards for Quality and Fairness*. Princeton, NJ: Author.

The thirteen standards presented in this guide are designed to help ETS design, develop, and deliver technically sound, fair and useful products and services and to help auditors evaluate these products and services.

Component 1	Component 2	Component 3	Component 4	Component 5
				

ETS. (2009a). *ETS Guidelines for Fairness Review of Assessments*. Princeton, NJ: Author.

The *ETS guidelines for fairness review of assessments* provides a comprehensive manual used in the fairness review for test development staff. The document is organized around cognitive, affective, and physical sources of bias and examples for each bias category is provided.

Component 1	Component 2	Component 3	Component 4	Component 5
				

ETS. (2009b). *ETS International Principles for Fairness Review of Assessments.* Princeton, NJ: Author.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Educational Testing Service. (2010). *How ETS works to improve test accessibility*. Princeton, NJ: Educational Testing Service.

This document describes enhancements to assessments for test takers with disabilities. It provides practical guidance for ensuring that the test taker can interact appropriately with the content, presentation, and response mode of the test. Grounded in a philosophy of inclusiveness for all test takers, the case is made that content and format of assessments should allow all students to demonstrate their mastery of the knowledge, skills, and abilities being assessed.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Eells, K. (1948). *Social status in intelligence test items.* Unpublished doctoral dissertation. University of Chicago.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Eisner, E. W. (1999). The uses and limits of performance assessment. *Phi Delta Kappan*, *80*(9), 658-660.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Eklöf, H. (2006). *Motivational beliefs in the TIMSS 2003 context: Theory, measurement and relation to test performance*. Unpublished doctorate, Umea University.

One of the main aims of this thesis was to explore issues related to student test-taking motivation and achievement in the large-scale testing context. Swedish students made ratings of their mathematics test-taking motivation before participating in the mathematic component of TIMSS 2003. Their ratings of test-taking motivation were positively but rather weakly related to achievement. Exploratory factor analysis suggests that the test-taking motivation construct is distinct from general attitudes towards a specific academic subject or domain.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing*, 7(3), 311-326.

This study explores the reported level of test-taking motivation and the relation between test-taking motivation and mathematics achievement in a sample of Swedish eighth-grade students participating in TIMSS 2003. A majority of students reported that they were motivated to do their best in TIMSS. Test-taking motivation was positively related to mathematics achievement with a small effect. Interestingly, gender comparisons showed that test-taking motivation was positively, but not significantly related to achievement in boys, and was unrelated to achievement in girls. This result was probably due to the larger variability in the ratings by boys on test-taking motivation.

Component 1	Component 2	Component 3	Component 4	Component 5
✓				

Elliott, R. (1987). *Litigating Intelligence*. Dover, MA: Auburn House, 139 ff.

Component 1	Component 2	Component 3	Component 4	Component 5
			✓	

Elliott, S., Kettler, R., Beddow, P., Kurz, A., Compton, E., McGrath, D., Bruen, C., Hinton, K., Palmer, P., Rodriguez, M., Bolt, D., Roach, A. (2010). Effects of using modified items to test students with persistent academic difficulties. ***Exceptional Children, 76***(4).

This study investigated the effects of using modified items in achievement tests to enhance accessibility. An experiment determined whether tests composed of modified items would reduce the performance gap between students eligible for an alternate assessment based on modified achievement standards (AA-MAS) and students not eligible, and the impact on student proficiency levels. Three groups of eighth-grade students (N = 755) from four states took original and modified versions of reading and mathematics tests. Findings indicate modified item conditions were significantly easier for all students and modifications would result in more AA-MAS eligible students meeting proficiency status. Study limitations and follow-up research on item modifications and the performance of students with disabilities are discussed.

Component 1	Component 2	Component 3	Component 4	Component 5
✓	✓			✓

Erdogan, Y, (2008). Legibility of Websites Which are Designed for Instructional Purposes. ***World Applied Sciences Journal 3***(1).

In this study, legibility of web pages which are designed for instructional purposes were compared according to font types and foreground/background color combinations. Three different font styles which are sans-serif (Verdana), serif (Times New Roman) and monotype (Courier New) were used. 15 background/foreground color combinations were investigated; basic colors that are available on most browsers were selected. Also four different background/foreground color contrast combinations were chosen (dark text on dark ground, light text on light ground, light text on dark ground and dark text on light ground). In the current study a survey method was used to investigate the attitudes of the students towards the legibility of web pages. The sample consisted of 124 students of the Computer and Instructional Technologies Department, Istanbul. All the students were capable of visual literacy and web technologies. At the end of the study it was found that Verdana is regarded as the most legible font type and white ground/black text is the best color combination for

web pages. Also, there was a significant difference according to color contrast; web pages which are prepared with dark text on light ground are more legible than other combination.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Fedorchak, G., & Russell, M. (2007). Universal Access to Assessment. Enhanced Assessment Grant. Retrieved from http://www2.ed.gov/programs/eag/awards07.html#nh.

The project seeks to examine the feasibility, effect, and capacity to deliver state achievement tests using a computer-based test delivery system specifically designed to provide universal access to test content for students with disabilities or special needs. The project is a direct outgrowth of prior work supported by EAG funding conducted by New Hampshire, Vermont, and Rhode Island in which the feasibility of using computers to provide specific test accommodations was examined. Based on this prior work, members of the New Hampshire Department of Education Curriculum and Assessment program conducted a statewide pilot test in which the interface used for the prior EAG project was used to provide a read aloud accommodation to students for its 2006 grade 10 mathematics test. This successful pilot led to a collaborative effort with Nimble Assessment Systems to develop a comprehensive test delivery system that employed principles of universal design to flexibly meet the accessibility and accommodation needs of individual students. The project brings together 11 states to examine the feasibility and effect of using this comprehensive test delivery system to improve test validity for students with disabilities and special needs who are believed to benefit from one or more of the accessibility and accommodation tools built into the system. Specifically, members of this collaborative project include: New Hampshire, Vermont, Rhode Island, South Carolina, North Carolina, Georgia, Montana, Iowa, Connecticut, Maryland, and Florida. In addition, the proposed project includes partnerships with the National Center for Educational Outcomes, Nimble Assessment Systems, and the NECAP state contractor (currently Measured Progress).

Component 1	Component 2	Component 3	Component 4	Component 5
				

Fife, J. H. (2011). Automated Scoring of CBAL Mathematics Tasks with *m-rater*. ETS Research Manuscript #RM-11-12.

For the past several years, ETS has been engaged in a research project known as Cognitively Based Assessment of, for, and as Learning (CBAL). The goal of this project is to develop a research-based assessment system that provides accountability testing and formative testing in an environment that is a worthwhile learning experience in and of itself. An important feature of the assessments in this system is that they are computer-delivered, with as many of the tasks as possible scored automatically. For the automated scoring of mathematics items, ETS has developed m-rater scoring engine. In the present report, I discuss the m-rater–related automated scoring work done in CBAL Mathematics in 2009. Scoring models were written for 16 tasks. These models were written in Alchemist, a software tool originally developed for the writing of c-rater™ scoring models (c-rater is ETS's scoring engine for scoring short text responses for content). In 2009 the c-rater support team completed a collection of enhancements to Alchemist that enables the user to write m-rater scoring models. This collection of enhancements is known as KeyBuilder. In 2009 I reviewed the literature to see to what extent problem solutions that are expressed in the form of a structured sequence of equations can be automatically evaluated.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Folling-Albers, M., & Hartinger, A. (1998). Interest of girls and boys in elementary school. In L. Hoffmann, A. Krapp, K. A. Renninger & J. Baumert (Eds.), Interest and learning. *Proceedings of the Seeon-Conference on interest and gender* (pp. 175-183). Kiel: IPN.

The primary goal of this study was to explore interests that elementary school children have in various in-school and out-of-school contexts and activities. Individual studies were conducted to describe the interests of elementary school students and to investigate the stability of focal points of interest. The second part of the study investigated whether taking specific teaching variables into consideration can advance the interest of children in a prescribed syllabus topic. Findings indicate that certain specific teaching/learning variables, derived from pedagogical interest theory, can be influenced by children's subject-specific interests.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Fuchs, L. S., & Fuchs, D. (1999). Fair and unfair testing accommodations. *School Administrator, 56*, 24–29.

Test accommodations are changes in standardized test conditions to equalize opportunities between students with or without disabilities by achieving valid scores. The Individuals with Disabilities Act 1997 amendments require states and districts to include disabled students in accountability programs. Assumptions, practical implications, methodologies, and resources are discussed.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Gallagher, A. M., Bennett, R. E., & Cahalan, C. (2000). *Detecting construct-irrelevant variance in open-ended computerized mathematics tasks (*RR-00-18). Princeton, NJ: Educational Testing Service.

The purpose of this study was to evaluate whether variance due to computer-based presentation was associated with performance on a new constructed-response type--Mathematical Expression--that requires examinees to build mathematical expressions using a mouse and an on-screen tool palette. Participants took parallel computer-based and paper-based tests consisting of Mathematical Expression items, plus a test of their skill in entering and editing data using the computer interface. Comparisons of mean performance, reliability, speededness, and relations with external indicators were conducted across the paper-based and computer-based tests; also, computer-based math score was regressed on edit/entry score after controlling for paper-and- pencil math score and background information. Although no statistical evidence of construct-irrelevant variance was detected, some examinees reported mechanical difficulties in responding and indicated a preference for the paper-and-pencil test.

Component 1	Component 2	Component 3	Component 4	Component 5

				

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.

Examined data from several national testing programs to determine whether the change from paper-based administration to computer-based tests influences group differences in performance. Results from four college and graduate entrance examinations and a professional licensing test show that African Americans and, to a lesser degree, Hispanics, appear to benefit from the shift. Discusses implications of these results.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Garcia, T., Short, C. (2008). Legibility and readability on the World Wide Web. Downloaded from http://bigital.com/english/files/2008/04/web_legibility_readability.pdf Bigital: Buenos Aires, Argentina.

This study compared the results of students taking tests online using tests formatted with different typefaces. For this test, the size and leading of the texts was established as 12/15 pixels and 14 pixels for titles; black text on white background. The typefaces tested were are:- Verdana- Helvetica / Arial- Georgia- Times / Times New Roman- Trebuchet MS,- Courier / Courier New- Comic Sans MS. Students performed better on the test formatted with Verdana for both average reading time and comprehension.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Garden, R. (1999). Development of TIMSS Performance Assessment Tasks. *Studies in Educational Evaluation, 25*(3), 217-41.

Describes the development of the performance assessment tasks of the Third International Mathematics and Science Study. The challenge was to produce tasks that would measure the achievement of curricular objectives while being sufficiently reliable to allow comparisons between countries and of groups within countries.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Gibson, W. (2001). Johnny mnemonic. In D. Trend (Ed.), *Reading Digital Culture* (pp. 57-69). Malden, MA: Blackwell.

Computer technology has transformed many fundamental parts of life: work and leisure environments, communication and consumption patterns, knowledge creation and learning practices and participation in politics and public life. Reading Digital Culture is a comprehensive collection of the most influential essays on digital media written on the brink of the new millennium and foreshaows many issues relevant to the implementation of a large-scale digitally delivered assessment program including the culture of digital reading.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Goldberg, A., Russell, M., & Cook, A. (2003). Effects of computers on student writing: A meta-analysis of research 1992–2002. *Journal of Technology, Learning and Assessment, 2*(1). Available: http://www.bc.edu/research/intasc/jtla/journal/v2n1.shtml.

Meta-analyses were performed including 26 studies conducted between 1992–2002 focused on the comparison between K–12 students writing with computers vs. paper-and-pencil. Significant mean effect sizes in favor of computers were found for quantity of writing (d=.50, n=14) and quality of writing (d= .41, n=15). Studies focused on revision behaviors between these two writing conditions (n=6) revealed mixed results. Others studies collected for the meta-analysis which did not meet the statistical criteria were also reviewed briefly. These articles (n=35) indicate that the writing process is more collaborative, iterative, and social in computer classrooms as compared with paper-and-pencil environments. For educational leaders questioning whether computers should be used to help students develop writing skills, the results of the meta-analyses suggest that on average students who use computers when learning to write are not only more engaged and motivated in their writing, but they produce written work that is of greater length and higher quality.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Gong, B. (2009). *Innovative assessment in Kentucky's KIRIS system: Political considerations.* Paper presented at the Best Practices in State Assessment workshop sponsored by the Board of Testing and Assessment, National Academy of Sciences. Washington, DC.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Graf, E. A. (2008). Approaches to the design of diagnostic item models (ETS Research Report No. RR-08-07). Princeton, NJ: Educational Testing Service.

Quantitative item models are item structures that may be expressed in terms of mathematical variables and constraints. An item model may be developed as a computer program from which large numbers of items are automatically generated. Item models can be used to produce large numbers of items for use in traditional, large-scale assessments. But they have potential for use in other areas as well, including diagnostic assessment. In this report, I first review research on diagnostic

assessment and then discuss how approaches to diagnostic assessment can inform the design of diagnostic item models.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Graf, E. A. (2009). *Defining mathematics competency in the service of cognitively based assessment for grades 6 through 8* (RR-09-42). Princeton, NJ: Educational Testing Service.

This report makes recommendations for the development of middle-school assessment in mathematics, based on a synthesis of scientific findings in cognitive psychology and mathematics education. The focus is on background research, rather than test specifications or example tasks. Readers interested in early development and pilot efforts associated with the Cognitively Based Assessment of, for, and as Learning (CBAL) project in mathematics (for which this review helped provide a theoretical foundation) should consult Graf, Harris, Marquez, Fife, and Redman (2009).The organization of the report is motivated by the evidence-centered design (ECD) approach to assessment developed by Mislevy and colleagues (e.g., see Mislevy, Steinberg, & Almond, 2003).

Component 1	Component 2	Component 3	Component 4	Component 5
				

Graf, E. A., Peterson, S., Steffen, M., & Lawless, R. (2005). Psychometric and cognitive analysis as a basis for the design and revision of quantitative item models (ETS Research Report No. RR-05-25). Princeton, NJ: Educational Testing Service.

We describe the item modeling development and evaluation process as applied to a quantitative assessment with high-stakes outcomes. In addition to expediting the item-creation process, a model-based approach may reduce pretesting costs, if the difficulty and discrimination of model-generated items may be predicted to a predefined level of accuracy. The development and evaluation of item models represents a collaborative effort among content specialists, statisticians, and cognitive scientists. A cycle for developing and revising item models that generate items with more predictable statistics is described. We review the goals of item modeling from different perspectives and recommend a method for structuring families of models that span content and generate items with more predictable psychometric parameters.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Gregg, N., & Nelson, J. M. (2012). Meta-analysis on the effectiveness of extra time as a test accommodation for transitioning adolescents with learning disabilities: More questions than answers. *Journal of Learning Disabilities.* 45, 17-30.

The accommodation of students with learning disabilities (LD) on mandatory high stakes tests continues to heighten concern over the equity and effectiveness of current practices. As students transition from high school, they are required to complete timed graduation tests and postsecondary entrance examinations. The most common accommodation accessed by transitioning adolescents with LD is extended time. In order to inform test accommodation practices, a meta-analysis was

conducted to address whether test scores from accommodated (i.e., extended time only) and standard test administrations are comparable for transitioning adolescents with LD as compared to their normally achieving peers. The results of the meta-analyses raised more questions than answers and highlighted the need for future research in this area.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Grisay, A. and C. Monseur, (2007). Measuring the equivalence of item difficulty in the various versions of an international test. ***Studies in Educational Evaluation*** **33, 69-86.**

In this article, PISA data from the Reading Literacy study conducted in 2000 and 2001 were analyzed in order to explore equivalence issues across the 47 test adaptations in various languages of instruction that were used by the participating countries. On average, about 82% of the variance in relative item difficulty was found to be common across the various national versions. However, the index of equivalence appeared to be lower than desirable in certain categories of countries. Tentative analyses were conducted to better understand the reasons behind these differences.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Haas, C. & Hayes, J. (1986). What did I just say? Reading problems in writing with the machine. ***Research in the Teaching of English, 20*****(1), 22–35.**

Relates how 16 "computer writers" felt about how they use the computer for writing tasks and reports on three experimental studies that compared the performance of college students reading texts displayed on a computer terminal screen and on a printed hard copy. Findings showed that visual/spatial factors influenced locational recall, information retrieval, and appropriate reordering of text.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Haertel, E. H. (1999). Performance assessment and education reform. ***Phi Delta Kappan*****,** ***80*****(9), 662-666.**

Component 1	Component 2	Component 3	Component 4	Component 5
				

Haladyna, T. (1992). The Effectiveness of Several Multiple-Choice Formats. ***Applied Measurement in Education, 5*****(1), 73-88.**

Several multiple-choice item formats are examined in the current climate of test reform. The reform movement is discussed as it affects use of the following formats: (1) complex multiple-choice; (2) alternate choice; (3) true-false; (4) multiple true-false; and (5) the context dependent item set.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item writing rules. *Applied Measurement in Education*, 2(1), 37-50.

This article presents and discusses a taxonomy of multiple-choice item-writing rules covering procedural and content concerns related to item writing and guidelines for stem and option construction. The taxonomy derives from an analysis of 46 authoritative textbooks and other sources in the educational measurement literature. The analysis also leads to a 'validity by consensus' for each rule. The taxonomy is viewed as a complete and authorative set of guidelines for writing multiple-choice items

Component 1	Component 2	Component 3	Component 4	Component 5
				

Hansen, E. G., Forer, D., & Lee, M. (2004). *Toward accessible computer-based tests: Prototypes for visual and other disabilities (*RR-04-25). Princeton, NJ: Educational Testing Service.

There is a great need to explore approaches for developing computer-based testing systems that are more accessible for people with disabilities. This report explores three prototype test delivery approaches, describing their development and formative evaluations. Fifteen adults, 2 to 4 from each of the six disability statuses--blindness, low vision, deafness, deaf-blindness, learning disability, and no disability--participated in a formative evaluation of the systems. Each participant was administered from 2 to 15 items in each of one or two of the systems. The study found that although all systems had weaknesses that should be addressed, almost all of the participants (13 of 15) would recommend at least one of the delivery methods for high-stakes tests, such as those for college or graduate admissions. The report concludes with recommendations for additional research that testing organizations seeking to develop accessible computer-based testing systems can consider.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Hansen, E.G., Laitusis, C.C., Frankel, L., & King, T. (2010). *Improving the accessibility of computer-based reading task: A focus group of teachers of students with visual impairments* (RR Number pending). Princeton, NJ: Educational Testing Service.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Hansen, E. G., Laitusis, C. C., Frankel, L., & King, T. (in press). Designing accessible technology-enabled reading assessments: Recommendations from teachers of students with visual impairments. *Journal of Blindness Innovation Research.*

There is a great need to ensure that innovative technology-enabled assessments are accessible for students with disabilities. This study examined the severity of accessibility challenges that students with visual disabilities (ranging from low vision to complete blindness) would encounter in accessing prototype reading tasks from ETS's Cognitively-Based Assessments of, for, and as Learning (CBAL) research system. This focus group study involved presenting prototype tasks to six teachers of students with visual impairments. The teachers (a) examined the prototype tasks, (b) evaluated the severity of accessibility problems that would be encountered by students having different primary methods for accessing text (e.g., braille, audio, visual enhancement), and (c) offered suggestions about how to improve the accessibility of the tasks. The report summarizes the results of this study and provides recommendations for improving the accessibility of innovative reading tasks.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Hansen, E. G., & Mislevy, R. J. (2006). Accessibility of computer-based testing for individuals with disabilities and English language learners within a validity framework. In M. Hricko & S. Howell (Eds.), *Online assessment and measurement: Foundation, challenges, and issues*. Hershey, PA: Idea Group Publishing, Inc.

There is a great need for designers of computer-based tests and testing systems to build accessibility into their designs from the earliest stages, thereby overcoming barriers faced by individuals with disabilities and English language learners. Some important potential accessibility features include text-to-speech, font enlargement and screen magnification, online dictionaries, and extended testing time. Yet accessibility features can, under some circumstances, undermine the validity of test results. Evidence centered assessment design (ECD) is offered as a conceptual framework—providing sharable terminology, concepts, and knowledge representations—for representing and anticipating the impact of accessibility features on validity, thus helping weigh the consequences of potential design alternatives for accessible computer-based tests and testing systems.

Component 1	Component 2	Component 3	Component 4	Component 5
				

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

There is a great need to help test designers determine how to make tests that are accessible to individuals with disabilities. This report takes design patterns, which were developed at SRI for assessment design, and uses them to clarify issues related to accessibility features for individuals with disabilities—such as low-vision and blindness—taking a test of reading. Design patterns appear useful in clarifying how variable features of a test design need to be matched to disability-related characteristics of test takers in order to ensure accessibility. Giving consideration to accessibility issues during the development and use of design patterns may help improve the validity and fairness of tests, as well as their accessibility for individuals with disabilities.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Hansen, E. G., Mislevy, R. J., & Steinberg, L. S. (2008). *Evidence centered assessment design for reasoning about testing accommodations in NAEP reading and mathematics* (ETS Research Report RR-08-28). Princeton, NJ: ETS.

Accommodations play a key role in enabling individuals with disabilities to participate in the National Assessment of Educational Progress (NAEP) and other large-scale assessments. However, it can be difficult to know how accommodations affect the validity of results, thus making it difficult to determine which accommodations should be allowed. This study describes recent extension of evidence-centered assessment design (ECD) for reasoning about the impact of accommodations and other accessibility features (e.g., universal design features) on the validity of assessment results, using examples from NAEP reading and mathematics. The study found that the ECD-based techniques were useful in analyzing the effects of accommodations and other accessibility features on validity. Such design capabilities may increase assessment designers' capacity to employ accessibility features without undermining validity.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Hansen, E. G., Mislevy, R. J., Steinberg, L. S., Lee, M. J., & Forer, D. C. (2005). Accessibility of tests for individuals with disabilities within a validity framework. *System: An International Journal of Educational Technology and Applied Linguistics, 33*(1), 107-133.

There is a great need to ensure that language tests are accessible to individuals with disabilities. Yet accessibility features can sometimes conflict with the validity of test scores. In some cases the nature of the conflict seems obvious, yet in other cases there is controversy, such as that concerning the use of a "readaloud" accessibility feature on tests of reading. What is needed is a more rigorous approach for reasoning about the validity implications of accessibility features. The approach described in this article seeks to integrate thinking about accessibility, task design, and validity – all in a framework of sharable terminology, concepts, and knowledge representations. We believe that such a framework can allow one to more accurately and quickly identify the validity-related consequences of design changes that are intended to improve accessibility for individuals with disabilities. Such a framework may permit greater inclusion of individuals with disabilities or other sub-populations without invalidating test results.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Hansen, E. G., Shute, V. J., & Landau, S. (2010). An Assessment-for-Learning System in Mathematics for Individuals with Visual Impairments. *Journal of Visual Impairment & Blindness, 104*(5), 275–286.

This study examined the usability of an assessment-for-learning (AfL) system that provides audio-tactile graphics for algebra content (geometric sequences) for individuals with visual impairments--two who are blind and two with low vision. It found that the system is generally usable as a mathematics AfL system.

Component 1	Component 2	Component 3	Component 4	Component 5

				

Hansen, E. G. & Steinberg, L.Ss. (2004). *Evidence Centered Assessment Design for reasoning about testing accommodations in NAEP reading and math.* Paper commissioned by the Committee on Participation of English Language Learners and Students with Disabilities in NAEP and Other Large-Scale Assessments of the Board on Testing and Assessment (BOTA) of the National Research Council.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Hansen, E. G., & Zapata-Rivera, D. (2010). *Evidence centered design for an accessible game-based assessment-for-learning system for middle school mathematics.* Presentation at the annual meeting of the National Council on Measurement in Education (NCME) on May 3, 2010 in Denver, Colorado.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Harmon, M., Smith, T., Martin, M., Kelly, D., Beaton, A., Mullis, I., Gonzalez, E., & Orpwood, G. (1997). Performance Assessment in IEA's Third International Mathematics and Science Study (TIMSS). Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

This report presents the initial findings from the TIMSS performance assessment. Some 1,500 schools and 15,000 students from 21 countries participated, making it the largest international performance test yet conducted. This report describes the TIMSS performance assessment and provides a detailed summary of the performance of the students in each participating country on every item of every task.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Harris, K., Bauer, M. I. (2009). *Using assessment to infuse a rich mathematics disciplinary pedagogy into classrooms*. Proceedings of the 35th International Association for Educational Assessment (IAEA) Annual Conference. Brisbane, Australia, September 13–18.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Hegarty, M. (2004). Commentary: Dynamic visualizations and learning: Getting to the difficult questions. *Learning and Instruction, 14*, 343–351.

The rapid development of computer graphics technology has made possible an easy integration of dynamic visualizations into computer-based learning environments. This study examines the relative

effectiveness of dynamic visualizations, compared either to sequentially or simultaneously presented static visualizations. Moreover, the degree of realism in the visualizations was manipulated experimentally. One hundred-and-twenty university students were randomly assigned to one of six conditions (3 × 2; between-subjects; presentation format × realism). Learners' visuo-spatial abilities were considered as a continuous moderator for the presentation format. Learning outcomes were measured by a pictorial locomotion pattern classification test. Dynamic conditions outperformed static-sequential ones, but not static-simultaneous conditions, in classification performance. Realism had no main effect and did not interact with the presentation format as expected. Learners' visuo-spatial abilities had a positive effect on learning outcomes, but did not moderate the effects of the presentation format. Implications of the results for the design of instructional materials are discussed.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Henderson, P., & Karr-Kidwell, P. J. (1998). *Authentic assessment: An extensive literary review and recommendations for administrators* (Report No. TM 028 235). (ERIC Document Reproduction Service No. ED418140).

The purpose of this literature review was to explore the use and characteristics of authentic assessment and provide recommendations to administrators on how to implement such assessment types. The impetus for the review came from a growing concern that standardized, multiple-choice tests were jeopardizing classroom instruction and limiting the teaching of higher order thinking skills and problem solving. The authors begin their review with an historical summary of United States testing (from the 1800's thru 1990) and associated criticisms. They then use conceptual literature and some research studies to summarize characteristics and benefits of authentic assessment (e.g. real-world tasks, encouragement of student self-assessment, and linked to classroom instruction). The last two sections of this review provide suggestions for how teachers develop authentic assessments and implement them. It should be noted these recommendations, link to other literature, focus only on classroom-based performance assessment development. In conclusion, additional recommendations are made specifically to administrators. Examples of recommendations include purchasing resources, planning staff developing, and allocating time for teachers to create tasks.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Herman, J. (1997). Assessing new assessments: How do they measure up? *Theory into Practice*, 36(4), 196-204.

In this theoretical piece, Herman argues that with expectations of student learning and knowledge changing, so are the types of assessments used to assess achievement. The current role of assessment, qualities for assessment, and an analysis of the implications of classroom and large-scale assessment are discussed. To begin the article, the author summarizes the power of assessment used with instruction. She then summarizes the negative consequences of multiple-choice tests (e.g. not assessing higher level thinking skills) and calls for the "Promise" of alternative assessment. Qualities of alternative assessments are described, to include: alignment with standards, fairness issues, opportunity to learn, concerns for bias, technical accuracy, consistency across tasks, and utility for improving instruction and learning. Cost and the importance of public

support are also mentioned. In closing, Herman discusses the "Road Ahead" and how alternative assessment might fit into a large scale context.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Hess, K. (2011). *Content Specifications with Content Mapping for the Summative Assessment of the Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects*. [White Paper]. Retrieved from http://www.k12.wa.us/SMARTER/ContentSpecs/ELA-LiteracyContentSpecifications.pdf.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Hess, K., & Biggam, S. (2004). *A Discussion of 'Increasing Text Complexity,' New Hampshire, Rhode Island, and Vermont Departments of Education.*

This paper outlines factors that can influence the level of complexity of a text. The authors consider word difficulty and language structure, text structure, discourse style, genre and characteristic features of the text, background knowledge/familiarity with content, level of reasoning required, format and layout of the text and text length as key factors in a model of text complexity. The proposed model demonstrates features of increasing text complexity by grade levels and demonstrates this hierarchy with examples from literature.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Hessinger, J., Bridgeman, B., & Cline, F. (2003). *Effect of Extra Time on GRE Quantitative and Verbal Scores,* (ETS Research Report Series RR-03–13). Princeton, NJ: Educational Testing Service.

The purpose of this study was to test the assumption that the Graduate Record Examinations® (GRE®) General Test is a measure of academic reasoning abilities in which speed of responding plays at most a minor role. In addition to completing the operational GRE General Test, participants each completed a research version of either the GRE verbal or quantitative test within either the standard time limit or within one-and-a-half times the standard time limit. Scores obtained from 15,948 examinees indicate that extra time added about seven points (on the 200-800 score scale) to examinees' verbal scores and seven points to their quantitative scores. Scores under the different timing conditions were generally comparable across gender and ethnic groups, but quantitative scores were slightly higher for lower-ability examinees who had more time.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Hidi, S., & McLaren, J. (1990). The effect of topic and theme interestingness on the production of school expositions. In H. Mandl, E. D. Corte, N. Bennett & H. F. Friedrich (Eds.), *Learning and instruction: European research in an international context* (Vol. 2.2, pp. 295-308). Oxford: Pergamon.

The study investigated the relationship between interest generated by topic and theme for elementary school students in the context of expository writing. The findings showed that children in grades 4 and 6 had interest ratings of topics and themes that only moderately correlated with adult's ratings. A subsequent qualitative analysis of the topical rating differences showed that children tended to be generally less interested in social science topics than adults. Additionally, children were most interested when they had moderate knowledge of the topics. This study is relevant in terms of the high incidence of adult selection of stimulus materials for assessment projects.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Higgins, D., Futagi, Y., & Deane, P. (2005). *Multilingual generalization of the modelcreator software for math item generation* (RR-05-02). Princeton, NJ: Educational Testing Service.

This paper reports on the process of modifying the ModelCreator item generation system to produce output in multiple languages. In particular, Japanese and Spanish are now supported in addition to English. The addition of multilingual functionality was considerably facilitated by the general formulation of our natural language generation system, which proceeds in three stages, whereby the first two stages are largely language-independent. This multilingual capability is very promising for the domain of educational assessment because, when combined with item modeling, it provides the potential to produce a large pool of automatically constructed items, with difficulty estimates and balanced forms for speakers of different languages.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Higgins, J. & Katz, M. (2011). MeTRC Accessible Assessment Research Study Report. University of Oregon.

The learning and assessment research strand of MeTRC investigates the effects of using accessibility tools within an electronic content delivery environment on students' math performance in relationship to the skills and knowledge being developed and assessed. The goal is to evaluate strategies for altering or supporting the presentation of text within math learning and assessment activities to increase accessibility and understanding. This report summarizes research that seeks to understand the extent to which the application of different scripting rules to generate alternate audio-based representations of mathematics expressions affect student understanding of that content or the test questions designed to measure that content.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Higgins, J., Russell, M., & Hoffmann, T. (2005). Examining the effect of computer-based passage presentation on reading test performance. *Journal of Technology, Learning, and Assessment, 3*(4). Available from http://www.jtla.org.

To examine the impact of transitioning 4th grade reading comprehension assessments to the computer, 219 fourth graders were randomly assigned to take a one-hour reading comprehension assessment on paper, on a computer using scrolling text to navigate through passages, or on a computer using paging text to navigate through passages. This study examined whether presentation form affected student test scores. Students also completed a computer skills performance assessment, a paper based computer literacy assessment, and a computer use survey. Results from the reading comprehension assessment and the three computer instruments were used to examine differences in students test scores while taking into account their computer skills. ANOVA and regression analyses provide evidence of the following findings: 1. There were no significant differences in reading comprehension scores across testing modes. On average, students in the paper group (n=75) answered 58.1% of the items correctly, students in the scrolling group (n=70) answered 52.2% of the items correctly, and students in the whole page group (n=74) answered 56.9% of the items correctly. Te almost a 6% point difference in scores between the paper and scrolling groups was not significant at the p less than 0.05 or p less than 0.1 level. Although the results suggest that, across all students, the modal effect is not statistically significant, this finding may be due in part to the unusually high computer access and higher socio-economic status of the sample. 2. There were no statistically significant differences in reading comprehension scores based on computer fluidity and computer literacy, but a pattern in performance suggests that students are disadvantaged by the scrolling text mode, particularly students with lower computer skills. 3. The majority of students who took the reading test on a computer indicated that they would prefer to take the test on computer. Although this sample did not include many students who had limited prior computer experience, the survey responses, completion rates, and student observations provide evidence that computer anxiety generally did not interfere with students' ability to take the assessment. 4. Providing highlighters and review markers is useful for some students. The results of this study suggest that further research is warranted to understand differences in scores when reading comprehension assessments are administered via computer to a larger and more diverse group of students.

Component 1	Component 2	Component 3	Component 4	Component 5
✓	✓			

Hoffmann, T. (2007). Final report for universal assessment system phase I research. Wellesley, MA: Nimble Assessment Systems.

The primary goal of our Phase I project was to develop and pilot a flexible test delivery system that was capable of providing students with several accommodations. These accommodations included Read Aloud of text, Magnification of test items, Masking of test items, and access to a digital talking calculator. In addition, the project aimed to develop an interface that would enable teachers to activate or deactivate specific tools for individual students depending upon their accommodation needs.

Component 1	Component 2	Component 3	Component 4	Component 5
				✓

Holland, P., & Thayer, D. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test Validity*. Hillsdale, NJ: Lawrence Erlbaum.

This chapter provides a detailed explanation of the use of the Matel-Haenszel procedure for differential item functioning.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *The Journal of Technology, Learning and Assessment, 5* (2).

This study investigated the comparability of scores for paper and computer versions of a writing test administered to eighth grade students. Two essay prompts were given on paper to a nationally representative sample as part of the 2002 main NAEP writing assessment. The same two essay prompts were subsequently administered on computer to a second sample also selected to be nationally representative. Analyses looked at overall differences in performance between the delivery modes, interactions of delivery mode with group membership, differences in performance between those taking the computer test on different types of equipment (i.e., school machines vs. NAEP-supplied laptops), and whether computer familiarity was associated with online writing test performance. Results generally showed no significant mean score differences between paper and computer delivery. However, computer familiarity significantly predicted online writing test performance after controlling for paper writing skill. These results suggest that, for any given individual, a computer-based writing assessment may produce different results than a paper one, depending upon that individual's level of computer familiarity. Further, for purposes of estimating population performance, as long as substantial numbers of students write better on computer than on paper (or better on paper than on computer), conducting a writing assessment in either mode alone may underestimate the performance that would have resulted if students had been tested using the mode in which they wrote best.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Hume, A. (2005). The Anatomy of Web Fonts. Downloaded from http://www.sitepoint.com/anatomy-web-fonts/

The variable boldness and fine extra strokes of the serif fonts, particularly at smaller sizes of body text, often appear pixilated and untidy. This is still the case even with the most modern anti-aliasing techniques. With anti-aliasing enabled, the serif fonts look blurred (which is exactly what they are) around their curves and terminals. On the other hand, the straight, low contrast, open strokes of a sans-serif font, such as Verdana, will always leave a good impression on-screen.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Johnson, E., Kimball, K., & Brown, S.O. (2001). American sign language as an accommodation during standards-based assessments. *Assessment For Effective Intervention, 26*(2), 39–47.

A study investigated whether the use of American Sign Language as an accommodation affected the validity of standards-based assessments given in 12 classrooms of students with hearing impairments. Findings indicate sign language translation can result in the omission of pertinent information required to answer test items correctly. Suggestions are provided.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Johnstone, C. J., Altman, J., Thurlow, M. L., & Thompson, S. J. (2006). *A summary of research on the effects of test accommodations: 2002 through 2004* (Technical Report 45). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

The No Child Left Behind Act of 2001 (NCLB) requires the reporting of participation in assessments overall and by subgroup, including students with disabilities. As states and school districts strive to meet the goals for adequate yearly progress required by NCLB, the use of individual accommodations continues to be scrutinized for effectiveness, threats to test validity, and score comparability. This report summarizes 49 empirical research studies completed on test accommodations between 2002 and 2004, and provides direction in the design of critically needed future research on accommodations. NCEO found that studies during this three-year period had the following characteristics: (1) Purpose: The primary purpose of the 2002-2004 accommodations research was to determine the effects of accommodations use on the large-scale test scores of students with disabilities; (2) Types of assessment, content areas, and accommodations: The majority of the studies tested students using norm-referenced or criterion-referenced tests, on math or reading/language arts; (3) Participants: Equal numbers of research studies involved between 1-100 participants, 100-1,000 participants, and more than 1,000 participants of multiple age categories. Participants were varying percentages of students without disabilities and students with disabilities. Students with learning disabilities were studied most frequently among students who receive special education services; (4) Findings: Findings shared no common theme, with various accommodations shown to have both a positive and non-positive effect on scores. Individual accommodations showed either differential item functioning or no differential item functioning depending on the study. The lack of consistent findings points to a need for further research; and (5) Limitations: Most often, authors noted that studies were too narrow in scope, involved a small sample size, or provided confounding factors. These limitations and other considerations led researchers to recommend investigating the characteristics of accommodations in further detail. Important overall observations from the NCEO analysis include a need in future research for a clear definition of the constructs tested, a reduction in confounding factors, increased study of institutional factors affecting accommodations judgment, and exploration of the desirability and perceived usefulness of accommodations by students themselves. Future research should focus on improvement in these areas but also on the positive effects of field-testing potential items in accommodated formats in addition to standard formats. The following are appended: (1) Summary of Research Purpose; (2) Summary of Type of Assessment; (3) Subject Area Studied (by Author); (4) Type of Accommodation Studied (by Author); (5) Summary of Participants; (6) Summary of Research Results; (7) Summary of Limitations Cited by Researchers; and (8) Summary of Suggestions for Future Research (as recommended by authors).

Component 1	Component 2	Component 3	Component 4	Component 5

				

Johnstone, C., Rogers, C., Wu, Y., Fedorchak, G. & Katz, M. (in press). Rules for Audio Representation of Science Items on a Statewide Assessment: Results of a Comparative Study. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Josephson, S. (2008). Keeping Your Readers' Eyes on the Screen: An Eye-Tracking Study Comparing Sans Serif and Serif Typefaces. *Visual Communication Quarterly 15*(1-2).

An exploratory eye-tracking study was designed to compare the onscreen legibility of sans serif and serif typefaces. The typefaces selected for this study included the serif font Times New Roman and the sans serif font Arial—both originally designed for printing on paper—and the serif font Georgia and the sans serif font Verdana—both designed in recent years especially for reading onscreen. Six participants read four short news stories, each displayed in a different typeface, while their eye-movement behavior was recorded. Overall, Verdana performed best. Participants were able to read more quickly and experienced fewer regressions (backward movements) when the type was set in Verdana. They also expressed a strong preference for this font on the computer screen.

Component 1	Component 2	Component 3	Component 4	Component 5
				

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

Component 1	Component 2	Component 3	Component 4	Component 5
				

Kane, M., & Mitchell, R. (1996). *Implementing performance assessment: Promises, problems, and challenges*. Mahwah, NJ: Lawrence Erlbaum Associates.

Commissioned by the Pelavin Research Institute of American Insitutes for Research, this edited book consisted of 10 chapters by some of the leaders researchers of the 1990's (e.g. Robert Linn, Eva Baker, Edward Haertal and Daniel and Lauren Resnick). Each chapter explores in-depth issues, characteristics, and strengths and weaknesses of performance assessments. The editors note in the Forward that this book assumes the reader will have some knowledge of performance assessment. The chapter titles include: Assessment Reform: Promises and Challenges, Performance Assessment and the Multiple Functions of Educational Measurement, Evaluating Progress with Alternative Assessments: A Model for Title 1, Extended Assessment Tasks: Purposes, Definitions, Scoring, and Accuracy, Linking Assessments, Examining the Costs, Conceptualizing the Costs of Large-scale Pupil Performance Assessment, Change has Changed: Implications for Implementation of Assessments

from the Organizational Change Literature, Arizona's Educational Reform: Creating and Capitalizing of the Conditions for Policy Development and Implementation, and Performance Assessment and Equity.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Katz, I. R. (2001). *Development of constructed-response engineering items using a rapid prototyping tool* (RR-01-02). Princeton, NJ: Educational Testing Service.

This report is a case study of the process of creating experimental items for a new Graduate Record Examinations (GRE) Engineering Test. Using a prototyping tool called the Free-Response Authoring and Delivery System (FRADS), a single test developer generated 55 computer-based test items that require either complex constructed responses or multiple-choice answers. Many of the items underwent significant evolution from their beginning as specifications written by engineering professors. Using FRADS, the test developer crafted new versions of items by exploring alternative response modes (constructed-response vs. multiple-choice, numeric input vs. icon movement, or other alternative tool choices) in addition to creating computer-based versions of the initially specified items. The creation of the constructed-response items is particularly noteworthy, because without the prototyping software, this project would likely have required programming staff. Instead, one test developer, working alone, was able to generate, evaluate, and modify these complex constructed-response items. Although presented in the context of a specific exam, the case study is relevant to other testing programs as a model of how new item types can be quickly and inexpensively explored using rapid prototyping tools.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Katz, I. R. (2005). Beyond technical competence: Literacy in information and communication technology. *Educational Technology Magazine, 45*(6), 44-47.

This article informs that there is growing consensus that, despite coming of age with the Internet and other technology, today's college students might not have the information and communication technology (ICT) literacy skills--the ability to effectively research and communicate using technology--necessary to navigate and make good use of the overabundance of information available today. The risks for college students who leave higher education without ICT literacy skills are substantial. More than in the past, skills in dealing with information via technology are part of what is needed to function in society. Information literacy includes many of the skills associated with conducting research and communicating information--in the past, what might have been termed "library skills"--but with the Internet, these skills are probably more often plied outside of the actual library building. ICT literacy is a specialization of information literacy, focusing on information competence as demonstrated through technology.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Katz, I.R. (2007). Testing information literacy in digital environments: The ETS iSKILLS ™assessment. *Information Technology and Libraries,* 26(3), 4-13.

Despite coming of age with the Internet and other technology, many college students lack the information and communication technology (ICT) literacy skills necessary to navigate, evaluate, and use the overabundance of information available today. This paper describes the development and early administrations of ETS's iSkills assessment, an Internet-based assessment of information literacy skills that arise in the context of technology. From the earliest stages to the present, the library community has been directly involved in the design, development, review, field trials, and administration to ensure the assessment and scores are valid, reliable, authentic, and useful.

Component 1	Component 2	Component 3	Component 4	Component 5
✓	✓			

Katz, I. R., Elliot, N., Attali, Y., Scharf, D., Powers, D., Huey, H., Joshi, K., & Briller, V. (2008). *The assessment of information literacy: A case study* (RR-08-33). Princeton, NJ: Educational Testing Service.

This study presents an investigation of information literacy as defined by the ETS iSkills™ assessment and by the New Jersey Institute of Technology (NJIT) Information Literacy Scale (ILS). As two related but distinct measures, both iSkills and the ILS were used with undergraduate students at NJIT during the spring 2006 semester. Undergraduate students (n = 331), first through senior years, took the iSkills and submitted portfolios to be judged by the ILS. First-year students took the Core iSkills assessment, which was designed to provide administrators and faculty with an understanding of the information and communication technology (ICT) literacy of a student doing entry-level coursework (n = 155). Upper classmen took the more difficult Advanced iSkills assessment, appropriate for rising juniors (n = 176). Across all class levels, iSkills scores varied as expected. First-year basic skills writing students performed at lower levels than first-year students enrolled in traditional composition and cultural history courses; seniors performed at higher levels than sophomores and juniors. Because the NJIT ILS scores were designed to be curriculum sensitive, portfolio scores did not similarly follow grade levels. Analyses revealed weak correlations between portfolio and Core iSkills scores and moderate correlations between portfolio and Advanced iSkills scores. As two associated yet distinct systems of inquiry designed to explore undergraduate student performance, the ETS iSkills assessment and the NJIT ILS—taken both individually and together—yield important information regarding student performance.

Component 1	Component 2	Component 3	Component 4	Component 5
	✓			

Katz, I. R., Martinez, M. E., Sheehan, K., & Tatsuoka, K. (1993). *Extending the rule space model to a semantically-rich domain: Diagnostic assessment in architecture* (RR-93-42-ONR). Princeton, NJ: Educational Testing Service.

This paper presents a technique for applying the Rule Space Model of cognitive diagnosis (Tatsuoka, 1983) to assessment in a semantically rich domain. Responses of 122 architects to 22 architecture test items developed to assess a range of architectural knowledge were analyzed using Rule Space. Verbal protocol analysis guided the construction of a model of examinee performance, consisting of processes for constructing an initial representation of an item (labeled "understand"), forming goals and performing actions based on those goals ("solve"), and determining whether goals have been

attempted and satisfied ("check"). Item attributes derived from these processes formed the basis for diagnosis. Successful diagnostic classifications were obtained for approximately 65%, 90%, and 40% of examinees based, respectively, on attributes associated with the "understand," "solve," and "check" processes of the problem-solving model. The findings support the effectiveness of Rule Space in a complex domain and suggest directions for developing new architecture items by using attributes particularly effective at distinguishing among examinees of different ability levels. Nine tables and three figures present study data.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Katz, I. R., & Smith-Macklin, A. S. (2007). Information and communication technology (ICT) literacy: Integration and assessment in higher education*. Systemics, Cybernetics and Informatics*. *5*(4).

Despite coming of age with the Internet and other technology, many college students lack the information and communication technology (ICT) literacy skills—locating, evaluating, and communicating information—necessary to navigate and use the overabundance of information available today. This paper presents a study of the validity of a simulations-based assessment of ICT literacy skills. Our overall goals for the assessment are to support ICT literacy instructional initiatives at colleges and universities.

Component 1	Component 2	Component 3	Component 4	Component 5
				

***Keehner, M., Hegarty, M., Cohen, C., Khooshabeh, P., Montello, D. R.* (2008). Spatial Reasoning With External Visualizations: What Matters Is What You See, Not Whether You Interact. *Cognitive Science, 32,* 1099-1132.**

Three experiments examined the effects of interactive visualizations and spatial abilities on a task requiring participants to infer and draw cross sections of a three-dimensional (3D) object. The experiments manipulated whether participants could interactively control a virtual 3D visualization of the object while performing the task, and compared participants who were allowed interactive control of the visualization to those who were not allowed control. In Experiment 1, interactivity produced better performance than passive viewing, but the advantage of interactivity disappeared in Experiment 2 when visual input for the two conditions in a yoked design was equalized. In Experiments 2 and 3, differences in how interactive participants manipulated the visualization were large and related to performance. In Experiment 3, non-interactive participants who watched optimal movements of the display performed as well as interactive participants who manipulated the visualization effectively and better than interactive participants who manipulated the visualization ineffectively. Spatial ability made an independent contribution to performance on the spatial reasoning task, but did not predict patterns of interactive behavior. These experiments indicate that providing participants with active control of a computer visualization does not necessarily enhance task performance, whereas seeing the most task-relevant information does, and this is true regardless of whether the task-relevant information is obtained actively or passively.

Component 1	Component 2	Component 3	Component 4	Component 5

				

Kenney, J. F. (1997). New testing methodologies for the Architect Registration Examination. *CLEAR Exam Review, 8*(2), 23-28.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Ketterlin-Geller, L.R., & Tindal, G. (2008). Embedded technology: Current and future practices for increasing accessibility for all students. *Journal of Special Education Technology, 22*(4), 1-15.

This article highlights the technological solutions available for increasing the accessibility of educational materials for students with disabilities. Special attention is paid to design features for supporting the needs of students with disabilities in computer-based tests; however, many of the same technologies can be applied to instructional environments. Current technologies such as assistive devices are discussed as well as emerging strategies for embedding flexibility into learning and testing interfaces. In addition, advances in the field of ambient intelligence are described as they might apply to educational environments.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Kettler, R., Rodriguez, M., Bolt, D., Elliott, S., Beddow, P., & Kurz, A. (2008). Modified multiple-choice items for alternate assessments: Reliability, difficulty, and the interaction paradigm. Unpublished manuscript. Peabody College of Vanderbilt University.

The final regulations for the *No Child Left Behind* act (U.S. Department of Education, 2007) indicate that a small group of students with disabilities may be counted as proficient through an alternate assessment based on modified achievement standards (AA-MAS). This new policy inspired the Consortium for Alternate Assessment Validity and Experimental Studies (CAAVES) project, a four-state collaboration with the goal of investigating item modification strategies for constructing future alternate assessments. An experimental design was used to determine whether tests composed of modified items would have the same level of reliability as tests composed of original items, and also help reduce the performance gap between students who would be eligible for an AA-MAS and students who would not be eligible. Three groups of eighth-grade students (N = 755) defined by eligibility and disability status took Original and Modified versions of reading and mathematics tests. In a third condition, the students were provided limited reading support. Changes in reliability across groups and conditions for both the reading and mathematics tests were determined to be minimal. Mean item difficulties within the Rasch model were shown to decrease more for students who would be eligible for the AA-MAS than for non-eligible groups, revealing evidence of an interaction paradigm. Exploratory analyses indicated that shortening the question stem may be a highly effective modification, and that adding graphics to reading items may be a poor modification.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Khattri, N., Reeve, A., & Kane, M. (1998). *Principles and practices of performance assessment*. Manwah, NJ: Lawrence Earlbaum Associates.

This book is based on the findings of a qualitative case study in which 16 schools were examined that were part of a district, state or national assessment reform movement. Chapters are organized around: history of assessment reform, characteristics of performance assessment, facilitators and barriers in assessment reform, teacher appropriation, and impact of PAs on student and teacher learning. The books' Appendices summarize research methods and school cases.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Khattri, N. & Sweet, D. (1996). Assessment Reform: Promises and Challenges. In Kane, M., & Mitchell, R. (1996). *Implementing performance assessment: Promises, problems, and challenges*. (pp. 1-21). Mahwah, NJ: Lawrence Erlbaum Associates.

In this book chapter, the authors highlight the history of performance assessment, promises and challenges, and gaps in knowledge of PA. They also comment on PA contextualized within organizational change, highlighting relations between PA and curriculum and assessment, professional development, and community support. It should be noted, within the chapter section titled, "Thrust for Reform" the authors argue that PAs will have a positive effect on students because of the nature of the real-world tasks and integration of multiple subject areas.

Component 1	Component 2	Component 3	Component 4	Component 5
				

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

Including English language learners (ELLs) in large-scale assessments raises questions about the validity of inferences based on their scores. Test accommodations for ELLs are intended to reduce the impact of limited English proficiency on the assessment of the target construct, most often mathematic or science proficiency. This meta-analysis synthesizes research on the effectiveness and validity of such accommodations for ELLs. Findings indicate that none of the seven accommodations studied threaten the validity of inferences. However, only one accommodation--providing English dictionaries or glossaries--has a statistically significant effect on ELLs' performance, and this effect equates to only a small reduction in the achievement score gap between ELLs and native English speakers. Findings suggest that accommodations to reduce the impact of limited language proficiency on academic skill assessment are not particularly effective. Given this, we posit a hypothesis about the necessary role of academic language skills in mathematics and science assessments.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Kim, D. & Huynh, H. (2008). Computer-Based and Paper-and-Pencil Administration Mode Effects on a Statewide End-of-Course English Test. *Educational and Psychological Measurement, 68*(4), 554-570.

The current study compared student performance between paper-and-pencil testing (PPT) and computer-based testing (CBT) on a large-scale statewide end-of-course English examination. Analyses were conducted at both the item and test levels. The overall results suggest that scores obtained from PPT and CBT were comparable. However, at the content domain level, a rather large difference in the reading comprehension section suggests that reading comprehension test may be more affected by the test administration mode. Results from the confirmatory factor analysis suggest that the administration mode did not alter the construct of the test.

Component 1	Component 2	Component 3	Component 4	Component 5
				

King, T. (in progress) *Examining the Accessibility of Grade 8 CBAL Math Items for Native Spanish-Speaking English Language Learners and Students with Motor Disabilities*. Princeton, NJ: Educational Testing Service.

Component 1	Component 2	Component 3	Component 4	Component 5
				

King, T. C., & Young, J. W. (2010). *How do students interpret linguistically modified test items? Cognitive interviews with Spanish-speaking ELs and English only students.* California Educational Research Association, San Diego, CA.

Component 1	Component 2	Component 3	Component 4	Component 5
				

King, T. C., and Young, J. W. (2011). *Examining the validity of linguistically modified items for English language learners through cognitive interviews*. American Educational Research Association, New Orleans, LA.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Kirsch, I. (2001). *The International Adult Literacy Survey: Understanding What Was Measured*.Princeton, NJ: Educational Testing Service.

This paper offers a framework adopted for developing the tasks used to measure adult literacy. The framework consists of six parts that represent a logical sequence of steps, from needing to define and represent a particular domain of interest, to identifying and operationalizing characteristics used to construct items, to providing an empirical basis for interpreting results. The various parts of the framework are seen as important in that they help to provide a deeper understanding of the construct of literacy and the various processes associated with it. A processing model is proposed

and variables associated with performance on the literacy tasks are identified as accounting for a large component of variance in task difficulty.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Kirsch, I. and P. B. Mosenthal. (1990). Exploring document literacy: Variables underlying the performance of young adults. *Reading Research Quarterly*, 25(1), 5-30.

The purpose of this study was to identify critical variables that underlie the performance of young adults on a diverse set of document literacy tasks. The authors addressed this question using the scores on document tasks from the Young Adult Literacy study conducted by NAEP. The 61 tasks on the document scale of the NAEP assessment were analyzed using a grammar developed by the authors. Based upon the analyses, the authors identified document variables (based on the structure and complexity of the document), task variables (based on the structural relation between the document and the accompanying question or directive), and process variables (based on strategies used to relate information in the question or directive to information in the document) that underlie the difficulty of the tasks for the total population and for major subgroups. The identification of these variables and the grammar of documents on which they are based provide a foundation for building an explanatory model that could systematically account for the constructs underlying document processing. Such a model could be used by test developers in constructing new tasks with predictable characteristics.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Klenowski, V. (1995). Student self-evaluation process in student-centered teaching and learning contexts of Australia and England. *Assessment in Education*, 2, 145–163.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Kopriva, R. (2010). Meaningfully Assessing English Learners in Local and Statewide Academic Assessments: What Does It Entail? Presentation at George Washington University, March 17, 2010.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Kopriva, R., & Carr, T. (2009). Building Comparable Computer-Based Science Items for English Learners: Results and Insights from the ONPAR Project. Presentation at the National Conference on Student Assessment, Los Angeles.

Component 1	Component 2	Component 3	Component 4	Component 5

				

Koretz, D. (1998). *Large-Scale portfolio assessments in the US: Evidence pertaining to the quality of measurement*. Assessment in Education: Principles, Policy, and Practice, 5(3), 309.

Portfolio assessment, that is, the evaluation of performance by means of a cumulative collection of student work, has figured prominently in recent US debate about education reform. Proponents hope not only to broaden measurement of performance, but also to use portfolio assessment to encourage improved instruction. Although portfolio assessment has sparked considerable attention and enthusiasm, it has been incorporated into only a few of the nearly ubiquitous large-scale external assessment programmes in the US. This paper evaluates the quality of the performance data produced by several large-scale portfolio efforts. Evaluations of reliability, which have focused primarily on the consistency of scoring, have yielded highly variable results. While high levels of consistency have been reached in some cases, scoring has been quite inconsistent in others, to the point of severely limiting the utility of scores.

Information about other aspects of validity is more limited and generally discouraging. For example, scores from portfolio assessments often do not show anticipated relationships with other achievement data, and teachers report practices in the implementation of portfolio assessment that are appropriate for instructional purposes but threaten the validity of inferences from portfolio scores. While other studies show positive effects of portfolio programmes (see Stecher, this issue), these findings suggest that portfolio assessment at its current state of development is problematic for many of the uses to which large-scale external assessments are now put in the US.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Koretz, D., Stecher, B., Klein, S., McCaffrey, D. (1994). *The Vermont portfolio assessment program: Findings and implications.* Educational Measurement: Issues and Practice, 13(3), 5-16.

How effective has the Vermont program been in meeting its goals? What goals appear to be in conflict in programs of this type? Why is there a need for “caution and moderate expectations”?

Component 1	Component 2	Component 3	Component 4	Component 5
				

Krapp, A., & Fink, B. (1992). The development and function of interests during the critical transition from home to preschool. In K. A. Renninger, S. Hidi & A. Krapp (Eds.), *The role of interest in learning and development,* (pp. 397-430). Hillsdale, NJ: Lawrence Erlbaum.

This chapter is an exploratory study based on individual case studies that examines the development of interests during the transition from home to pre-school. The findings reinforce the notion of the role of personal interest in managing critical transitions. Some theoretic considerations in the origins of personal interests and interests as a condition of development are discussed.

Component 1	Component 2	Component 3	Component 4	Component 5

				

Krygier, J., Reeves, C., Cupp, J., & DiBiase, D. (1997). *Multimedia in geographic education: Design, implementation, and evaluation*. Retrieved March 16, 2006, from http://horizon.unc.edu/projects/monograph/CD/Science Mathematics/Krygier.asp.

This paper describes an educational application of multimedia for geography and earth science education based on the assumption that multimedia is more than mere technology. This paper argues for an approach to educational multimedia design focused on a coherent set of multimedia design guidelines informed by a broad array of evaluation functions. Further, it is argued that such design and evaluation guidelines must be shaped by broader educational and content (geography and earth science) goals. It is suggested that this approach to the design, implementation, and evaluation of educational multimedia resources may guide other similar projects.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Lafontaine, D. and C. Monseur (2006), Impact of Test Characteristics on Gender Equity. *Indicators in the Assessment of Reading Comprehension*, University of Liège, Liège.

Downloaded at http://www.intestcom.org/Downloads/ITC2006Brussels/Session2.3.4lafontaineandmonseur.pdf on Jan 26 2012.

The hypothesis of an interaction between gender and item format in reading is well supported by findings in the literature. This presentation hypothesizes that there is a possible interaction between item format and the reading process. The impact of item format is larger for the aspect where students reflect on what they have read. However, the result needs to be treated cautiously due to the small number of items. The hypothesis of an interaction between type of text and item format is not supported by the data. The type of text has the more substantial impact on the gender gap (53 % of the variance), followed by the aspect (24 %) and the item format (16%). The author suggests that the large impact of the type of text can be related to reading practices where females read more fiction eg novels while males tend to read more non fiction, eg magazines.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Laitusis, C. C. (2008). State reading assessments and inclusion of students with dyslexia. *Perspectives on Language and Literacy, 33* (31-33). Available at http://www.interdys.org/ewebeditpro5/upload/Laitusis_from_Fall_08_United_Litho_Proof-6.pdf

In an effort to preserve the validity of the inferences made from scores on the third grade reading test, Maryland policy treats the responses of a student who has received a verbatim reading accommodation as “missing data,” thereby basing the MSA score on the student's responses on Standard Two (comprehension of informational text) and Standard Three (comprehension of literary text). While the use of a specific assistive technology product, such as Kurzweil 3000, may not be universally accepted for all test takers, this specification has two simultaneous advantages for students with dyslexia.\n In addition, students must receive accommodations in the classroom that

address their difficulties in word-level reading and all accommodations decisions must be documented in the student's IEP, 504 Plan, or official record..

Component 1	Component 2	Component 3	Component 4	Component 5
				

Laitusis, C. C. (2010). Examining the Impact of Audio Presentation on Tests of Reading Comprehension. *Applied Measurement in Education, 23 (2) 153-167*.

This study examined the impact of a read-aloud accommodation on standardized test scores of reading comprehension at grades 4 and 8. Under a repeated measures design, students with and without reading-based learning disabilities took both a standard administration and a read-aloud administration of a reading comprehension test. Results show that the mean score on the audio version was higher than scores on the standard version for both groups of students at both grade levels. Students with reading-based learning disabilities at both levels benefited differentially more than students with no disability. This finding continues to hold after controlling for reading fluency and ceiling effects at both grades. The results also examined the relationship between test scores and teachers' ratings of reading comprehension to determine which measures are the better predictors of teachers' ratings of reading comprehension by grade and disability classification.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Laitusis, C. C., & Cook, L. L. (Eds.) (2007). Large Scale Assessment and Accommodations: What Works? Washington, DC: Council for Exceptional Children

Component 1	Component 2	Component 3	Component 4	Component 5
				

Laitusis, C. C., & Cook, L. L. (2008). Reading aloud a test of reading comprehension. *ETS Research Spotlight*. Princeton, NJ: Educational Testing Service.

ETS is involved in research sponsored by educational institutions and government agencies. ETS researchers are selective about the projects they pursue to ensure that the work is in alignment with ETS's mission to "help advance quality and equity in education by providing fair and valid assessments and related services." The following article describes research supported in part by the U.S. Department of Education to help make large scale assessments of reading proficiency more accessible for students who have disabilities that affect reading. Through such efforts ETS is providing research-based principles and guidelines that help to further the field of education.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Laitusis, C. C, Maneckshana, B., & Monfils, L. (2009). Differential Item Functioning Comparisons on a Performance-Based Alternate Assessment for Students with Severe Cognitive Impairments, Autism and Orthopedic Impairments. Journal of Applied Testing Technology.

The purpose of this study was to examine Differential Item Functioning (DIF) by disability groups on an on-demand performance assessment for students with severe cognitive impairments. Researchers examined the presence of DIF for two comparisons. One comparison involved students with severe cognitive impairments who served as the reference group and students with autism and severe cognitive impairments who served as the focal group. The other comparison compared students with severe cognitive impairments (reference group) and students with severe cognitive impairments and orthopedic impairments (focal group). Results indicated a moderate amount of DIF for the autism comparison and a negligible amount of DIF for the orthopedic impairment comparison. In addition researchers coded all test items based on characteristics likely to favor one of the three groups. Although several of the hypothesized coding categories resulted in accurate prediction of DIF, the study was limited to items from one testing program for students in one state. More research is needed to see if these hypotheses can be replicated across testing programs and populations.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Laitusis, C. C., Morgan, D., Bridgeman, B., Zanna, J., & Stone, E. (2007). *Examination of Fatigue Effects from Extended Time Accommodations on the SAT I: Reasoning Test.* (College Board Report No. 2007–31). New York, NY: College Entrance Examination Board.

This study examined operational data from the SAT Reasoning Test™ to determine if students who tested under extended-time conditions were suffering from excessive fatigue effects relative to students who tested under standard-time conditions. Excessive fatigue was defined by significant (a) increases in differential item functioning (DIF) and (b) decreases in item completion rates, for items at the end of testing compared to the beginning of testing. Both of these factors were examined by comparing the performance of students who tested with extended time on items administered early (section position 2 or 3) and different items administered late (section position 8, 9, or 10) during the 10-section test administration. The sample included students with learning disabilities and/or Attention-Deficit Hyperactivity Disorder (ADHD) who tested with extended time (time and a half or double time) and students without disabilities who tested under standard-time conditions. Analyses were conducted on the critical reading and writing sections of the SAT® and examined item difficulty as well as item completion rates. Results indicated few changes in levels of DIF (early in the test compared to late in the test). In addition, item completion rates for students who received extra time were comparable to (or in some cases higher than) test-takers without disabilities who tested under standard time on both early and late sections.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Laitusis, C. C., Stone, E., Steinberg, J., & Cook, L. (in press). Technical Report.

Component 1	Component 2	Component 3	Component 4	Component 5

				

Laitusis, C. C., & Thurlow, M. (2008). Measuring literacy skills for students with visual impairments. Innovations Quarterly, 4(4).

Component 1	Component 2	Component 3	Component 4	Component 5
				

Landau, S., Russell, M., & Erin, J. (2006). The use of the Talking Tactile Tablet as a testing accommodation. *RE:view: Rehabilitation Education for Blindness and Visual Impairment, 38*(1), 7–21.

Over the past decade, testing has become an important component of education reform efforts. Currently, 49 states have formal programs that annually test students in all public schools. Although the subject areas and grade levels tested vary widely across states, the most recent federal education legislation requires all states to administer annual mathematics and language arts tests to students in Grades 3-8. Federal legislation also requires that education agencies, including local schools and state departments of education, provide accommodations for students with special needs so that tests measure the capabilities of disabled applicants and not their impairments. For students who are blind or have visual impairments, several types of accommodations may be provided, including Braille or large-print versions of the test, assistive devices, and test administrators who read items and responses aloud to the test taker. Earlier research in special education indicates that these accommodations have mixed results on the performance of students. This article describes a device, the "Talking Tactile Tablet" (TTT) that attempts to deliver a multi-sensory presentation of standardized test items to students who are blind or visually impaired. The TTT is a peripheral device that allows the user to interact with the computer through textured plastic sheets placed on top of the device. These sheets can contain graphics, buttons for a calculator, maps, diagrams, or other types of images. As the user presses on an object or segment of an object, information is transmitted to the computer as x, y coordinates. These coordinates are then used to perform various prescribed functions, such as making a menu selection, verbally identifying a diagram or portion of a diagram, or selecting a multiple-choice response. In the study reported, it was found that, although the TTT had no effect on student performance, the findings suggested four advantages to its use: (1) It may allow students to complete tests more quickly without diminishing their performances; (2) It provides students with more independence when performing a test that involves graphic elements; (3) It eliminates the opportunity for a test administrator to assist students inadvertently during testing; and (4) It increases the standardization of test delivery by presenting items in the same format, allowing students access to the same tools, and eliminating the need for a human proctor to inadvertently influence test performance.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Landau, S., Russell, M., Gourgey, K., Erin, J., & Cowan, J. (2003). Use of the Talking Tactile Tablet in mathematics testing. *Journal of Visual Impairment and Blindness, 97*(2), 85–96.

This article describes an experimental system for administering multiple-choice math tests to students who are visually impaired or have other print disabilities. Using a new audio-tactile

computer peripheral device called the Talking Tactile Tablet, a preliminary version of a self-voicing test was created for eight participants.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Lazarus, S.S., Thurlow, M.L., Christensen, L.L., & Cormier, D. (2007). States' alternate assessments based on modified achievement standards (AA-MAS) in 2007 (Synthesis Report 67). National Center on Educational Outcomes.

Federal legislation requires that all students, including students with disabilities, be included in all state and district-level accountability systems. Many students can take the regular assessment with or without accommodations, but some students with disabilities need alternate ways to access assessments. For the past several years, states have had alternate assessments based on alternate achievement standards. In April 2007, No Child Left Behind regulations were finalized that gave states the option to develop an alternate assessment based on modified achievement standards (AA-MAS). This assessment option is for a small group of students with disabilities who can make significant progress, but who may not reach grade-level achievement within the time period covered by their Individualized Education Program (IEP) (U.S. Department of Education, 2007). Prior to the finalization of this regulation a few states had developed, or were developing, an assessment they considered to be an AA-MAS--though none had yet been through the U.S. Department of Education's peer review process. This study compiles and summarizes publicly available information about these assessments. The purpose of this report is to provide a snapshot of the characteristics of the AA-MAS in these states at a time shortly after the April 2007 regulations were finalized. Because these states developed their assessments prior to the final regulations, some of the characteristics of these early AA-MAS may not fully comply with the regulations. Contents include: (1) Overview; (2) Process Used to Find Information about States' AA-MAS; (3) Results; (4) Eligibility Criteria; and (5) Discussion. Appended are state documents used in analysis and AA-MAS characteristics by state.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Linn, R. L. (1973). Fair test use in selection. ***Review of Educational Research.*** **43, 139-161.**

There is more than one reasonable definition of test fairness and these definitions are in conflict; alternate formulations are discussed here.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Linn, R.L., Baker, E.L. & Dunbar, S.B. (1991). Complex, Performance-Based Assessment: Expectations and Validation Criteria. Educational Researcher, 20(8), 15-21.

Increasing emphasis on assessment and concern about assessment techniques have stirred interest in alternative assessment forms, for which evidence is needed about consequences, transfer of performance on specific assessment tests, and assessment fairness. Criteria concerning

consequences, fairness, transfer-generalizability, cognitive complexity, content quality, content coverage, meaningfulness, and cost efficiency are presented.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Lissitz, R. W., & Samuelson, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36(8), 437-448.

This article raises a number of questions about the current unified theory of test validity that has construct validity at its center. The authors suggest a different way of conceptualizing the problem of establishing validity by considering whether the focus of the investigation of a test is internal to the test itself or focuses on constructs and relationships that are external to the test. They also consider whether the perspective on the test examination is theoretical or practical. The resulting taxonomy, encompassing both investigative focus and perspective, serves to organize a reconceptualization of the field of validity studies. The authors argue that this approach, together with changes in the rest of the terminology regarding validity, leads to a more understandable and usable model.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Liu, O. L., Schedl, M., Malloy, J., & Kong, N. (2009). *Does content knowledge affect TOEFL iBT reading performance? A confirmatory approach to differential item functioning (*RR-09-29). Princeton, NJ: Educational Testing Service.

The TOEFL iBT™ has increased the length of the reading passages in the reading section compared to the passages on the TOEFL® computer-based test (CBT) to better approximate academic reading in North American universities, resulting in a reduced number of passages in the reading test. A concern arising from this change is whether the decrease in topic variety increases the likelihood that an examinee's familiarity with the particular content of a given passage will influence the examinee's reading performance. This study investigated differential item functioning and differential bundle functioning for six TOEFL iBT reading passages, three involving physical science and three involving cultural topics. The majority of items displayed little or no differential item functioning (DIF). When all of the items in a passage were examined, none of the passages showed differential functioning at the passage level. Hypotheses are provided for the DIF occurrences. Implications for fairness issues in test development are also discussed.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1*(1), 95-100.

Two parallel forms of a broad-range tailored test of verbal ability have been built. The test is appro priate from fifth grade through graduate school. Simulated test administrations indicate that the 25-item tailored test is at least as good as a compar able 50-item conventional test. At most ability

levels, the tailored test measures much better. An offer is made to provide upon request item charac teristic curve parameters for 690 widely used Coop erative Test items, in order to facilitate research.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Lowe, R. K. (2004). Interrogation of a dynamic visualization during learning. *Learning and Instruction, 14*, 257–274.

Because animations can depict situational dynamics explicitly, they have the potential to help learners build coherent, high-quality mental models of complex change processes. Further, "interactive" animations provide opportunities for learners to deal with available information selectively and so avoid excessive processing demands. However, to be instructionally effective, the selected subsets of information must have high domain and task relevance. Approaches used by domain novices to interrogate an interactive animation of a complex dynamic system as they prepared for a subsequent prediction task were explored. Subjects searched the animation in order to learn generalizations upon which to base their predictions. Spatial and temporal strategies employed tended to be narrowly focused upon individual graphic features or localized groups while broader relational aspects required for coherence were neglected. The findings suggest that in order to build satisfactory mental representations from interactive animations, learners may require specific guidance regarding search strategies and targets.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Lumley, T. September 2010. *Reading literacy in the 21st century: Lessons from PISA: Electronic Reading Assessment (ERA).* Presentation given at GAVE. Lisbon, Portugal.

Retrieved from http://www.gave.min-edu.pt/np3content/?newsId=314&fileName=Lumley_Lecture_Part_2.PDF.pdf on Jan 26, 2012.

This presentation gives some background underpinning the introduction of an Electronic Reading Assessment as part of the PISA cycle. A definition of electronic reading is given plus key features that distinguish electronic reading from print reading. A model showing the relationship between the reading literacy dimensions of task, text and aspects in the electronic and print medium is developed. For PISA, electronic reading is conceptualized in terms of the components of text processing and navigation.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Mandinach, E. B., Bridgeman, B., Cahalan, C., & Trapani, C. (2005). *The impact of extended time on SAT I Reasoning test performance for students with and without learning disabilities.* (College Board Report 2005-8). New York, NY: College Entrance Examination Board.

Component 1	Component 2	Component 3	Component 4	Component 5

				

Mandinach, E. B., Bridgeman, B., Cahalan, C., & Trapani, C. (2005). *The impact of extended time on SAT test performance,* (ETS Research Report Series RR-05–20). Princeton, NJ: Educational Testing Service.

The effects of extended time on SAT Reasoning Test™ performance are examined. The study explored the impact of providing standard time, time and a half (1.5 time) with and without specified section breaks, and double time without specified sections breaks on the verbal and mathematics sections of the SAT®. Difference among ability, disability, and gender groups were examined. Results indicated that time and a half with separately timed sections benefits students with and without disabilities. Some extra time improves performance, but too much may be detrimental. Extra time benefits medium- and high-ability students but provides little or no advantage to low-ability students. The effects of extended time are more pronounced for the mathematics section of the SAT. The implications for potential changes to the SAT and the need for future research are discussed.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Martinez, M. E., & Katz, I. R. (1992). Cognitive processing requirements of constructed figural response and multiple-choice items in architecture assessment. *Educational Assessment, 3*(1), 83-98.

Contrasts between constructed-response items and multiple-choice counterparts have yielded but a few weak generalizations. Such contrasts typically have been based on the statistical properties of groups of items, an approach that masks differences in properties at the item level and may lead to inaccurate conclusions. In this article, we examine item-level differences between a certain type of constructed-response item (called figural response) and comparable multiple-choice items in the domain of architecture. Our data show that in comparing two item formats, item-level differences in difficulty correspond to differences in cognitive processing requirements and that relations between processing requirements and psychometric properties are systematic. These findings illuminate one aspect of construct validity that is frequently neglected in comparing item types, namely the cognitive demand of test items.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Martinez, M. M., Ferris, J. J., Kraft, W., Manning, W. H. (1990). *Automatic scoring of paper and pencil figural responses* (RR-90-23). Princeton, NJ: Educational Testing Service.

Large-scale testing is dominated by the multiple-choice question format. Widespread use of the format is due, in part, to the ease with which multiple-choice items can be scored automatically. This paper examines automatic scoring procedures for an alternative item type: figural response. Figural response items call for the completion or modification of figural material, including illustrations, diagrams, and graphs. Twenty-five science items were written in cooperation with the National Assessment of Educational Progress (NAEP) and printed with an ink that was invisible to scanning

equipment. The items were answered with pencils, response sheets were scanned, and the resulting data were processed by computer-based scoring algorithms. The paper describes the technology that leads to successful pilot scoring of seven items for ten subjects. Implications of this technology for the future of large-scale testing are discussed.

Component 1	Component 2	Component 3	Component 4	Component 5
	✓			

Matters, G., & Burnett, P. C. (1999). Multiple-choice versus short-response items: Differences in omit behaviour. *Australian Journal of Education*, 43(2), 117-128.

The overall rate of omission of items for a large sample of 17-year-old Australian students on a high-stakes test of achievement of cognitive skills of the high school curriculum is reported for a subtest in multiple-choice format and a subtest in short-response format. For the multiple item format, the omit rates were minuscule and there was no significant difference by gender or by type of school attended. For the short response format, the omit rates were between 10 and 20 times that for multiple-choice and the difference between male and female omit rates was significant as was the school difference. For both formats, females from single-sex schools omitted significantly fewer items than did females from coeducational schools.

Component 1	Component 2	Component 3	Component 4	Component 5
✓	✓	✓		

Mattson, D., Alcaya, C., & Russell, M. (2008). Accessible Portable Item Protocol. Enhanced Assessment Grant. Retrieved on 8/9/11 from: http://www2.ed.gov/programs/eag/awards08.html#mn.

Computer-based test delivery holds promise to increase the efficiency with which tests are administered and the speed with which results are returned to schools. Two challenges to computer-based delivery, however, are the provision of test accommodations and the ability to easily deliver test items across different delivery systems. The Accessible Portable Item Protocol (APIP) Project brings together a consortium of states (MN, FL, MD, MT, NH, SC, UT, & VT) to develop the capacity of all states to use a standard item mark up language for accessible computer-based test items. As a result of this project, the APIP will allow all states to ensure that our test items are accessible for students with a variety of needs and that our items are portable across computer-based delivery systems that apply the APIP standards.

Component 1	Component 2	Component 3	Component 4	Component 5
✓	✓	✓		✓

Mattson, D., & Russell, M. (2010). Interoperability and Accessibility Requirements by Adopting the Accessible Portable Item Profile (APIP) Standards, White Paper. Retrieved on 8/10/11 from: (http://education.state.mn.us/MDE/Account-ability_Programs/Assessment_and_Testing/APIP/index. html).

The APIP project is developing open-source, open-license interoperability standards that any assessment program, vendor, or other organization can freely adopt. If adopted by a consortium, the

APIP standards allow the consortium to provide assurances that items and assessment instruments developed with USDE funds will fulfill interoperability requirements. As a USDE federally-funded project (Grant #S368A090010), the APIP standards are intended to define a single standard for accessibility and interoperability of assessment items throughout the country. For a consortium that wishes to employ the APIP standards, the following text may be freely used to state how interoperability requirements will be fulfilled:

"To maximize the interoperability of assessments across technology platforms and allow assessment programs that adopt Common Assessments developed by this consortium to efficiently and cost-effectively change technology-based delivery platforms, we will adopt the Accessible Portable Item Profile (APIP) standards. The development of the APIP standards is supported by USDE funding to establish an industry format standard for highly accessible interoperable test items. The APIP standards incorporate key elements of established Question and Test Item specifications, Access for All specifications, and the National Instructional Materials Accessibility Standards to create a single standard for accessible item file format, accompanied by documentation of intended behaviors when the standardized APIP tagging structure is applied to test items."

Component 1	Component 2	Component 3	Component 4	Component 5
				

Mazzeo, J. & Harvey, A.L. (1998). The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature (College Board Rep. No. 88-8).

A literature review was conducted to determine the current state of knowledge concerning the effects of computer administration of standardized educational and psychological tests on the psychometric properties of these instruments. Students were grouped according to a number of factors relevant to the administration of tests by computer. Based on the studies reviewed, it seems that: (1) the rate at which examinees omit items in an automated test may differ from the rate at which they omit items in a conventional presentation; (2) scores on automated personality inventories are lower than scores obtained using the conventional testing format; (3) scores from automated versions of speed tests are not likely to be comparable with scores on paper-and-pencil versions; (4) presentation of graphics in an automated test may affect score equivalence; (5) tests containing items based on reading passages can become more difficult when presented via computer; and (6) the possibility of asymmetric practice effects may make it wise to avoid equating studies based on single-group counterbalanced designs.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice, 11*(1), 3-9, 20.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement.* Washington, DC: American Council on Education.

The major concern of validity, as of science more generally, is not to explain any single isolated event, behavior, or item response, because these almost certainly reflect a confounding of multiple determinants. Rather, the intent is to account for consistency in behaviors or item responses, which frequently reflects distinguishable determinants.

Component 1	Component 2	Component 3	Component 4	Component 5
✓	✓	✓	✓	✓

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

Authentic and direct assessment of performance and products are examined in light of contrasting functions and purposes with implications for validation, especially those of specialized validity criteria for performance assessment. The roles of positive and negative consequences of validation are underscored, along with the need for evidence of construct validity.

Component 1	Component 2	Component 3	Component 4	Component 5
		✓		

Miranda, H., Russell, M., & Hoffmann, T. (2004). *Examining the feasibility and effect of a computer based read aloud accommodation on mathematics test performance.* Chestnut Hill, MA: Technology and Assessment Study Collaborative, Boston College.

The purpose of the pilot study presented here was to compare student performance using a human reader to deliver the read aloud accommodation versus using computer-based text-to-speech technology. Specifically, the research questions for this study were:

1. How does computer-based digital speech compare with human readers as a means of providing the read-aloud accommodations?"
2. What is the effect of delivering the read aloud technology through CBT on students' test scores?
3. What is the effect of delivering the read-aloud accommodation using a human reader on students' test scores?
4. How is the accommodation effect related to students' computer skills, computer literacy, and computer use?

Results from this research will provide further evidence about the effect of computer delivery of the read-aloud accommodation on students' math performance compared with the effect of delivering the read aloud-accommodation using a human reader. This research is federally funded through the Enhancing State Assessment grant program and conducted collaboratively with Vermont, Rhode Island, New Hampshire, Maine, the Education Development Center (EDC) and CAST.

Component 1	Component 2	Component 3	Component 4	Component 5
✓	✓			✓

Miranda, H., Russell, M., Seeley, K., & Hoffmann, T. (2005). *Examining the feasibility and effect of a computer based read aloud accommodation on mathematics test performance.* Chestnut Hill, MA: Technology and Assessment Study Collaborative, Boston College.

The study presented here sought to address the need for further investigation into the effect of the verbal-response accommodation in testing situations for ELL students and students with disabilities students as well as to investigate the validity of accommodated scores by comparing accommodated and non-accommodated scores for ELL, students with disabilities, and general education (GE) students.

The purpose of this study was to examine the feasibility and effect of allowing students to provide verbal responses that are recorded using digital technologies. Specifically, the study examined the effects of allowing students to respond to items about a reading passage using: a) paper delivery and written response; b) computer delivery and keyboard response; c) computer delivery and verbal response recorded by the computer.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Mislevy, R., Almond, R., and Lukas, J. (2003). A Brief Introduction to Evidence-Centered Design (College Park, MD: College of Education at the University of Maryland, 2003), Downloaded on 8/9/11 at: www.education.umd.edu/EDMS/mislevy/papers/BriefIntroECD.pdf.

Evidence-centered assessment design (ECD) is an approach to constructing educational assessments in terms of evidentiary arguments. This paper provides an introduction to the basic ideas of ECD, including some of the terminology and models that have been developed to implement the approach. In particular, it presents the high-level models of the Conceptual Assessment Framework and the Four-Process Architecture for assessment delivery systems. Special attention is given to the roles of probability-based reasoning in accumulating evidence across task performances, in terms of belief about unobservable variables that characterize the knowledge, skills, and/or abilities of students. This is the role traditionally associated with psychometric models, such as those of item response theory and latent class models. To unify the ideas and to provide a foundation for extending probability-based reasoning in assessment applications more broadly, however, a more general expression in terms of graphical models is indicated. This brief overview of ECD provides the reader with a feel for where and how graphical models fit into the larger enterprise of educational and psychological assessment. A simple example based on familiar large-scale standardized tests such as the GRE is used to fix ideas. The document contains two appendices: (1) further reading about the ECD Project; and (2) a glossary of evidence-centered design terms.

Component 1	Component 2	Component 3	Component 4	Component 5
				

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

Evidence-centered assessment design (ECD) provides language, concepts, and knowledge representations for designing and delivering educational assessments, all organized around the evidentiary argument an assessment is meant to embody. This article describes ECD in terms of layers for analyzing domains, laying out arguments, creating schemas for operational elements such as tasks and measurement models, implementing the assessment, and carrying out the operational processes. We argue that this framework helps designers take advantage of developments from measurement, technology, cognitive psychology, and learning in the domains. Examples of ECD tools and applications are drawn from the Principled Assessment Design for Inquiry (PADI) project.

Attention is given to implications for large-scale tests such as state accountability measures, with a special eye for computer-based simulation tasks.

Component 1	Component 2	Component 3	Component 4	Component 5
✓	✓	✓		

Mislevy, R.J., Steinberg, L.S., & Almond, R.A. (2003). On the structure of educational assessments. Measurement: Interdisciplinary Research and Perspectives, 1, 3-67.

In educational assessment, educators observe what students say, do, or make in a few particular circumstances and attempt to infer what they know, can do, or have accomplished more generally. A web of inference connects the two. Some connections depend on theories and experience concerning the targeted knowledge in the domain, how it is acquired, and the circumstances under which people bring their knowledge to bear. Other connections may depend on statistical models and probability-based reasoning. Still others concern elements and processes involved in test construction, administration, scoring, and reporting. This paper describes a framework for assessment that makes explicit the interrelationships among substantive arguments, assessment designs, and operational processes. The work was motivated by the need to develop assessments that incorporate purposes, technologies, and psychological perspectives that are not well served by familiar forms of assessments. However, the framework is equally applicable to analyzing existing assessments or designing new assessments within familiar forms.

Component 1	Component 2	Component 3	Component 4	Component 5
✓	✓	✓	✓	✓

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*(3), 229-258.

In this theoretical paper, the problem identified is for validity researchers to find the appropriate criteria for a validity argument within the context of validity. In the first section of the paper, construct validity and assessment use are featured as two important concepts for validity-based arguments. Second, a synthesis is presented of the questions, criteria, and evidence used for assessment in general, and performance assessment. In the final section, an overview of concerns from interpretive researchers is summarized and Moss argues for a more "expansive" conception of validity theory that doesn't favor standardized assessments.

Within the section, "Components of Validity Inquiry for Performance Assessments in Particular," it is noted that validity researchers tend to rely on traditional sources of evidence (e.g. internal consistency and correlations to other assessments). Moss uses several seminal papers to argue these sources are not sufficient for performance assessment. Further sources of evidence include components of a testing system such as, a) exemplars and scoring processes, b) standards such as reliability and transparency, and c) criteria such as fairness and cognitive complexity (see pages 244-248 for additional references).

Component 1	Component 2	Component 3	Component 4	Component 5
		✓		

Moss, P. A., Girard, B. J., & Haniford, L. C. (2007). Validity in educational assessment. *Review of Research in Education*, 30, 109-162.

This article provides an overview of validity theories categorized in three waves from the early educational measurement theories, to the notion of validity as scientific inquiry. These validity theories placed primary emphasis on explaining the meaning of test scores by situating them in a larger theoretical framework. The latest wave sees validity theorized as practical argument where the focus of validity inquiry is directly on the purposes for which a test is used. These validity theories have made significant contributions to educational assessment in its evolution from traditional to digital media.

Component 1	Component 2	Component 3	Component 4	Component 5
				

National Center on Educational Outcomes. (2011). Developing common accommodations policies: Discussion points for consortia (NCEO Brief #2). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

As the Race-to-the-Top Assessment Consortia develop their assessments, they will need to develop shared accommodations policies to ensure that their tests are used in consistent ways across the participating states. Developing a common set of accommodations policies will require that Consortium members recognize the divergent viewpoints that currently exist in their states on the use and misuse of accommodations. This Brief addresses the need for the Consortia to develop shared accommodations policies. It presents information that shows the variability in use of accommodations and policies across the states within each Consortium. The Brief identifies ways to address the different perspectives on accommodations that underlie this variability, and provides several questions for the Consortia to use as discussion points as they develop their common accommodations policies.

Component 1	Component 2	Component 3	Component 4	Component 5
				

National Assessment Governing Board. (2009). *Reading Assessment and Item Specifications,* NAGB.

This document, the *Reading Assessment and Item Specifications for the 2009 National Assessment of Educational Progress*, provides information to guide passage selection, item development, and other aspects of test development. It accompanies the *Reading Framework for the 2009 National Assessment of Educational Progress,* which presents the conceptual base for the assessment.

Component 1	Component 2	Component 3	Component 4	Component 5
				

National Assessment Governing Board. (Sept 2010). *Reading Framework for 2011 National Assessment of Educational Progress.* U.S. Government Printing Office Washington, DC.

This framework provides a definition of reading as a complex process that involves understanding written text, developing and interpreting meaning and using meaning as appropriate to type of text, purpose, and situation. The cognitive skills articulated in the framework include locating/recalling,

integrating/interpreting, and critiquing/evaluating from both literary and information texts. To measure these cognitive skills, students respond to both multiple-choice and constructed-response items with varying distributions of question type by grade level. Students in grade 4 spend approximately half of the assessment time responding to multiple-choice items and half responding to constructed-response items. Students in grades 8 and 12 spend a greater amount of time on constructed-response items.

Component 1	Component 2	Component 3	Component 4	Component 5
✓		✓		

National Research Council. (2000). *How people learn: Brain, mind, experience, and school.* National Academy Press, Washington, DC.

How People Learn examines these findings and their implications for what we teach, how we teach it, and how we assess what our children learn. The book uses exemplary teaching to illustrate how approaches based on what we now know result in in-depth learning. This new knowledge calls into question concepts and practices firmly entrenched in our current education system. Topics include:

- How learning actually changes the physical structure of the brain.
- How existing knowledge affects what people notice and how they learn.
- What the thought processes of experts tell us about how to teach.
- The amazing learning potential of infants.
- The relationship of classroom learning and everyday settings of community and workplace.
- Learning needs and opportunities for teachers.
- A realistic look at the role of technology in education.

Component 1	Component 2	Component 3	Component 4	Component 5
✓	✓	✓		

National Research Council. (2004). *Keeping score for all: The effects of inclusion and accommodation policies on large-scale educational assessment.* 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. Washington, DC: National Academies Press.

U.S. public schools are responsible for educating large numbers of students with disabilities and English language learners - some 20 person of the nation's 46 million public school students fall into one or both of these categories. Both of these populations have been increasing, and the demand for evidence of their academic progress has also grown. In response to both changing public expectations and legal mandates, the federal government, states, and districts have attempted to include more such students in educational assessments.

Testing these two groups of students, however, poses particular challenges. Many of these students have attributes - such as physical, emotional, or learning disabilities or limited fluency in English - that may prevent them from readily demonstrating what they know or can do on a test. In order to allow these students to demonstrate their knowledge and skills, testing accommodations are used. For the purpose of this report, we have defined testing accommodations by drawing from the

definition in the AERA/APA/NCME Standards for Educational and Psychological Testing (American Educational Research Association et al., 1999). Our adapted definition is as follows: accommodation is used as the general term for any action taken in response to a determination that an individual's disability or level of English language development requires a departure from established testing protocol.

Component 1	Component 2	Component 3	Component 4	Component 5
				

New England Common Assessment Program Accommodations Guide. (2009). Retrieved on 10/5/11 from http://www.education.nh.gov/instruction/assessment/necap/admin/documents/accomm_guide09.pdf.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Nunnally, J. C. (1978). Psychometric theory. New York, NY: McGraw-Hill.

This book is a seminal text on psychometric theory which includes substantive findings in the area of measurement of psychological and educational variables, considers the broad measurement problems that arise in these areas and related methods of statistical analyses. It also combines classical procedures that explain variance with modern inferential procedures.

Component 1	Component 2	Component 3	Component 4	Component 5
				

OECD. (1999). *Measuring student knowledge and skills: A new framework for assessment*. Paris: OECD.

This report describes the conceptual framework on which the PISA 2000 assessment is based. It defines the domains of reading literacy, mathematical literacy and scientific literacy forming the core of PISA. It also describes the methods developed to ensure that the assessment tasks are valid across countries, are strong at measuring relevant skills and are based on authentic life situations.

Component 1	Component 2	Component 3	Component 4	Component 5
				

OECD. (2002). *Reading for change Performance and engagement across countries*. Paris: OECD.

This report takes a cross-national look at performance in reading. Students from more advantaged backgrounds perform better on average, but the gap varies greatly across countries. Female students perform better than male students in every country. But the most striking result reported here is the difference between students who are more engaged in reading and those who are less so. Those who express positive attitudes to reading, who read a variety of materials, and who spend time reading for pleasure, are on average much better readers. The analysis also indicates that reading engagement can to some extent compensate for the disadvantage in students' social

background. This result underlines the critical importance to school systems of developing curricula that will interest students as well as instruct them.

Component 1	Component 2	Component 3	Component 4	Component 5
				

OECD. (2005a). *Longer-term strategy of the development of PISA*. Paris: OECD.

The original plan for PISA was that, after the completion of the first three assessments in 2000, 2003 and 2006, the cycle would repeat itself with three-yearly assessments in the areas of reading, mathematics and science. However, in light of newly emerging policy priorities and the experience gained with PISA so far, the PISA Governing Board began in March 2004 to review the objectives and design of the PISA data strategy for the period 2009 and beyond. This paper sets out ideas for the structure of future PISA assessments including the assessment of the reading of electronic texts.

Component 1	Component 2	Component 3	Component 4	Component 5
				

OECD. (2005b). *Technical report for the OECD Programme for International Student Assessment 2003*. Paris: OECD Publications.

The PISA 2003 Technical Report describes the complex methodology underlying PISA 2003, along with additional features related to the implementation of the project at a level of detail that allows test developers to understand its processes of item construction, review, bias monitoring and field testing. It presents information on the sample design, methodologies used to analyze the data, technical features of the project and quality control mechanisms.

Component 1	Component 2	Component 3	Component 4	Component 5
				

OECD. (2007). *Assessing Scientific, Reading and Mathematical literacy. A Framework for PISA 2006*. Paris: OECD Publications.

This report presents the conceptual framework underlying the PISA 2006 survey. It includes a re-developed and expanded framework for scientific literacy, an innovative component on the assessment of students' attitudes to science and the frameworks for the assessment of reading and mathematics. Within each domain, the framework defines the content that students need to acquire, the processes that need to be performed, and the contexts in which knowledge and skills are applied. The domains and their aspects are also illustrated with sample tasks.

Component 1	Component 2	Component 3	Component 4	Component 5
				

OECD. (2009). *PISA 2009 Assessment Framework Key Competencies in Reading, Mathematics and Science*. Paris: OECD Publications.

The PISA 2009 Assessment Framework presents the theory behind the development of the 2009 survey. The re-worked and expanded framework for reading literacy includes not only the assessment of reading and understanding printed texts, but also an innovative component to assess how well students read, navigate and understand electronic texts. The framework also includes the basis for measuring mathematical and scientific competencies and presents the theory behind the questionnaires used to gather information from students, schools and parents on students' home backgrounds, students' approaches to learning and school learning environments.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Office of Educational Accountability. (2006). *Wisconsin's Model Academic Standards for Language Arts.* Wisconsin Department of Public Instruction.

This document describes the content and performance standards for Language Arts in the state of Wisconsin. The standards cover Reading/Literature, Writing, Oral Language, Language, Media and Technology and Research and Inquiry. Although the domain of language arts is divided into six sets of standards, these standards are seen to be connected. Additionally, most performance standards specified by grade level expect students to achieve a level of proficiency in more than one content standard.

Component 1	Component 2	Component 3	Component 4	Component 5
				

O'Hara, K. & Sellen, A. (1997). "A comparison of reading paper and online documents." *Conference on Human Factors in Computing Systems. Proceedings of the SIGCHI conference on Human factors in computing systems*, 335–342. Atlanta, Georgia. Retrieved October 10, 2008 from http://www.sigchi.org/chi97/proceedings/paper/koh.htm?searchterm=anticipatory.

We report on a laboratory study that compares reading from paper to reading on-line. Critical differences have to do with the major advantages paper offers in supporting annotation while reading, quick navigation, and flexibility of spatial layout. These, in turn, allow readers to deepen their understanding of the text, extract a sense of its structure, create a plan for writing, cross-refer to other documents, and interleave reading and writing. We discuss the design implications of these findings for the development of better reading technologies.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Olson, J. F., & Dirir, M. (2010). *Technical report for studies of the validity of test results for test accommodations. Establishing the validity of test accommodations and score interpretations for students with disabilities: A collaboration of state-based research.* Washington, DC: Council of Chief State School Officers.

Component 1	Component 2	Component 3	Component 4	Component 5

				

Oltman, P. (1994). *The effect of complexity of mouse manipulation on performance in computerized testing* (RR-94-22). Princeton, NJ: Educational Testing Service.

The purpose of the study was to explore the possibility that requirements for complex mouse manipulation in computer-based testing might make more of a negative impact on certain groups than on others. Very little evidence of such differential effects was observed in this study.

Component 1	Component 2	Component 3	Component 4	Component 5
				

O'Reilly, T. & Sheehan, K. (2008). *Cognitively Based Assessment of, for and as Learning: A 21st century approach for assessing reading competency* (RM-09-04). Princeton, NJ: Educational Testing Service.

This paper describes a new approach for assessing reading comprehension in an accountability setting, called Cognitively Based Assessment of, for, and as Learning (CBAL). The CBAL approach uses evidence centered design to develop a competency model that drives the development of summative, formative, and professional support aspects of the assessment. The assessment is designed to measure the complete range of the reading construct including the pre-requisite reading skills that are necessary to decode and recognize words, to the higher level critical thinking skills needed to ensure success in the 21st century. This paper presents a prototype assessment that illustrates critical features of the CBAL approach. Psychometric and concurrent validity analyses suggest that the CBAL test development approach is a promising method for measuring reading competency.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Passonneau, R., Hemat, L., Plante, J., & Sheehan, K. (2002). *Electronic sources as input to GRE reading comprehension item development: SourceFinder prototype evaluation* (RR-02-12). Princeton, NJ: Educational Testing Service.

This evaluation study compares the performance of a prototype tool called SourceFinder against the performance of highly trained human test developers. SourceFinder - a specialized search engine developed to locate source material for Graduate Record Examinations (GRE) reading comprehension passages - employs a variety of shallow linguistic features to model the search criteria employed by expert test developers, to automate the source selection process, and to reduce source-processing time. The current evaluation provides detailed information about the aspects of source variation that are not well modeled by the current prototype. Approaches for enhancing performance in identified areas are discussed. The present study also provides a more explicit description of the source selection task, and a rich data set for developing a less subjective, more explicit definition of the types of documents preferred by test developers.

Component 1	Component 2	Component 3	Component 4	Component 5

				

Pearlman, M., Berger, K., & Tyler, L. (1993). *An application of multimedia software to standardized testing in music* (RR-93-36). Princeton, NJ: Educational Testing Service.

The results of a project in which multimedia software and programming tools were used to construct a proto- type aural skills test for undergraduates in music are reported. The test was delivered and scored by the software program, which also recorded and stored test- taker's responses. All components of the test were computer-based and computer-delivered. The results of the iterative tryouts are reported; the test screens are presented; and directions for future research are suggested.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Pecheone, R., Kahl, S., Hamma, J., Jaquith, A. (2010). *Through a Looking*

Glass: Lessons Learned and Future Directions for Performance Assessment. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

This chapter begins with background and historical perspectives on assessment and how performance assessment came to be (e.g. Multiple-choice tests fell short of measuring student achievement in the areas of problem-solving, critical thinking and the other 21st Century Skills). The chapter also highlights states that have been involved in large-scale performance assessment initiatives (e.g. Kentucky, New York, New Jersey, and the New England states). Specific frameworks and examples of each state's past assessment system are provided. The next section of the chapter provides lessons learned from past performance assessment endeavors and suggestions for future development and implementation. Quality controls for taking performance assessment to scale such as examples of task quality and measurement quality are provided. The chapter concludes with examples of current performance assessment system projects and Appendices of PA component examples.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Pellegrino, J. W., and Quellmalz, E.S. (2010). Perspectives on the integration of technology and assessment. *Journal of Research on Technology in Education*, *43*, (2), 119–134.

This paper considers uses of technology in educational assessment from the perspective of innovation and support for teaching and learning. It examines assessment cases drawn from contexts that include large-scale testing programs as well as classroom-based programs, and attempts that have been made to harness the power of technology to provide rich, authentic tasks that elicit aspects of integrated knowledge, critical thinking, and problem solving. These aspects of cognition are seldom well addressed by traditional testing programs using paper and pencil or computer technologies. The paper also gives consideration to strategies for developing balanced, multilevel assessment systems that involve articulating relationships among curriculum-embedded, benchmark, and summative assessments that operate across classroom, district, state, national, and international levels. It discusses the multiple roles for technology in an assessment-based

information system in light of the decision support needed from the multiple actors who operate across levels of the education system. The paper concludes with a consideration of the current state of the field as well as the potential for technology to help launch a new era of integrated, learning centered assessment systems.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Pennock-Roman, M., & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. Educational Measurement: Issues and Practice, 30, 9–25.

The objective was to examine the impact of different types of accommodations on performance in content tests such as mathematics. The meta-analysis included 14 U.S. studies that randomly assigned school-aged English language learners (ELLs) to test accommodation versus control conditions or used repeated measures in counter-balanced order. Individual effect sizes (Glass's d) were calculated for 50 groups of ELLs and 32 groups of non-ELLs. Individual effect sizes for English language and native language accommodations were classified into groups according to type of accommodation and timing conditions. Means and standard errors were calculated for each category. The findings suggest that accommodations that require extra printed materials need generous time limits for both the accommodated and unaccommodated groups to ensure that they are effective, equivalent in scale to the original test, and therefore more valid owing to reduced construct-irrelevant variance. Computer-administered glossaries were effective even when time limits were restricted. Although the Plain English accommodation had very small average effect sizes, inspection of individual effect sizes suggests that it may be much more effective for ELLs at intermediate levels of English language proficiency. For Spanish-speaking students with low proficiency in English, the Spanish test version had the highest individual effect size (+1.45).

Component 1	Component 2	Component 3	Component 4	Component 5
				

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

Component 1	Component 2	Component 3	Component 4	Component 5
				

Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models for culture fair selection. ***Journal of Educational Measurement,*** **13, 3-29.**

Compares and evaluates models for bias in selection. Strategies are compared and evaluated as to their advantages and disadvantages in the areas of business and education. Some suggested formats for establishing culture fair selection are felt, by the authors, to be inadequate for their task and require a more complex analysis.

Component 1	Component 2	Component 3	Component 4	Component 5

				

Pitoniak, M., Young, J. W., Martiniello, M., King, T., Buteux, A., and Ginsburgh, M. (2009). *Guidelines for the assessment of English language learners.* Princeton, NJ: Educational Testing Service.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Pitoniak, M. J., & Royer, J. M. (2001). Testing accommodations for examinees with disabilities: A review of psychometric, legal, and social policy issues. *Review of Educational Research, 71*(1), 53–104.

Summarizes issues related to the provision of testing accommodations for examinees with disabling conditions, considering the history of such accommodations, the disabilities typically addressed, legal issues and precedents, and psychometric issues. Also considers social policy questions related to equity and fairness in testing.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Pommerich, M., & Burden, T. (2000). From Simulation to Application: Examinees React to Computerized Testing. Paper presented at the annual meeting of the National Council on Measurement in Education.

A small-scale study was conducted to compare test-taking strategies, problem-solving strategies, and general impressions about the test across computer and paper-and-pencil administration modes. Thirty-six examinees (high school students) participated in the study. Each examinee took a test in one of the content areas of English, Mathematics, Reading, and Science. In spite of the small sample, observations from the study highlight issues test developers might want to consider in determining how to present a test. Several factors were identified that might lead an examinee to respond to more than just item content when giving an answer: page and line breaks, passage and item layout features, highlighting, and item characteristics. Other factors include navigational features such as scrolling, item review, item preview, and omit capability. Examinee characteristics contributed to many of the observed mode effects, especially examinee carelessness. Care should be taken to ensure that the examinee is responding to item content only and not to inherent features associated with the test administration mode.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Powers, D. (1999). *Test anxiety and test performance: Comparing paper-based and computer adaptive versions of the GRE General Test,* (ETS Research Report Series RR-99–15). Princeton, NJ: Educational Testing Service.

Tests the hypothesis that the introduction of computer-adaptive testing may help to alleviate test anxiety and diminish the relationship between test anxiety and test performance. Compares a sample of Graduate Record Examinations (GRE) General Test takers who took the computer-adaptive version of the test with another sample who took the paper-based version. There was no support for the study's major hypothesis.

Component 1	Component 2	Component 3	Component 4	Component 5
✓	✓			

Powers, D. E., (1992). *Will They Think Less of My Handwritten Essay If Others Word Process Theirs? Effects on Essay Scores of Intermingling Handwritten and Word-Processed Essays.* The Praxis Series: Professional Assessments for Beginning Teachers™. Retrieved from EBSCOhost.

The effects on essay scores of intermingling handwritten and word-processed student essays were studied with 32 students who produced handwritten and word-processed essays. Essays were converted to the other format and rescored. Results reveal higher average scores for handwritten essays. Implications for scoring are considered.

Component 1	Component 2	Component 3	Component 4	Component 5
✓	✓	✓		

Powers, D. E., O'Neill, K. A. (1992). *Inexperienced and anxious computer users: Coping with a computer-administered test of academic skills* (RR-92-75). Princeton, NJ: Educational Testing Service.

The objective of this study was to assess the degree to which the mode of administration of the computer-based Academic Skills Assessments of The Praxis Series: Professional Assessments for Beginning Teachers contributes to performance differences among test takers. To make this determination, inexperienced or anxious computer users were recruited to take the assessments. The degree to which the test design and test familiarization procedures effectively minimized variation due to comfort and familiarity with computers was examined from three perspectives: (1) the extent to which the availability of a personal, information- providing test center supervisor influenced test performances, beyond the help provided by a computerized test familiarization tutorial, (2) the effect of within-test practice on later performance on a subsequent section of the test, and (3) the relationship of computer-based test performance to attitudes toward computers and experience in using them. The conclusion was that performance on the tests is not unduly affected by computer administration

Component 1	Component 2	Component 3	Component 4	Component 5
	✓			

Puhan, G., Boughton, K., & Kim, S. (2005). *Evaluating the comparability of paper and pencil and computerized versions of a large scale certification test* (RR-05-21). Princeton, NJ: Educational Testing Service.

The study evaluated the comparability of two versions of a teacher certification test: a paper-and-pencil test (PPT) and computer-based test (CBT). Standardized mean difference (SMD) and

differential item functioning (DIF) analyses were used as measures of comparability at the test and item levels, respectively. Results indicated that effect sizes derived from the SMD were small ($d < 0.20$) and not statistically significant ($p > 0.05$), suggesting no substantial difference between the two test versions. Moreover, DIF analysis revealed that reading and mathematics items were comparable for both versions. However, five writing items were flagged for DIF. Substantive reviews failed to identify format differences that could explain the performance differences, so the causes of DIF could not be identified.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Quellmalz, E., DeBarger, A., Haertel, G., Schank, P., Buckley, B., Gobert, J., Horwitz, P., and Ayala, C. (2008). *Exploring the role of technology-based simulation in science assessment: The Calipers Project.* In Assessing Science Learning: Perspectives for Research and Practice. NSTA, 2008.

Not in Eric

Component 1	Component 2	Component 3	Component 4	Component 5
				

Quellmalz, E. S. (1991). Developing criteria for performance assessments: The missing link. *Applied Measurement in Education, 4*(4), 319-331.

In this article, the author recommends criteria for evaluating performance assessment and argues for each criterion to be well established during the assessment development…not after the assessment has been implemented. Six characteristics for sound criteria presented are: significance, fidelity, generalizability, developmental appropriateness, accessibility, and utility. The article concludes with, "Tactics for Specifying Criteria." These include using exemplars to develop criteria, work on criteria development early, and frequent examination of criteria.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Race to the Top (RTTT) Assessment Program Application for New Grants. (2010). Comprehensive Assessment Systems CFDA Number: 84.395B. Submitted by Washington State on behalf of the SMARTER Balanced Assessment Consortium.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Ragosta, M., & Wendler, C. (1992). *Time Eligibility Issues and Comparable Time Limits for Disabled and Nondisabled SAT Examinees*, (ETS Research Report Series RR-92–35). New York, NY: Educational Testing Service.

Data from test administration timing records, the Scholastic Aptitude Test (SAT) history file, and a survey questionnaire were used to investigate the issues of comparable testing time and eligibility for special test accommodations for SAT examinees with disabilities. In 1986-87 and 1987-88, 17,632 special test administrations were given. Comparable testing time for disabled examinees was found to be between 1.5 and 2 times that of non-disabled examinees. Time limits in that range would assure that approximately equal percentages of disabled and non-disabled examinees would complete each SAT section. Additional time was found necessary for blind students using Braille or cassette versions of the test. Eligibility for special test administration is tied to the severity of disability and documentation of disability. Some levels of disability could be distinguished for those with sensory disabilities, but the severity of disability could not be defined for physically disabled or learning-disabled examinees. It was difficult to isolate the need for test accommodations based on school practices. Alternatives to the current eligibility policy are discussed, including a change to school-based criteria and the use of individualized testing programs, or a change to empirically derived testing times. Five graphs and 40 tables present study data. Thirteen references are included. Appendix A presents the student disability questionnaire, and Appendix B presents the draft application form for special test administration.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Ravitch, D. (2003). *The Language Police: How pressure Groups Restrict What Students Learn.* New York, NY: Knopf.

Textbook publishers and state education agencies have sought to root out racist, sexist, and elitist language in classroom and library materials. But according to Diane Ravitch, a leading historian of education, what began with the best of intentions has veered toward bizarre extremes. At a time when we celebrate and encourage diversity, young readers are fed bowdlerized texts, devoid of the references that give these works their meaning and vitality. With forceful arguments and sensible solutions for rescuing American education from the pressure groups that have made classrooms bland and uninspiring, *The Language Police* offers a powerful corrective to a cultural scandal.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Reed, L. (1993). Achieving the aims and purposes of schooling through authentic assessment. *Middle School Journal*, 25(2), 11-13.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Renninger, K. A. (1992). Individual interest and development: Implications for theory and practice. In K. A. Renninger, S. Hidi & A. Krapp (Eds.), *The role of interest in learning and development,* (pp. 361-396). Hillsdale, N J: Lawrence Erlbaum.

The author positions interest as a critical bridge between cognitive and affective issues in both learning and development. This chapter addresses how individual interest and interest inherent in stimuli (books, text, toys, etc.) across subjects affect cognitive performance. Interest influences

student performance on assigned tasks such as reading and problem solving. Interest has a positive influence on students' abilities to recall and comprehend sentences and texts, and on the level of difficulty of texts read and problems worked. Although the presence of an attention grabbing context will not automatically guarantee student learning, it can provide a forum for learning if the context which includes teaching style, resources and technical support, is adjusted to incorporate student interests.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Renninger, K. A. (1998). The roles of individual interest(s) and gender in learning: An overview of research on preschool and elementary school-aged children/students. In L. Hoffmann, A. Krapp, K. A. Renninger & J. Baumert (Eds.), ***Interest and Learning*****. Proceedings of the Seeon-Conference on Interest and Gender (pp. 165 -174). Kiel: IPN.**

Research on individual interest among preschool and elementary-school students has addressed the content of student interests and the role of interest in the ways students access, process, and complete tasks eg reading a text, solving a math word problem or molding shapes with play dough. This chapter addresses the role of interests and gender, in the ways young children, and elementary school students, learn and summarizes findings that have emerged across studies of interest and issues that remain open to further research.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Rieber, L. P., Tzeng, S.-C., & Tribble, K. (2004). Discovery learning, representation, and explanation within a computer-based simulation: Finding the right mix. ***Learning and Instruction, 14*****, 307–323.**

The purpose of this research was to explore how adult users interact and learn during an interactive computer-based simulation supplemented with brief multimedia explanations of the content. A total of 52 college students interacted with a computer-based simulation of Newton's laws of motion in which they had control over the motion of a simple screen object--an animated ball. Two simulation conditions were studied, each differing in how the feedback of the ball's speed, direction, and position was represented: graphical feedback consisted of animated graphics and textual feedback consisted of numeric displays. In addition, half of the participants were given simulations supplemented with brief multimedia explanations of the content modeled by the simulation in order to investigate how to promote referential processing, a key component of dual coding theory. Results showed significant differences for both the use of the explanations and simulations containing graphical feedback in helping participants gain both implicit and explicit understanding of the science principles.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Rivera, C., Acosta, B., & Willner, L. (2008). Guide for refining state assessment policies for accommodating English language learners. Arlington, VA: The George Washington University Center for Equity and Excellence in Education. Retrieved from http://ells.ceee.gwu.edu/.

The aim of this Guide is to support states in refining assessment policies so they are more responsive to the linguistic needs of ELLs. It is designed to help state education agencies build policies that coherently address ELLs, and that clearly distinguish the accommodation of ELLs from the accommodation of students with disabilities.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Rivera, C., Shafer Willner, L. & Acosta, B. (2008). Guide for refining state assessment policies for accommodating English Language Learners. Arlington, VA: The George Washington University Center for Equity and Excellence in Education.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Rivera, C., Shafer Willner, L., & Acosta, B. (2009). Improving the Selection of Accommodations for English Language Learners in Content Assessments, NCELA Newsletter. The George Washington University.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Rocky Mountain Braille Associates. (2011). *Design Principles for Tactile Graphics* http://www.tactilegraphics.org/.

This is a work-in-progress, providing some basic guidelines on designing and producing tactile graphics.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Rose, D. H., Meyer, A., Rappolt, G., & Strangman, N. M. (2002). Teaching Every Student in the Digital Age: Universal Design for Learning. Alexandria, VA: ASCD Press.

This guide to universal design in the classroom is divided into two sections. The first addresses the concept of universal design for learning (UDL); the second addresses the practical application of UDL in the classroom. Each chapter opens with a summary of key ideas and a graphic organizer that illustrates how the concepts fit together. The eight chapters address the following topics: (1) "Education in the Digital Age;" (2) "What Brain Research Tells Us about Learner Differences;" (3) "Why We Need Flexible Instructional Media;" (4) "What Is Universal Design for Learning?;" (5) "Using UDL To Set Clear Goals;" (6) "Using UDL To Support Every Student's Learning;" (7) "Using UDL To Accurately Assess Student Progress;" and (8) "Making Universal Design for Learning a Reality". An

appendix offers four classroom templates to help teachers apply the UDL framework. The templates address: a class learning profile, curriculum barriers, UDL solutions, and creating systematic change. Each template includes an introduction and three parts: an example of how the template might be used, collected sample items to use in the blank template, and a framework for applying UDL.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Routitsky, A., and Turner, R. (2003). *Item format types and their influences on cross-national comparisons of student performance*. Paper presented at the annual meeting of the American Educational Research Association (AERA).

The authors analyzed PISA 2003 field trial data using mathematics items for a population of 15 year-olds in 42 countries. They found mixed results regarding the interaction between item format and gender. Findings indicate that extended open-constructed response favor girls and short answer questions favor boys. However, as the item difficulty increases, both open constructed response and short answer items tend to favor boys. In addition, analyses conducted across all countries show that students of lower ability do better on the multiple-choice items than on both extended open constructed response and short answer items. This finding could explain why the interaction between gender and item format is nearly always observed in subject areas in which girls traditionally outperform boys (notably, reading and writing) and that results are less conclusive for subject areas in which boys traditionally outperform girls (mathematics and science).

Component 1	Component 2	Component 3	Component 4	Component 5
				

Russell, M. (1999). Testing writing on computers: A follow-up study comparing performance on computer and on paper. *Educational Policy Analysis Archives, 7*(20). Retrieved from http://epaa.asu.edu/epaa/v7n20/.

Russell and Haney (1997) reported that open-ended test items administered on paper may underestimate the achievement of students accustomed to writing on computers. This study builds on Russell and Haney's work by examining the effect of taking open-ended tests on computers and on paper for students with different levels of computer skill. Using items from the Massachusetts Comprehensive Assessment System (MCAS) and the National Assessment of Educational Progress (NAEP), this study focuses on language arts, science and math tests administered to eighth grade students. In addition, information on students' prior computer use and keyboarding speed was collected. Unlike the previous study that found large effects for open-ended writing and science items, this study reports mixed results. For the science test, performance on computers had a positive group effect. For the two language arts tests, an overall group effect was not found. However, for students whose keyboarding speed is at least 0.5 or one-half of a standard deviation above the mean, performing the language arts test on computer had a moderate positive effect. Conversely, for students whose keyboarding speed was 0.5 standard deviations below the mean, performing the tests on computer had a substantial negative effect. For the math test, performing the test on computer had an overall negative effect, but this effect became less pronounced as keyboarding speed increased. Implications are discussed in terms of testing policies and future research.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Russell, M. (2010). Accessible Test Design. Chapter in M. Russell & M. Kavanaugh (Eds), *Assessing Students in the Margins: Challenges, Strategies, and Techniques,* Information Age Publishing.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Russell, M. (2011). *Digital Test Delivery: Empowering-Accessible Test Design to Increase Test Validity for All Students.* Retrieved from http://www.acarseries.org/papers/Michael_Russell-Digital_Test_Delivery.pdf.

The development of digital test content, coupled with computer-based test delivery, provides an important opportunity to improve the accessibility of test items. By applying principles of accessible test design, the next-generation assessment systems will deliver more valid inferences about student learning based on test scores for all students. Rather than developing assessment content for the general population of students and then making post hoc changes to accommodate the needs of sub- groups of students, accessible test design provides a framework for making careful decisions about the methods used to tailor test administration to maximize the measurement of targeted constructs for each student. In turn, the Accessible Portable Item Profile (APIP) standards provide a tool for implementing accessible test design. The APIP standards empower next-generation assessments to solve three challenges. First, APIP provides a structure for specifying and storing the access needs of each student. Second, APIP provides a structure for augmenting item content with a variety of supplemental and alternate accessibility information designed to ensure that a test item functions properly for students with a variety of access needs. Third, APIP provides specifications for developing test delivery systems that can use a student access profile to tailor the provision of access tools (such as magnification, color contrast, masking) and the presentation of supplemental accessibility information (audio, Braille, tactile, or signed versions of item content). Collectively, the tools provided by APIP enable next-generation assessments to capitalize on the flexibility of digital technologies to maximize test validity for all students.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Educational Policy Analysis Archives, 5*(3). Retrieved from http://epaa.asu.edu/epaa/v5n3.html.

The effect that mode of administration, computer versus paper and pencil, had on the performance of 120 middle school students on multiple choice and written test questions was studied. Results show that, for students accustomed to writing on computers, responses written on the computer were more successful. Implications for testing are discussed.

Component 1	Component 2	Component 3	Component 4	Component 5

				

Russell, M., Higgins, J., & Hoffmann, T. (2009). Meeting the Needs of All Students: A Universal Design Approach to Computer-Based Testing. ***Innovate: Journal of Online Education, 5***(4).

Michael Russell, Thomas Hoffmann, and Jennifer Higgins describe how the principles of universal design were applied to the development of an innovative computer-based test delivery system, NimbleTools, to meet the accessibility and accommodation needs of students with a wide range of disabilities and special needs. Noting the movement to computer-based testing and the concerns facing assessment of special needs students, Russell, Hoffmann, and Higgins discuss how NimbleTools, with its multiple avenues for accommodations, provides a concrete example of how to improve access and achievement for students with special needs.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Russell, M., Hoffmann, T., & Higgins, J. (2009). NimbleTools: A universally designed test delivery system. ***Teaching Exceptional Children, 42***(2), 6–12.

Students with disabilities and special needs have faced challenges in accessing educational content, and in taking traditional pen-and-paper tests. How might technology improve the process, while making statewide tests truly accessible to all students? NimbleTools is the first computer-based test delivery system that incorporates principles of universal design by building in multiple accessibility tools designed to meet various needs of students. The program flexibly adjusts the accessibility tools available to students; students select when and how to use each tool while performing a given test item. Its success in two pilot studies suggests that computer-based test delivery can improve students' access to tests and their ability to demonstrate knowledge. The development of NimbleTools illustrates how the principle of universal design can be applied to ensure that all students have access to the tools they need to demonstrate what they actually know and can do.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Russell, M., Johnstone, C., Higgins,J., & Hoffmann, T. (2008). FCAT computer accommodations pilot study final report. Report prepared for the Florida Department of Education, Tallahassee, FL.

This report presents findings from a study that examined the use of NimbleTools to deliver the FCAT on computer with accommodations provided to students with disabilities and special needs. Commissioned by the state of Florida and funded through state legislative action, the study focused on the feasibility, usability, and effect of using NimbleTools to deliver multiple sections of a test composed of released FCAT items to students in grades 6 and 9.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Russell, M., Kavanaugh, M., Masters, J., Higgins, J., & Hoffmann, T. (2009). Computer-Based Signing Accommodations: Comparing a Recorded Human with an Avatar. *Journal of Applied Testing Technology, 10*(3). Retrieved from http://www.testpublishers.org/Documents/090727Russelletal.pdf.

Many students who are deaf or hard-of-hearing are eligible for a signing accommodation for state and other standardized tests. The signing accommodation, however, presents several challenges for testing programs that attempt to administer tests under standardized conditions. One potential solution for many of these challenges is the use of computer-based test delivery that integrates recordings of signed presentation of test content into the test. In addition to standardizing conditions, computer-based delivery holds potential to decrease the cost of developing recordings of signed presentation by using avatars rather than humans. However, because avatars are relatively new and are not as expressive or lifelike as humans, they may not be as affective as humans in presenting content in a clear and interpretable manner. The study presented here employed a randomized trial to compare the effect that a computer-based provision of the signed accommodation using a recorded human versus a signing avatar had on students' attitudes about performing a mathematics test and on their actual test performance. This study found that students generally reported that it was easy to perform a mathematics test on computer, and that both the recorded human and the signing avatar tools were easy to use and to understand. Students also reported a strong preference for performing future tests on computer, and generally preferred using the recorded human and the avatar for future tests rather than a DVD. While students also reported that they preferred the recorded human rather than the signing avatar, this preference did not affect test performance. The use of the recorded human and the avatar did not have affects on either the amount of time required to complete the test items or on students' performance on the test items. Implications for future research are discussed in light of these findings and the shortcomings of this study.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Russell, M., Mattson, D., Higgins, J., Hoffmann, T., Bebell, D., & Alcaya, C. (2011). *A Primer to the Accessible Portable Item Profile (APIP) Standards.* [White Paper]. Retrieved from http://apipstandard.org/archive/papers/APIP%20Primer%20-%20Final.pdf.

The APIP Standard project was initially developed by eight state testing programs and leaders in accessible test design to provide assessment programs and item developers a tool for standardizing the file format of digital test items. An important goal of APIP was to provide a tool that item developers can use to specify all the information and resources required to make a test item accessible for students with a variety of disabilities and special needs.

The APIP framework is comprised of two information models that allow test delivery engines to tailor the presentation of items to meet individual examinees' access needs. The first information model focuses on Examinee Access Information. During test delivery, the Examinee Access Information model performs two functions. First, the model provides information that allows a test delivery engine to activate specific tools that tailor the presentation of item content to the examinee. These embedded access features may include magnification, alternate contrast, increased white space and answer masking. Second, the Examinee Access Information model provides information that specifies which accessibility information embedded within the item model is pertinent to the examinee. APIP allows item developers to place a variety of types of access information within an item, including specifications for how an item is to be presented in auditory, Braille, sign, or tactile forms. In addition, the item information model allows an item developer to point to alternate versions

of the item that are presented in an alternate language (e.g., Spanish), in simplified English (e.g., with negatives removed), or in some other form.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Russell, M., & Plati, T. (2001). Effects of computer versus paper administration of a state-mandated writing assessment. *Teachers College Record.* Retrieved from http://www.tcrecord.org/Content.asp?ContentID=10709.

This study builds on two previous studies to examine the effect administering extended composition test items on paper and on computer has on student performance. This study employs writing items from the 1999 Massachusetts Comprehensive Assessment System (MCAS) to examine the mode of administration effect in grades eight and ten. Like the previous studies, this article reports that open-ended Language Arts items that require students to generate responses using paper and pencil severely underestimate the achievement of students accustomed to writing using a computer. Combining the effects found in this study with those found in a prior study, this article estimates that students accustomed to writing using a computer under-perform on the MCAS Language Arts test by four to eight points on an eighty point scale. This article concludes by recommending that state testing programs that employ open-ended items in Language Arts provide students with the option of composing responses on paper or on computer.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Russell, M., & Plati, T. (2002). Does it matter with which I write: Comparing performance on paper, computer and portable writing devices. *Current Issues in Education, 5*(4). Retrieved from http://cie.ed.asu.edu/volume5/number4/.

This study builds on three previous studies (Russell, 1999; Russell & Haney, 1997; Russell & Plati, 2001) to examine the effect of administering extended composition test items on paper, on computer, or on a portable writing device has on student performance. This study employs writing items from the 1999 Massachusetts Comprehensive Assessment System (MCAS) to examine the mode of administration effect in grades four and eight. Similar to previous studies, this article finds that open-ended Language Arts items that require students to generate responses using paper and pencil, severely underestimate the achievement of fourth grade students accustomed to writing using a computer. This study also finds that open-ended tests administered on paper underestimate the achievement of eighth grade students accustomed to writing with an eMate (a portable writing device). Combining the effects found in this study with those found in Russell's 1999 study, this article estimates that the MCAS Language Arts test underestimates the performance of students accustomed to writing using a computer by four to eight points on an eighty point scale. This article concludes by recommending that state testing programs that employ open-ended items in Language Arts provide students with the option of composing responses using the writing tools with which they are accustomed to working.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Sato, E., Rabinowitz, S., Gallagher, C., & Huang, C.W. (2010). Accommodations for English language learners: The Effect of Linguistic Modification of Math Test Item Sets. San Francisco, CA: WestEd.

This study examined the effect of linguistic modification on middle school students' ability to show what they know and can do on math assessments. To do so, two item sets with 25 multiple-choice items each were developed, one containing original math items and one containing these items with linguistic modifications. Items were selected from two content strands: (1) measurement and (2) number sense and operations. Efforts were made to ensure that both sets of math test items met stringent guidelines for grade and population appropriateness, content rigor, and standardized administration. In developing the two item sets, researchers solicited input from experts and collected data through cognitive interviews and pilot testing.

Component 1	Component 2	Component 3	Component 4	Component 5
✓				✓

Sawaki, S., & Nissan, S. (2009). *Criterion-related validity of the TOEFL iBT listening section* (RR-09-02). Princeton, NJ: Educational Testing Service.

The study investigated the criterion-related validity of the Test of English as a Foreign Language™ Internet-based test (TOEFL® iBT) Listening section by examining its relationship to a criterion measure designed to reflect language-use tasks that university students encounter in everyday academic life: listening to academic lectures. The design of the criterion measure was informed by students' responses to a survey on the frequency and importance of various classroom tasks that require academic listening, and the relationship of these tasks to successful course completion. The criterion measure consisted of three videotaped lectures (in physics, history, and psychology) and included tasks created by content experts who are former university professors of the relevant content area. These tasks reflected what the content experts expected students to have comprehended during the lecture. The criterion measure and the TOEFL iBT Listening section were administered to nonnative speakers of English who were enrolled in undergraduate and graduate programs. Data from 221 participants were analyzed. Substantial correlations were observed between the criterion measure and the TOEFL iBT Listening section score for the entire sample and for subgroups (Pearson correlation coefficients ranging from .56 to .74 and disattenuated correlations ranging from .62 to .82). Moreover, the analysis of the mean scores on the criterion measure for different ability groups indicated that participants who scored at or above typical cut scores for international student admission to academic programs (i.e., TOEFL iBT Listening section score of 14 or above) scored, on average, nearly 50% or more on the criterion measure, demonstrating reasonable comprehension of the academic lectures.

Component 1	Component 2	Component 3	Component 4	Component 5
	✓			

Scalise, K., and Gifford, B. (2006). A Framework for Constructing Intermediate Constraint Questions and Tasks for Technology Platforms. *Journal of Technology, Learning and Assessment*, *4*(6), 4-43.

This article examines a potential limitation for harnessing the benefits of computer-based assessment in large scale testing-the design of questions and tasks which computers can effectively score and report while still gathering meaningful measurement evidence. A taxonomy of 28 innovative item types is presented and organized by the degree of constraint on the respondent's

options for answering or interacting with the item or task. The 28 types of items are based on seven categories of ordering involving successively decreasing response constraints from fully selected to fully constructed items. The taxonomy provides a practical resource for item developers.

Component 1	Component 2	Component 3	Component 4	Component 5
✓	✓			

Schaeffer, G. A., Bridgeman, B., Golub-Smith, M. L., Lewis, C., Potenza, M. T., & Steffen, M. (1998). *Comparability of paper-and-pencil and computer adaptive test scores on the GRE General Test,* (ETS Research Report Series RR-98–38). Princeton, NJ: Educational Testing Service.

This report summarizes a study conducted to assess the comparability of paper-and-pencil (P&P) and computer adaptive test (CAT) scores on the GRE General Test. Volunteer examinees from around the country were randomly assigned to take the test in either CAT mode or P&P mode. Results indicated that for each measure (verbal, quantitative, analytical), mean scores on the CAT were higher than mean scores on P&P. It was hypothesized that CAT examinee test-taking behavior with regard to the CAT scoring method that was in place at the time of the study (allowing student to receive a score by answering only 80% or more of the questions) may have affected CAT scores. Investigation of this hypothesis indicated that CAT examinees who did not complete their CATs obtained higher mean sores than would be predicted. However, mean scores for examinees who completed their CATs were similar to mean scores for P&P examinees. A new psychometrically defensible CAT scoring method, called proportional scoring, was developed in which it is to the examinees' advantage, in terms of maximizing their scores, to carefully consider and answer as many items as they can. The proportional scoring method has been implemented in the operational GRE CAT program, and it is believed that this scoring method will result in CAT scores that are comparable to P&P scores. GRE CAT scores will continue to be closely monitored.

Component 1	Component 2	Component 3	Component 4	Component 5
✓	✓			

Shavelson, R., Baxter, G., Gao, X. (1993). Sampling Variability of Performance Assessments. Journal of Educational Measurement, *30*(3), 215-232.

Evidence is presented on the generalizability and convergent validity of performance assessments using data from six studies of student achievement that sampled a wide range of measurement facets and methods. Results at individual and school levels indicate that task-sampling variability is the major source of measurement error.

Component 1	Component 2	Component 3	Component 4	Component 5
		✓		

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education, 4*(4), 347-363.

Component 1	Component 2	Component 3	Component 4	Component 5

				

Shepard, L. A. (1991). Psychometricians' belief about learning. *Educational Researcher*, *20*(7), 2-16.

This article takes the perspective of the psychometrician in examining current views on learning. Based on a methodological approach and recent research on the interaction between teacher belief and classroom activities, the article argues that teachers bring their implicit theories about instruction and learning to their professional practice. This argument provides a strong foundation for the teacher belief that the constructed response item type allows for the production of higher order processing outputs.

Component 1	Component 2	Component 3	Component 4	Component 5
				

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

Formative assessment and instructional scaffolding are essentially the same thing. Formative assessment uses insights about a learner's current understandings to alter the course of instruction and thus support the development of greater competence. Scaffolding refers to supports that teachers provide the learner during problem solving--in the form of reminders, hints, and encouragement--to ensure successful completion of a task. Four strategies illustrate the strong connection between formative assessment and research on learning: eliciting prior knowledge, providing effective feedback, teaching for transfer of knowledge, and encouraging student self-assessment.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Shute, V. J., Graf, E. A., & Hansen, E. (2005). Designing adaptive, diagnostic math assessments for individuals with and without visual disabilities. In L. Pytlik Zillig, M. Bodvarsson & R. Bruning (Eds.), *Technology-based education: Bringing researchers and practitioners together* (pp. 169-202). Greenwich, CT: Information Age Publishing.

This report summarizes the design and development of an adaptive e-learning prototype for middle school mathematics for use with both sighted and visually disabled students. Adaptation refers to the system's ability to adjust itself to suit particular characteristics of the learner. The main parts of the report describe the system's theoretical foundation, architecture, models, and adaptive algorithm. We also review approaches for making assessment systems accessible to students with visual disabilities. Finally, we conclude with a summary of upcoming studies in relation to important research questions concerning micro- and macroadaptation. Using a design approach like the one described in this report may set a new precedent for environments that adapt to support student learning based on larger sets of incoming abilities and disabilities than have been considered previously.

Component 1	Component 2	Component 3	Component 4	Component 5

				

Singley, M. K., & Bennett, R. E. (1995). *Toward computer-based performance assessment in mathematics* (RR-95-34). Princeton, NJ: Educational Testing Service.

One of the main limitations of the current generation of computer-based tests is its dependency on the multiple-choice item. Our work is aimed at extending computer-based testing by bringing limited forms of performance assessment to it in the domain of mathematics. This endeavor involves not only building task types that better reflect valued problem solving, but also creating an integrated set of supports including easy-to-use interfaces, tutorials to teach novice computer users how to negotiate those interfaces, tools that help test developers create items quickly, and mechanisms for scoring constructed responses more efficiently.

Component 1	Component 2	Component 3	Component 4	Component 5
				

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.

Test accommodations are often given to students with disabilities as one means of removing construct-irrelevant barriers to proper measurement of their knowledge, skills, and abilities. However, the practice is controversial. This article reviews numerous studies that focused on the effects of accommodations on test performance. Consistent conclusions were not found across studies because of the wide variety of accommodations, the various ways in which they were implemented, and the heterogeneity of students to whom they were given. But two consistent findings emerged: Extended time tended to improve the performance of all students, although students with disabilities tended to exhibit relatively greater score gains; and oral accommodations on math tests were associated with increased test performance for some students with disabilities.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Slavin, R. E. (1996). Research on Cooperative Learning and Achievement: What we know, what we need to know. Contemporary Educational Psychology, *21*(0004), 43-69.

In this review of research on cooperative learning, Slavin highlights three theoretical perspectives: motivational, social cohesion, and cognitive. Each theory and corroborating empirical research is described within the context of cooperative learning. Some of the main findings include: each theory has differing views on whether or not extrinsic motivators should be given to student groups, all theories agree individual students should be held accountable with groups, and all theories agree group work (a.k.a. collaborative learning) is a positive instructional practice that increases student learning. Further, it's noted cognitive theorists argue that cooperative learning fits nicely within the Vygotskian and Piagetian theories of learning. Interactions amongst students in groups lead to an increase of higher level thinking skills, analysis skills, metacognition, and ultimately, student achievement.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Spencer, C. (2006). "Research on learners' preference for reading from printed text or from a computer screen." *Journal of Distance Education, 21*(1), 33-50.

In this study, 254 Royal Roads University School of Business learners (graduates and undergraduates) were surveyed on their online course-related reading habits and choices. Based on their responses and anecdotal comments and the data from follow-up interviews with six of the participants, learners preferred print copies of text materials for reasons of portability, dependability, flexibility, and ergonomics. Recommendations include providing an option in all online courses to print electronic text files in a format suitable for reading from paper. Further research is proposed on the effect of extended time spent in front of a computer screen on learners' preference for reading from paper.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Stiggins, R., & Chappuis, J. (2005). Using Student-Involved Classroom Assessment to Close Achievement Gaps. *Theory into Practice*, *44*(1), 11-18.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Stiggins, R. J. (1987). Design and development of performance assessments. *Educational Measurement: Issues and Practice*, *6*(3), 33-42.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Stocking, M. L., Smith, R., & Swanson, L. (2000). *An investigation of approaches to computerizing the GRE Subject Tests* (RR-00-04). Princeton, NJ: Educational Testing Service.

In this project, the authors explored the application of two common, computer-based test (CBT) designs, computerized adaptive testing (CAT) and linear-on-the-fly testing (LOFT) to two Graduate Record Examinations (GRE) Subject Tests chosen to represent the range of different test structures: the GRE Mathematics Test and the GRE Biology Test. The results were mixed. Some variations proved adequate in meeting minimum psychometric requirements. However, no variation proved adequate in the implementation of a complete test blueprint, nor did any variation provide strong support for meeting test security requirements in an environment of unrestricted continuous testing. The conversion of the GRE Subject Tests from paper-and- pencil administration to a computer-based testing format seems viable only if pretesting is implemented, item writing is expanded, test length is reduced, classification schemes are improved, and the roles of test specialists and the Committees of Examiners are substantially revised.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Stone, E., Cook, L., Laitusis, C. C., & Cline, F. (2010). Using Differential Item Functioning to Investigate the Impact of Testing Accommodations on an English Language Arts Assessment for Students Who are Blind or Visually Impaired. *Applied Measurement in Education, 23*(2), 132–152.

This validity study examined differential item functioning (DIF) results on large-scale state standards-based English-language arts assessments at grades 4 and 8 for students without disabilities taking the test under standard conditions and students who are blind or visually impaired taking the test with either a large print or braille form. Using the Mantel-Haenszel method, only one item at each grade was flagged as displaying large DIF, in each case favoring students without disabilities. Additional items were flagged as exhibiting intermediate DIF, with some items found to favor each group. A priori hypothesis coding and attempts to predict the effects of large print or braille accommodations on DIF were not found to have a relationship with the actual flagging of items, although some a posteriori explanations could be made. The results are seen as supporting the accessibility and validity of the current test for students who are blind or visually impaired while also identifying areas for improvement consisting mainly of attention to formatting and consistency.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Stout, W. (2002). Test Models for Traditional and Complex CTBs. In Mills, C., Potenza, M., Fremer, J., Ward, W. (Ed) *Computer-Based Testing*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

The evolution of assessment to a computer-based environment has prompted the input of measurement experts on whether the change in test medium has an influence on construct validity. In terms of test models the digital design offers potential for many innovations in item development through the use of sound and animation. The author cautions that any multi-media effects should not require extra mental or data processing on the part of the test-taker.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Stricker, L. J., & Rock, D. A. (2008). *Factor structure of the TOEFL internet-based test across subgroups* (RR-08-66). Princeton, NJ: Educational Testing Service.

This study assessed the invariance in the factor structure of the Test of English as a Foreign Language™ Internet-based test (TOEFL® iBT) across subgroups of test takers who differed in native language and exposure to the English language. The subgroups were defined by (a) Indo-European and Non-Indo-European language family, (b) Kachru's classification of outer and expanding circles of countries (based on prevalence of English use in educational and business contexts), and (c) years of classroom instruction in the English language. The same factor structure (four first-order factors corresponding to the test sections and a single higher-order factor encompassing these factors) was identified in each subgroup. The results support the present scoring scheme for the TOEFL iBT assessment and suggest that the test functions the same way for diverse subgroups of test takers.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Taylor, C., Jamiesen, J., Eignor, D., & Kirsch, I. (1998). *The relationship between computer familiarity and performance on the computer-based TOEFL test tasks* (RR-98-08) Princeton, NJ: Educational Testing Service.

The increasing use of computer-based testing raises concerns about equity and bias. Specifically, many in the field of language testing are concerned that the introduction of a computer-based TOEFL test in 1998 will confound language proficiency with computer proficiency and thus bring construct-irrelevant variance to the measurement of examinees' English-language abilities. In a Phase I study (Kirsch, Jamieson, Taylor, & Eignor, 1998), TOEFL examinees were surveyed regarding their computer familiarity and classified into one of three computer familiarity groups: low, moderate, and high. In this study, Phase II, more than 1,100 "low-computer-familiar" and "high-computer-familiar" examinees from 12 international sites were identified from the Phase I survey and administered a computer tutorial and a set of 60 computer-based TOEFL test items. The relationship between level of computer familiarity and performance on the computer-based items was then examined. The examinees in Phase II were largely representative of those in Phase I, who were representative of the general TOEFL test-taking population. The effect of computer familiarity after adjustments for language ability was examined by performing a series of analyses of covariance (ANCOVAs), using TOEFL paper-and-pencil test score as the covariate. These analyses were followed by a series of ANCOVAs involving the computer familiarity variable and a number of other variables: gender, reason for taking the TOEFL test, times the TOEFL test had been taken, and location where the TOEFL test was taken. After controlling for language ability, the researchers found no meaningful relationship between level of computer familiarity and level of performance on computerized language tasks among TOEFL examinees who had completed the computer tutorial.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Taylor, C., Kirsch, I, & Eignor, D. (1999). *Design and evaluation of a computer-based TOEFL tutorial* (RR-99-01). Princeton, NJ: Educational Testing Service.

In order to train examinees whose native language is not English to take a computerized TOEFL® test (Test of English as a Foreign Language™), a special set of tutorials was designed and developed. The TOEFL tutorials were field tested as a part of a computer familiarity study in 1996. This report describes the development of the tutorials. Also, the experiences of the 1,169 individuals who participated in the computer familiarity study are characterized in terms of timing and performance data, as well as self-reported attitudes. These analyses took into account computer familiarity and English ability, which both proved to be important in explaining some differences in time to complete the tutorials and perception of the tutorials' usefulness. Most examinees were successful and completing the practice items in the tutorials and thought that the tutorials were helpful. Some changes were subsequently made before operational implementation of the computerized TOEFL test in order to reduce the time needed to complete the TOEFL tutorials.

Component 1	Component 2	Component 3	Component 4	Component 5

				

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

The purpose of this paper is to explore the development of universal design and to consider its application to large scale assessments. Building on universal design principles presented by the Center for Universal Design, seven elements of universally designed assessments are identified and described in this paper.

With the shift to standards-based reform during the past decade, valid assessments for measuring the achievement of all students are essential. There is no longer an option for test developers to ignore the possibilities that universal design can bring to truly inclusive assessment systems. States that release requests for proposals for their state assessments have a similar obligation – to ensure that any proposal from test developers meets criteria that reflect the elements of universal design highlighted in this paper.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Thorndike, R. L. (1971). Concepts of culture fairness. *Journal of Educational Measurement,* 8, 63-70.

Fairness of a test relates to fair use. One definition of fair use states that a common qualifying score may be used with two groups if the regression line based on one group does not systematically over- or under-predict criterion performance in the other. However, it is shown that when the two groups differ appreciably in mean test score, the above procedure, which is "fair" to individual members of the group scoring lower on the test, is "unfair" to the lower group as a whole in the sense that the proportion qualified on the test will be smaller, relative to the higher-scoring group, than the proportion that will reach any specified level of criterion performance. An alternate definition would specify that the qualifying scores on a test should be set at levels that will qualify applicants in the two groups in proportion to the fraction of the two groups reaching a specified level of criterion performance.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Thurlow, M. L., Laitusis, C. C., Dillon, D. R., Cook, L. L., Moen, R. E., Abedi, J., & O'Brien, D. G. (2009). *Accessibility principles for reading assessments.* Minneapolis, MN: National Accessible Reading Assessment.

Within the context of standards-based educational systems, states are using large scale reading assessments to help ensure that all children have the opportunity to learn essential knowledge and skills. The challenge for developers of accessible reading assessments is to develop assessments that measure only those student characteristics that are essential parts of the reading proficiency the test intends to measure, and not those characteristics that could be related to the student's disability. The National Accessible Reading Assessment Projects (NARAP) have been conducting research to identify ways to increase the accessibility of reading assessments. This document is the

culmination of one of NARAP's goals: to develop evidence-based principles for making large scale assessments of reading proficiency more accessible for students who have disabilities that affect reading, while maintaining a high level of validity for all students taking the assessments. Some of the principles clarify and underscore the importance of well-accepted and widely used practices in designing reading assessments. Other principles have been developed from theory to respond to the needs of specific groups of students. Appendices include: (1) Support for Guidelines; (2) Principle 1; (3) Principle 2; (4) Principle 3; (5) Principle 4; and (6) Principle 5.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Thurlow, M. L., Quenemoen, R. F., Lazarus, S. S., Moen, R. E., Johnstone, C. J., Liu, K. K., Christensen, L. L., Albus, D. A., & Altman, J. (2008). *A principled approach to accountability assessments for students with disabilities* (Synthesis Report 70). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Building on research and practice, the National Center on Educational Outcomes (NCEO) has revisited and updated its 2001 document that identified principles and characteristics that underlie inclusive assessment and accountability systems. This report on a principled approach to accountability assessments for students with disabilities reflects lessons learned during the past seven years, presenting six core principles: (1) All students are included in ways that hold schools accountable for their learning; (2) Assessments allow all students to show their knowledge and skills on the same content; (3) High quality decision making determines how students participate; (4) Public reporting includes the assessment results of all students; (5) Accountability determinations are affected in the same way by all students; and (6) Continuous improvement, monitoring, and training ensure the quality of the overall system. Rationale and specific characteristics for each principle are included. A List of Groups and Individuals who Provided Review and Comment during Development of the 2008 NCEO Principles is appended.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Thurlow, M., Quenemoen, R., & Lazarus, C. (2011). Meeting the Needs of Special Education Students: Recommendations for the Race to the Top Consortia and States. Retrieved on 8/9/11 at: http://www.cehd.umn.edu/NCEO/OnlinePubs/Martha_Thurlow-Meeting_the_Needs_of_Special_Education_Students.pdf.

This paper identifies several actions for the Race to the Top assessment consortia to take to meet the needs of special education students. They are consistent with standards and principles for assessments, and they reflect evolving research and development activities directed toward supporting better assessments for every student.

1. Develop a set of common accommodation policies for the Race to the Top assessments
2. Follow accessibility principles in development, field testing, and implementation
3. Ensure that the design of computer-based tests is ap- propriate for special education students as well as other students
4. Develop formative and interim assessments to ensure inclusion of special education students in grade-level curricula focused on accelerated learning
5. Communicate and coordinate with the alternate assess- ment consortia

Component 1	Component 2	Component 3	Component 4	Component 5
				

Tindal, G. (2004). Large-scale testing of students with disabilities: Special Issue. *Exceptionality, 12*(2), 67-70.

Comments on the state policies affecting the large-scale testing of students with disabilities in the United States proposed by the National Center on Educational Outcomes (NCEO). Challenges in providing the right accommodations for examinees; Recognition of the possible setbacks in the modification of testing techniques; Explanation of the most efficient way of reporting outcomes.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Todt, E., & Schreiber, S. (1998). Development of interests. In L. Hoffmann, A. Krapp, K. A. Renninger & J. Baumert (Eds.), *Interest and learning. Proceedings of the Seeon-Conference on interest and gender,* (pp. 25-40). Kiel: IPN.

The authors present a heuristic model of interest development across the life-span. Their model is premised on the notion that interests reflect a person's developmental status. They hypothesized that topics of interest to children and students reflect the concerns associated with particular stages of development. They saw these topics of interest as a chronology that spanned the period from infancy through to 10 to 12 years of age. In the early years these topics ranged from interest in the structure of their physical and social environment to the gender appropriateness of their play and activities. In the later years interest developed in their competence or ability and expanded to include awareness of social relevance of possible and preferred engagements.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Tregidgo, A., & Plaute, J. (2004). *Listening communications skills assessment: The graphical user interface* (RM-04-02). Princeton, NJ: Educational Testing Service.

This ETS research memorandum documents the research and development work to build the graphical-user interface for the Listening Communications Skills Assessment for the Association of American Medical Colleges (AAMC). This new construct consists of vignettes that are in video format and delivered by computer. Usability studies were conducted to inform the interface design, establish the timing parameters, as well as type, functionality, and location of navigation tools, the Help button, volume control, replay function, and time display. The resulting interface has been approved by the AAMC for this new assessment. Recommendations for future work specific to the new assessment also are in this report.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Tucker, L. R. (1954). *Possible uses of electronic computers in mental testing* (RR-54-11). Princeton, NJ: Educational Testing Service.

This paper, which was presented by Leyard R Tucker at the International Congress of Pscyhology in Montreal Canada on June 8, 1954, looks at some of the ways in which electronic computers have impinged on mental testing activities.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Turnbull, W. W. (1968). Relevance in testing. *Science, 160*, 1424-1429.

Three stages mark the development of relevance in testing: relevance to the educational program, relevance to the individual student's past accomplishments, and relevance to the student's future accomplishments. If the College Entrance Examination Board should have to follow paths such as these, it has ahead of it the enormous job of transforming drastically the nature of its activities. Three possible stages are envisaged in the future development of the Board's tests: (1) for the immediate future, an extension of the recent trend toward the diversity of programs and of tests within programs (this being the stage of the multiplex external program); (2) a reduction of emphasis on external examinations and increased reliance on the record compiled by the student in his own school (this being the stage of the school-based program); and, (3) a system, eventually, of examinations in which each student is presented with the individual questions most pertinent to his past preparation and to his responses to test questions earlier in the sequence (this being the stage of the student-based program). It is held that the examination program should match both students and educational programs in diversity.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Turner, R., & Adams, R. J. (2007). The Programme for International Student Assessment: An Overview. *Journal of Applied Measurement*, *8*(3), 237-248.

This article provides a concentrated overview of the organizational and technical features of PISA. Organizational features cover the oversight role of the Secretariat within the OECD, the role of the national centre in each participating country and the role of the contractor. Technical features include test design and development, sampling, processes to ensure accurate translation and cultural appropriateness, field operations, quality monitoring, attribute scaling, analysis of data and the reporting of PISA outcomes.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Van der Linden W. (2002) On Complexity in CBT. In C Mills, Potenza, M., Fremer, J., Ward, W. (Ed) *Computer-Based Testing*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

In this chapter the author examines the issue of item complexity in the digital medium. From a measurement perspective he considers how the demands to produce assessment instruments that are authentic and engaging for students can impact on construct validity. He cautions that item

developers must be aware that any new variable added to the test that is not essential to the construct being tested can be a threat to construct validity.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Vendlinski, T., Stevens, R. (2002). Assessing Student Problem-Solving Skills with Complex Computer-Based Tasks. ***The Journal of Technology, Learning and Assessment, 1*****(3).**

Valid formative assessment is an essential element in improving both student learning and the professional development of educators. Various shortcomings in common assessment modalities, however, hinder our ability to make and evaluate such formative decisions. The diffusion of computer technology into American classrooms offers new opportunities to evaluate student learning and a rich, new source of data upon which to make inferences about the formative interventions that will improve learning. The path from data to inference, however, requires appropriate methodologies that can fully exploit the data without discarding or oversimplifying the behavioral complexity of student activity. This study used IMMEX™, a computerized simulation and problem-solving tool, along with artificial neural networks as pattern recognizers to identify the common types of strategies high school chemistry students used to solve qualitative chemistry problems. Then, based on the calculated probabilities that students would transition between these strategy types over time, Markov hidden chain analysis allowed us to develop a model of the capacity of the current curriculum to produce students able to apply chemistry content to a real-world problem.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Wade, S., Buxton, W., & Kelly, M. (1999). Using think-alouds to examine reader-text interest. ***Reading Research Quarterly*****,** ***34*****(2), 194-213.**

Using multiple measures, this investigation examined what text characteristics readers considered interesting and uninteresting. Based on think-alouds and postreading verbal reports, five text characteristics were most associated with interest: (a) information that was important, new, and valued; (b) information that was unexpected; (c) connections readers made between the text and their prior knowledge or experience; (d) imagery and descriptive language; and (e) author connections (e.g., comparisons and analogies). Characteristics that made the texts uninteresting involved problems related to comprehension--specifically, lack of adequate explanation and background information, difficult vocabulary, and lack of coherence. In addition, lack of credibility interfered with the value of the new information and theories presented in the texts.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K–12 mathematics tests. ***Educational and Psychological Measurement, 67*****, 219–238.**

This study conducted a meta-analysis of computer-based and paper-and-pencil administration mode effects on K-12 student mathematics tests. Both initial and final results based on fixed- and random-effects models are presented. The results based on the final selected studies with homogeneous effect sizes show that the administration mode had no statistically significant effect on K-12 student mathematics tests. Only the moderator variable of computer delivery algorithm contributed to predicting the effect size. The differences in scores between test modes were larger for linear tests than for adaptive tests. However, such variables as study design, grade level, sample size, type of test, computer delivery method, and computer practice did not lead to differences in student mathematics scores between computer-based and paper-and-pencil modes.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Washington State. (2010). Race to the Top Assessment Program Application for New Grants, Comprehensive Assessment Systems, CFDA Number 84.395B. Downloaded on 8/9/11 from: http://www.k12.wa.us/SMARTER/pubdocs/SBAC_Narrative.pdf.

Component 1	Component 2	Component 3	Component 4	Component 5
				

WestEd. (2010). *Simulations signal a new era in science assessment*. R&D Alert, *11*(2).

Component 1	Component 2	Component 3	Component 4	Component 5
				

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. Phi Delta Kappan, 70, 703-713.

Wiggins was an early and persuasive voice in the call for educators to move from standardized tests based on multiple choice items to testing regimes that are based on authentic assessments. He was an early proponent for the incorporation of performance tasks based on sophisticated criteria as a fundamental part of an authentic assessment. He acknowledged that using authentic standards and tests to judge educational achievement is labor-intensive and time-consuming, but provides students with a richer and more satisfying environment to demonstrate their learning.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Wiggins, G. (1998). Educative assessment: Designing assessments to inform and improve. San Francisco, CA: Jossey-Bass.

In this article, Wiggins argues for more authentic tasks as part of an assessment overhaul. He critiques norm-referenced testing and "shallow" thinking items and recommends authentic tasks be part of the national assessment agenda. Wiggins stresses the important of performance tasks (authentic tasks) consisting of extended responses that allow students to think critically, problem

solve, and the like. He compares these types of tasks to graduate student oral exams – "the student is given ample opportunity to explain his or her work…" (p. 712)

Component 1	Component 2	Component 3	Component 4	Component 5
				

Williams, J. D., & Jacobsen, S. (1990). Growth in mathematics skills during the intermediate years: Sex differences and school effects. *International Journal of Educational Research*, *14*, 157-174.

This article describes a cross-sectional study investigating the relationship between affective dispositions, cognitive proficiencies and gender on attitudes towards school-related activities and academic achievement. The findings indicate that in the domain of mathematics gender differences are mostly very small at the primary school level and in early secondary years, but that they tend to increase as students reach the end of high school. Therefore, test developers need to be aware of gender disparities in affective responses and academic achievement when selecting stimulus materials and designing prompts.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Williamson, D.M., Bauer, M., Steinberg, L.S., Mislevy, R.J., & Behrens, J.T. (2004). Design rationale for a complex performance assessment. *International Journal of Testing, 4*, 303–332.

In computer-based interactive environments meant to support learning, students must bring a wide range of relevant knowledge, skills, and abilities to bear jointly as they solve meaningful problems in a learning domain. To function effectively as an assessment, a computer system must additionally be able to evoke and interpret observable evidence about targeted knowledge in a manner that is principled, defensible, and suited to the purpose at hand (e.g., licensure, achievement testing, coached practice). This article describes the foundations for the design of an interactive computer-based assessment of design, implementation, and troubleshooting in the domain of computer networking. The application is a prototype for assessing these skills as part of an instructional program, as interim practice tests and as chapter or end-of-course assessments. An Evidence Centered Design (ECD) framework was used to guide the work. An important part of this work is a cognitive task analysis designed (a) to tap the knowledge computer network specialists and students use when they design and troubleshoot networks and (b) to elicit behaviors that manifest this knowledge. After summarizing its results, we discuss implications of this analysis, as well as information gathered through other methods of domain analysis, for designing psychometric models, automated scoring algorithms, and task frameworks and for the capabilities required for the delivery of this example of a complex computer-based interactive assessment.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Williamson, G. L. (2008). A Text Readability Continuum for Postsecondary Readiness. *Journal of Advanced Academics, 19*(4), 602-632.

This study investigates the gap between high school textbooks and various reading materials in the most widely chosen postsecondary domains of endeavor--the university, workplace, military and citizenship. Based on a large corpus of texts, significant readability gaps were found between high school texts and texts associated with citizenship, workplace, community college, GRE, and University text measures. This study provides sound evidence for the expectation that students need to engage with complex texts during middle and high school.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Willingham, W., Ragosta, M., Bennett, R., Braun, H., Rock, D., & Powers, D. (1988). *Testing handicapped people*. Boston: Allyn and Bacon.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Wise, L. L. (2010). Accessible reading assessments for students with disabilities: Summary and conclusions. *Applied Measurement in Education, 23*, 209–214.

The articles in this special issue make two important contributions to our understanding of the impact of accommodations on test score validity. First, they illustrate a variety of methods for collection and rigorous analyses of empirical data that can supplant expert judgment of the impact of accommodations. These methods range from internal analyses of reliability, differential item functioning, and factor structure to different ways for comparing score means from accommodated and non-accommodated assessments and ways of incorporating external criteria in assessing changes in validity. Equally important, these articles add significantly to our knowledge of the impact of specific reading accommodations for varying groups of students with disabilities. The accommodations reported included large print and Braille tests for blind and visually impaired students, segmented text for students with specific learning disabilities, and audio presentation of text for students with a variety of disabilities. The results reported are mixed, indicating some support for the validity of scores from accommodated assessments and also raising some questions for further exploration. Further meta-analytic studies are needed to sort out often conflicting results within and across different studies of reading accommodations.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Wolf, M. K., & Martiniello, M. (2010). Validity and fairness assessments for ELLs: The issue of language demands in content analysis. *AcceELLerate!, 2*(4), 9–10.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Wolf, M. K., Kao, J. C., Rivera, N. M, & Chang, S. M. (2012). Accommodation practices for English Language Learners in states' mathematics assessments. *Teachers College Record, 114*(3).

This article investigates the accommodation policies and practices for English language learners on large-scale, standards-based mathematics assessments.

Component 1	Component 2	Component 3	Component 4	Component 5
✓				✓

Wood, G. H., Darling-Hammond, L., Neill, M., & Roschewski, P. (2007). *Refocusing accountability: Using local performance assessments to enhance teaching and learning for higher order skills*. Briefing paper prepared for members of the Congress of the United States. Stewart, OH: Forum for Education and Democracy.

In this briefing paper prepared for the US Congress, the authors summarize: the definition of performance assessment, the benefits, and state and international programs with PA-based assessment systems. Recommendations are also given for using PAs within a national context. Although this paper predominantly discusses local performance assessment and systems, it does describe attributes of PAs that are generalizable to a national system: PAs measures higher order thinking skills, analysis and synthesis and tasks reflect real-world contexts. The authors also note that the attributes associated with performance assessment lead to student engagement and motivation.

Component 1	Component 2	Component 3	Component 4	Component 5
		✓		

Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). *Automated scoring of spontaneous speech using SpeechRater v1.0* (RR-08-62). Princeton, NJ: Educational Testing Service.

This report presents the results of a research and development effort for SpeechRaterSM Version 1.0 (v1.0), an automated scoring system for the spontaneous speech of English language learners used operationally in the Test of English as a Foreign Language™ (TOEFL®) Practice Online assessment (TPO). The report includes a summary of the validity considerations and analyses that drive both the development and the evaluation of the quality of automated scoring. These considerations include perspectives on the construct of interest, the context of use, and the empirical performance of the SpeechRater in relation to both the human scores and the intended use of the scores. The outcomes of this work have implications for short- and long-term goals for iterative improvements to SpeechRater scoring.

Component 1	Component 2	Component 3	Component 4	Component 5
	✓			

Young, J. W. (2008). Ensuring valid content tests for English Language Learners. *R&D Connections, 8*, 1–7.

Component 1	Component 2	Component 3	Component 4	Component 5

				

Young, J. W. (2009a). A Framework for Test Validity Research on Content Assessments Taken by English Language Learners. *Educational Assessment, 14*(3–4), 122–138.

In this article, I specify a conceptual framework for test validity research on content assessments taken by English language learners (ELLs) in U.S. schools in grades K-12. This framework is modeled after one previously delineated by Willingham et al. (1988), which was developed to guide research on students with disabilities. In this framework for research on ELLs, there are eight indicators of test comparability. Five of these indicators are measures of score comparability, while three indicators are measures of task comparability. To date, research has been conducted on six of the indicators of test comparability for content assessments taken by ELLs. For these indicators, findings from representative studies are summarized. For the remaining two indicators of test comparability for which no published research currently exists, I describe the types of research studies that are necessary to gather evidence to evaluate the comparability of content assessments for ELLs.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Young, J. W. (2009b). *Ensuring valid and fair content assessments for English language learners.* Presentation given at the 24th annual Texas Assessment Conference, Austin, TX.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Young, J. W., Cho, Y., Ling, G., Cline, F., Steinberg, J., & Stone, E. (2008). Validity and Fairness of State Standards-Based Assessments for English Language Learners. *Educational Assessment, 13*(2–3), 170–192.

English language learners (ELLs) constitute one of the fastest growing subpopulations of students in the United States. It is important to determine whether the assessments used by states in determining students' proficiencies are valid and fair for ELLs. This study focused on several standards-based assessments in mathematics and science administered to 5th and 8th graders. In assessing construct validity, all of the assessments were found to be essentially unidimensional in their underlying factor structure for native English speakers and for several ELL examinee groups. The use of translation glossaries/word lists as a testing accommodation was effective for supporting the unidimensionality of some of these assessments, specifically for the one at the 8th-grade level.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Young, J. W., Holtzman, S. B., & Steinberg, J. (2011). *Score comparability for language minority students on the content assessments used by two states,* (ETS Research Report Series RR-11–27). Princeton, NJ: Educational Testing Service.

In this research investigation of score comparability for language minority students (English language learners [ELLs] and former English language learners), we examined 3 indicators of score comparability (reliability, internal test structure, and differential item functioning) for 4th and 8th grade students who took the NCLB-mandated content assessments in English-language arts and mathematics in 2 different U. S. states. Overall, for the 8 assessments we examined, a high degree of score comparability was found for ELLs and former ELLs, when compared with native English speakers. The results from this study showed that although the assessments from the 2 states differed somewhat with respect to the 3 indicators, a high degree of score comparability was found for both states' content assessments.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Young, J. W. & King, T. C. (2008a). Testing accommodations for English language learners: A review of state and district policies (College Board Research Report No. 2008-6 and ETS Research Report RR-08-48). New York, NY: College Entrance Examination Board.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Young, J. W. & King, T. C. (2008b). *Testing accommodations for English Language Learners: A review of state and district policies,* (ETS Research Report Series RR-08–48). Princeton, NJ: Educational Testing Service.

This report is a review and summary of current information regarding testing accommodations currently used in different states and districts for English language learners (ELLs). The federal No Child Left Behind (NCLB) Act of 2001 requires the inclusion of ELLs in assessments used by the states for accountability purposes. This represents a federal education requirement that did not exist prior to the enactment of NCLB. However, the policies for identification and reclassification of ELLs, appropriate testing accommodations, and testing requirements are state-level decisions. In order to validly and fairly assess the skills of ELL students, testing accommodations are made available where necessary by the states. However, there is no common set of standards across the states as to what are appropriate accommodations permitted for ELLs. Similarities and differences among states regarding ELL testing accommodations are documented in this review. Special attention is given to the ELL accommodation policies for states with high school exit examinations because these are the high-stakes exams, which have the clearest relevance in designing accommodation policies for ELLs in taking the SAT®.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Young, J. W., Steinberg, J., Cline, F., Stone, E., Martiniello, M., Ling, G., & Cho, Y. (2010). Examining the Validity of Standards-Based Assessments for Initially Fluent Students and Former English Language Learners. *Educational Assessment, 15*(2), 87–106.

To date, assessment validity research on non-native English speaking students in the United States has focused exclusively on those who are presently English language learners (ELLs). However, little, if any, research has been conducted on two other sizable groups of language minority students: (a) bilingual or multilingual students who were already English proficient when they entered the school system (IFEPs), and (b) former English language learners, those students who were once classified as ELLs but are now reclassified as being English proficient (RFEPs). This study investigated the validity of several standards-based assessments in mathematics and science for these two student groups and found a very high degree of score comparability, when compared with native English speakers, for the IFEPs, whereas a moderate to high degree of score comparability was observed for the RFEPs. Thus, test scores for these two groups on the assessments we studied appear to be valid indicators of their content knowledge, to a degree similar to that of native English speakers.

Component 1	Component 2	Component 3	Component 4	Component 5
			✓	✓

Yu, L., Livingston, S. A., Larkin, K. C., & Bonett, J. (2004). *Investigating differences in examinee performance between computer-based and handwritten essays,* (ETS Research Report Series RR-04–18). Princeton, NJ: Educational Testing Service.

This study compared essay scores from paper-based and computer-based versions of a writing test for prospective teachers. Scores for essays in the paper-based version averaged nearly half a standard deviation higher than those in the computer-based version, after applying a statistical control for demographic differences between the groups of examinees taking the two versions. The statistical control was implemented by means of a propensity score, applying weights to members of one group to match the propensity score distribution of the other group. The score difference between the groups did not change substantially when the analysis was restricted to examinees taking the same mix of essay topics or to examinees taking one particular essay.

Component 1	Component 2	Component 3	Component 4	Component 5
✓	✓	✓		

Zakaluk, B. L., & Samuels, S. J. (1988). *Readability: Its past, present, and future*. Newark, NJ: International Reading Association.

This foundational collection of articles highlights the changes that have occurred in readability research from the 1950s to the late 1980s. Considering the current role of readability in stimulus material selection, the authors made some interesting predictions about future research in this field, particularly in issues related to qualitative measures of text readability. Topics covered include case studies examining the assigning of grade levels without formulas and new thinking directed at the assessment of text complexity.

Component 1	Component 2	Component 3	Component 4	Component 5
✓				

Zandvliet, D., & Farragher, P. (1997). A comparison of computer-administered and written tests. *Journal of Research on Computing in Education*, *29*(4), 423-438.

An early study using a computer assessment program which tested for equivalence with written responses to this test compared to electronic responses. Results of the comparative analysis indicated no significant differences between test scores but survey responses did indicate a student preference for the computer-based test. Analysis of student test-path data recorded by the computer indicated that the computerized tests took slightly longer for students to complete.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Zapata-Rivera, D., & Hansen, E. G. (2009, April). *Analyzing the learning potential of existing games using evidence-centered design and cognitive task analysis.* Paper presented at the annual meeting of the American Educational Research Association (AERA), San Diego, California, April 15, 2009.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Zenisky, A. L., & Sireci, S. G. (2007). *A summary of the research on the effects of test accommodations: 2005–2006* (Technical Report 47). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

The purpose of this report is to provide an update on the state of the research on testing accommodations, as well as to identify promising areas of research to further clarify and enhance understanding of current and emerging issues. The research described encompasses empirical studies of score comparability and validity studies, as well as investigations into accommodations use and perceptions of their effectiveness. Taken together, the current research explores many of the issues surrounding test accommodations practices in both breadth and depth. Insofar as reporting on the findings of current research studies is a primary goal of this analysis, a second goal is to also identify areas requiring continued investigation in the future.

Component 1	Component 2	Component 3	Component 4	Component 5
				

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.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Zieky, M. (2006). Fairness review. In Downing, S. & Haladyna, T. (Eds.) *Handbook of Test Development*. Mahwah, NJ: Erlbaum.

The chapter describes what fairness review is, why it is done, and how to do it. The chapter discusses 6 fairness review guidelines: Treat people with respect; Minimize the effects of construct-

irrelevant knowledge; Avoid material that is unnecessarily controversial, offensive or upsetting; Use appropriate terminology to refer to groups; Avoid stereotypes; Represent diversity in test materials.

Component 1	Component 2	Component 3	Component 4	Component 5
				

Zieky, M. J. (1993). Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.), *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaum.

The chapter explores the issues raised when DIF statistics are introduced into an ongoing, large-scale test development process for the first time. Issues discussed include: What should the matching criterion be? Should use of DIF be symmetrical? Should use of DIF to eliminate items be automatic or judgmental? Which group should be analyzed and how should those groups be defined? How should DIF data be presented to test developers? When and how should DIF data be used in the test development process? How large should sample sizes be?

Component 1	Component 2	Component 3	Component 4	Component 5
				

Zwick, R. (1991). Effects of Item Order and Context on Estimation of NAEP Reading Proficiency. Educational Measurement: Issues and Practice, *10*(3), 10-16.

Item parameter estimates derived through item response theory methods have been considered relatively robust to changes in item position and context, but the anomaly in reading scores from the 1986 National Assessment of Educational Progress (NAEP) illustrates problems with common population equating procedures when there are test form changes.

Component 1	Component 2	Component 3	Component 4	Component 5