

***Comparing Apples to Apples: Challenges and Approaches to
Establishing the Comparability of Test Variations***

Phoebe C. Winter
Pacific Metrics
April 9, 2009

Paper presented at the annual meeting of the National Council on Measurement in Education as part of the symposium, *Flight or Fancy: Innovations in Comparability, Computer-Interactive, and Other Things Testing* (Rebecca Kopriva, chair), April 2009, San Diego, CA.

Comparing Apples to Apples: Challenges and Approaches to Establishing the Comparability of Test Variations¹

Preface

This goal of this paper is to spark discussion, research, and critical thinking about what we mean when say “comparability,” what we want when we want score comparability, and how we can evaluate comparability. The paper came about because of a series of research studies that focused on investigating methods of evaluating comparability (Barton, March 2009, draft; DePascale, March 2009a, draft; DePascale, March 2009b, draft; Lottridge, Nicewander, and Mitzel, February 2009, draft; Sireci and Wells, March 2009, draft;).² As we designed and implemented the studies, we became aware of a need to better understand what we meant by comparability and especially its relationship to the type and validity of inferences we could make about examinee performance.

Background: Research Studies

The issue of comparability is not a new one. When states began shifting to the use of performance tasks in their large-scale assessments in the 1990s (now largely abandoned), measurement theorists called for a fresh look at how we determine the comparability of large-scale tests (see for example, Phillips, 1996). In 1999, Congress commissioned the National Research Council’s Board on Testing and Assessment to study the feasibility of linking scores from various state tests to each other and to the National Assessment of Educational Progress (National Research Council, 1999), as part of a series of studies intended to inform the debate about a voluntary national test. As states went back to using

¹ Much of this paper is an extension of writing done by the author for the grant proposal referenced below.

² These studies were funded by an Enhanced Assessment Grant from the US Department of Education, awarded to the North Carolina Department of Public Instruction, in partnership with the Council of Chief State School Officers and its SCASS TILSA and SCASS CAS. Publication of this document shall not be construed as endorsement of the views expressed in it by the US Department of Education, the North Carolina Department of Public Instruction, or the Council of Chief State School Officers.

more traditional testing formats, the issue of comparability returned to its traditional role, that of equating test forms designed to be parallel in content, difficulty, and format.

Test Variations

Under the No Child Left Behind Act (NCLB), which is the 2001 reauthorization of the Elementary and Secondary Education Act, the issue of comparability has become more complex again. A key feature of the law is its strengthened focus on the inclusion of students with disabilities and English language learners in state assessment and accountability systems. States no longer question whether to include these students, but rather *how* to include them in a way that will provide valid, reliable, and useful information about their knowledge and skills. To maximize the number of students assessed against grade-level content standards, states have developed variations of their general tests designed to increase accessibility for students who cannot demonstrate their knowledge and skills using the general state assessment.

Such variations include translations into students' native languages for subsets of English language learners, linguistically simplified versions of the general test for subsets of English language learners and some students with disabilities, and alternative formats referenced to grade-level standards, such as portfolio assessments and checklists, for both groups of students. For example, New York and New Mexico offer translations of their general tests to English learners who are literate in a language other than English. Virginia offers test forms with reduced linguistic complexity to English language learners. Massachusetts offers a portfolio assessment aligned to grade-level standards to students with disabilities and others who cannot take the general form, as well as dual-language (English-Spanish) math and science test booklets for students who qualify for this accommodation (Sireci and Khaliq, 2002).

A practical outcome of NCLB's testing requirements and its intent that scores be useful, meaningful, and used to improve schools, instruction, and student learning is that states have investigated ways of

providing more cost-effective testing and timely reporting. As a result, a steadily increasing number of states have begun offering their assessments via computer, allowing schools and districts to choose the mode of administration. Another feature that has affected state interest is that computer-based versions of paper-and-pencil tests can provide some accommodations that are difficult to provide within a standard administration, such as translated instructions or in-frame definitions (see for example, Dolan, Hall, Banerjee, Chun, and Strangman, 2005).

All of these test variations are intended to address the same content standards as the general assessment, with the same rigor, and yield scores that are referenced to the same achievement standards³. States use them to report on student achievement and aggregate the scores with their general test scores for accountability purposes. However, states have had difficulty demonstrating that the results from test variations are comparable in content and score meaning to their general assessment results. In fact, comparability was one of four topics on which the U.S. Department of Education provided technical assistance at a meeting designed to assist states in meeting the requirements of peer review of their standards and assessment systems.

The Studies

In 2006, the North Carolina Department of Education and partner states from the Council of Chief State School Officers' State Collaboratives on Assessment and Student Standards, working with a group of researchers, were awarded a grant from the US Department of Education to investigate methods of evaluating test score comparability (Winter, 2006). The research covers four types of test variations (the researchers for each topic are in parentheses):

1. oral native language versions, administered via a video of the English-language version of the test (Stephen Sireci and Craig Wells);

³ This paper uses the term "test variations" solely for tests designed to measure the same content and same levels of achievement. A test written to alternate or modified achievement standards is NOT a variation of the general test in our terminology.

2. clarified language (also known as sheltered English, simplified English, plain language) versions (Charles DePascale);
3. alternative formats —portfolios and checklists—or non-parallel native language forms (Karen Barton and Phoebe Winter); and
4. computer-delivered versions of the paper-and-pencil test (Susan Lottridge, Alan Nicewander, Howard Mitzel, and Matthew Schulz).

Some Results

The results of the studies are being compiled and reported. A variety of methods of evaluating comparability, from empirical studies to analyses of extant data were investigated. Some results of the investigations are summarized in Table 1. For more detail, readers should consult the project reports, which are listed in the Appendix.

The research team agreed that expert judgment plays a critical role in building in comparability when designing test variations and in evaluating whether scores are comparable, particularly when more than the mode of delivery has changed. For test variations based on item revisions (e.g., native language versions and clarified language versions), expert involvement in item revision so that the content, process, or skills targeted are maintained at the item level, followed by expert review of the results of revision, are necessary steps in the development and evaluation process. In their project reports, Abedi (2008, draft), DePascale (March 2009a, draft), Sireci (December 2007, draft), and Wright (March 2009, draft), discuss some specific considerations for item revision and review for native language and clarified language assessments. Research is needed on judgment-based approaches to evaluating the degree to which tests of different formats (e.g., a paper-based-test and an assessment based on collections of student work) measure comparable content, processes, and skills.

Table 1. Selected results from study of methods for evaluating test score comparability

Test Variations	Methods Evaluated	Conclusions
Oral/video translation <ul style="list-style-type: none"> • Grade 3 Science • Grade 5 Math • Grade 6 Social Studies • Spanish and English (Sireci and Wells, Feb. 2009)	Examine structure of assessments and items: <ul style="list-style-type: none"> • Confirmatory factor analysis • Multi-dimensional scaling • Differential item functioning Used replications to mitigate the effects of small sample size and disparate proficiency distributions	From Sireci , S. and Wells, C. (February 2009): <ul style="list-style-type: none"> • Replication methodology helpful when faced with small samples and widely different proficiency distributions <ul style="list-style-type: none"> ○ Gauge variability due to sampling (random) error ○ Gauge variability due to distribution differences • Multiple methods for evaluating structure are helpful • Effect size criteria helpful for DIF • Congruence b/w structural & DIF results
Computer-administered <ul style="list-style-type: none"> • End-of-course • Algebra ,I English II, Biology, Civics, History (Lottridge, Nicewander, and Schulz, Feb. 2009)	Compare results from a repeated measures (within subjects) administration to results from propensity score matching	<ul style="list-style-type: none"> • Propensity score matching produced similar results to studies using within-subjects samples. • Propensity score method provides a viable alternative to the difficult-to-implement repeated measures study. • Propensity score method is sensitive to group differences. For instance, the method performed better when 8th and 9th grade groups were matched separately.
Clarified language <ul style="list-style-type: none"> • Grades 3-8 mathematics (DePascale, Feb. 2009)	Examine development procedures and analyze item functioning	<ul style="list-style-type: none"> • Carefully documented and followed development procedures focused on maintaining the item construct can support comparability arguments. • Linking/equating approaches can be used to examine and/or establish comparability. • Comparing item statistics using the non-target group can provide information about comparability.
Checklist based on student work <ul style="list-style-type: none"> • Grades 3-8 reading and mathematics (Barton, Feb. 2009; Winter, Feb. 2009)	Evaluate results from a study comparing scores on the general test to checklist scores	<ul style="list-style-type: none"> • The burden of proof is much heavier for this type of test variation. • A study based on students eligible for the general test can provide some, but not solid, evidence of comparability. • Judgment-based studies combined with empirical studies are needed to evaluate comparability. • More research is needed in methods for evaluating what constructs each test type is measuring.

Comparability of Test Scores

As noted earlier, the primary reason states use test variations is to provide access to the test content for students who are unable to take the general test version and to make sure the test accesses these students' knowledge and skills; these test variations are intended to address the same content standards as the general assessment, with the same rigor, and yield scores that are referenced to the same achievement standards. For NCLB accountability purposes (and for other uses of the scores), scores from these tests should be comparable, at some level. In general, test scores can be considered comparable if they can be used interchangeably. The same interpretations should be able to be made, with the same level of confidence, from scores based on variations of the same test.

How comparability is defined, however, depends in part on what level of score (e.g., raw score, achievement-level score) is being used and how the score is being used. For example, if scores are being aggregated and reported at the scale score level, comparability is defined at that level; if scores are being interpreted at the achievement score level, they must be interchangeable at that level. This means that the test and its variations must, **at the level or grain size at which they are being compared**

- measure the same set of knowledge and skills (i.e., constructs),
- produce scores at the desired level of specificity that reflect the same degree of achievement on those constructs, and
- have similar technical properties (e.g., reliability, decision consistency, subscore relationships) in relationship to the level of score reported.

For example, Brenda takes the Spanish-language version of the state mathematics test, Sean takes the portfolio-based version, and Jacquetta takes the general version. All three students earn achievement level scores of "Above Proficient." If the test scores are comparable at the achievement score level, then this means that, relative to the same body of knowledge and skills, all three students meet the state's definition of "Above Proficient" performance. These scores can be included with

confidence into the “Above Proficient” category in aggregated reports for purposes such as school evaluation and accountability. That is, in this scenario, achievement level scores can be used interchangeably regardless of the test variation taken.

On the other hand, the scale scores each test yields may not be comparable. The portfolio version will not yield scale scores that are comparable to the scale scores from the general test. The Spanish language version may be developed and equated so that scores are reported on the same scale as the general version. If this is the case, then the scale scores of these two variations should be comparable.

Characterizing Comparability

“Comparability” is not understood as a technical measurement property in the same way that “validity” or “bias” are. To a greater extent than even “validity,” its meaning is defined by the context in which it is used. Not clearly defining what aspect or level of score comparability is under consideration can be a source of misunderstanding in discussions about the topic.

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

Standard 4.10

A clear rationale and supporting evidence should be provided for any claims that scores earned from different forms of a test may be used interchangeably. In some cases, direct evidence of score equivalence may be provided. In other cases, evidence may come from a demonstration that the theoretical assumptions underlying procedures for establishing score comparability have been sufficiently satisfied. The specific evidence and rationale required will depend in part on the intended uses for which score equivalence is claimed, (p. 57)

The *Standards* recognize that different types of evidence will be required for different types of variations: “Score equivalence is easiest to establish when different forms are established following identical procedures When this is not possible, for example, in cases where different test formats are used, additional evidence may be required to establish the requisite degree of score equivalence for the

intended context and purpose.” (p. 57.) In addition, an entire chapter of the *Standards* is devoted to issues related to testing students with diverse linguistic backgrounds.

In their works on linking educational assessments, Mislevy (1992) and Linn (1993) provide a useful framework for considering the comparability of different tests. Mislevy and Linn make distinctions among different levels or strengths of linking, and the connections they make between the purposes of tests and comparability issues can be used to structure the way we look at comparability among tests. The strongest linking procedure, equating, is appropriate for test forms that have been constructed from the same test blueprints. For tests designed to measure the same body of knowledge and skills but using different types of evidence, calibration may be most appropriate. Finally, social moderation techniques may be needed for tests that are designed to measure the same constructs but in ways different enough that equating or calibration studies are not possible.

Mislevy (1992) uses the terms equating, calibration, and social moderation to distinguish between different properties of each linking procedure as well as the methods themselves. As the result of *equating*, “Any question that could be addressed using [Test] X scores can be addressed in exactly the same way using [Test] Y scores (p. 21).” *Calibration* “...relates the results of different assessments to a common frame of reference, and thus to one another only indirectly (p. 24).” Finally, “*social moderation* uses judgment to match levels of performance on different assessments directly to one another (p. 25).”

Constructs and Inferences

While the linking distinctions outlined by Mislevy (1992) and Linn (1993) can be used to consider what we mean by comparability, we need to explicitly add construct comparability and grain size, or level, of the inference we wish to make from the scores to the framework (Winter, Feb. 2009). Figure 1 is a simple depiction of one way to consider these two aspects of comparability. “Content” is used instead of “construct” in the figure, because K-12 achievement tests tend to focus on content rather than defining the underlying constructs in a conceptual manner (Chudowsky and Pellegrino, 2003).

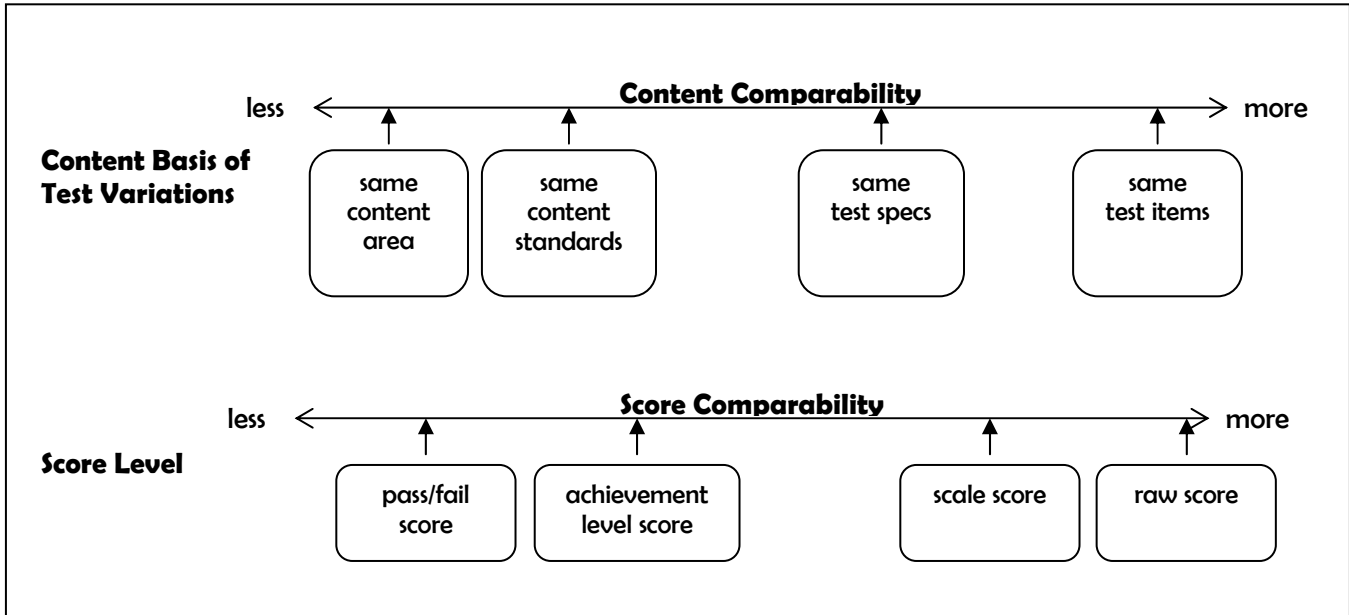


Figure 1. Comparability continuum

Comparability of Scores and Validity of Inferences

The series of studies described earlier began with the question of how to evaluate the comparability of scores from test variations. An important aspect of any evaluation, of course, is the set of criteria used to evaluate. Four criteria were developed based on NCLB peer review requirements as we understood them.

First, the test variation should support inferences that are more valid than those that the general test would support for the test variation’s targeted students.⁴ As an obvious example, on its face, a well-translated version of a math test would provide more valid inferences for a student who does not read English but is literate in his or her first language.

Second, the test variation must be aligned to the state grade-level content standards. It must measure the standards with at least the same breadth and depth as the general test⁵.

⁴ In the case of computer-based versions of paper-and-pencil tests, the target populations are the same; therefore the variation should support inferences that are equally as valid.

⁵ In some cases, the alignment of test variations to content standards, particularly those that sample student work from the classroom, may be better than the alignment of the general test to content standards.

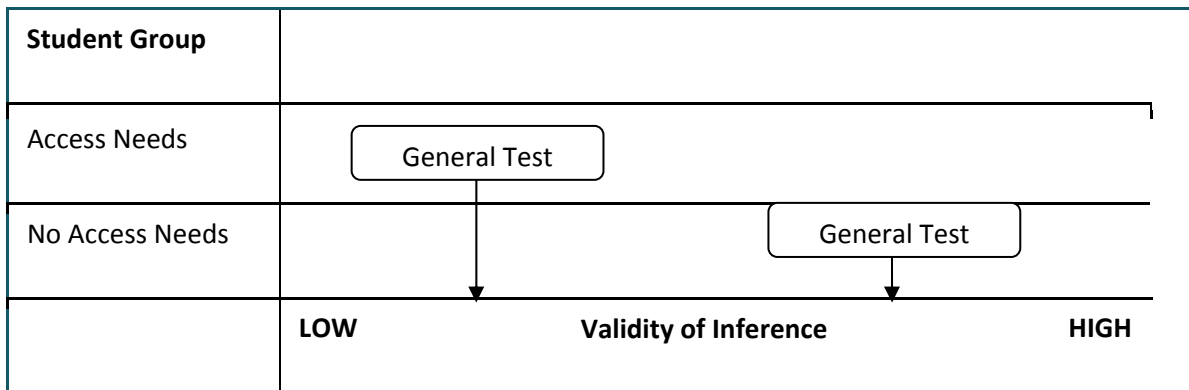
Third, the test variation should be scored and cut scores should be set so that the same degrees of knowledge and skills are required to meet each achievement level as are required by the general test.

Fourth, the test should provide results that are as reliable, at the desired level of score comparability, as the general test. For example, classification consistency into achievement levels should be as high for the test variation as it is for the general test.

What if a test variation yields more valid scores for a targeted group of students than the general test, but the scores cannot be said to be comparable, based on the other criteria above? That is, for reasons of lack of research or a mature understanding of how to measure the achievement of a group of students, the variation does not meet the requirements of comparability.

For example, say that there is a group of students with access needs for whom the general test provides inferences that are less valid than those for the rest of the student population, as illustrated in Figure 2.

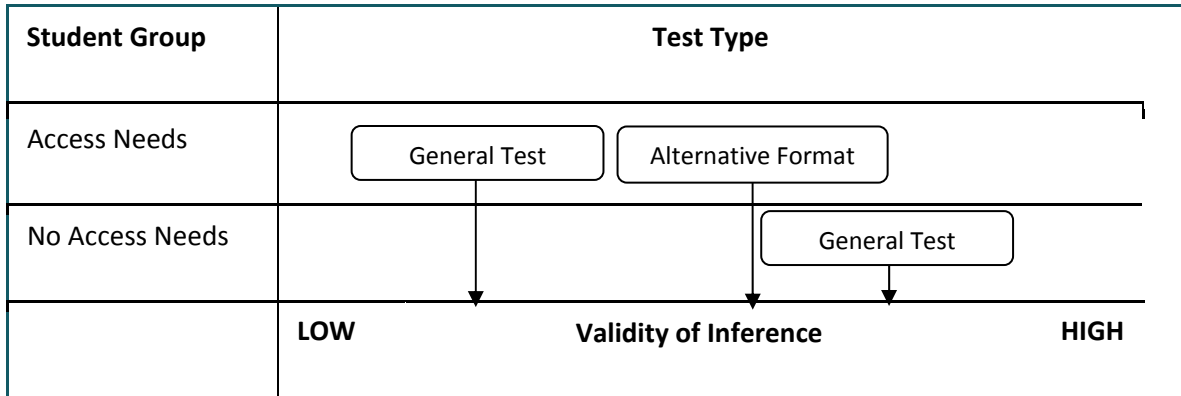
Figure 2: Illustration of Uneven Validity for Two Groups



Research and experience teaching and assessing these students provides a basis for assessing them using an alternative format (e.g., an observational scale backed up by evidence). The state develops such an assessment, working in as many features as possible to support technical quality and score integrity. However, when the state evaluates the alternative format it finds that the alternative format

does not support inferences for students with access needs that are as valid as inferences the general test supports for students without access needs, as illustrated in Figure 3.

Figure 3: Illustration of Uneven Validity for Two Groups, with Improvements for the Group with Access Needs



The alternative format clearly measure the targeted construct for the group of students with access needs better than the general test does. Although the alternative format does not meet the other three criteria for comparability, requiring students with access needs to take the general test yields a less valid and less comparable score for these students.

In deciding which type of test to use with students with access needs, we should be aware of the implications of using the less valid score from the general test. For example, a test variation may have lower classification consistency for its target population than the general test does for the general population. However, the test variation might produce higher classification consistency for students with access needs than the general test does. Thoughtful consideration of this type of issue must go into the decision of whether to use the variation. If the decision is made to use the test variation, the test developer and user have a responsibility to continue to conduct research and refine administration and scoring procedures as they learn how to make the variation more comparable to the general test.

Closing

As we move to new large-scale testing systems that may allow for variations beyond those designed to promote access for equity reasons, to variations designed to tap different types of knowledge, different constructs, in different modes of cognition and response, what we mean when say “comparability,” what we want when we want score comparability, and how we can evaluate comparability becomes even more complex and even more central to good measurement. We need to more carefully consider the constructs we are targeting in our measures and where there are similarities and differences across test variations. We need to consider scoring and linking models that reflect what we are learning about student knowledge and skills and how we are learning it.

The groundwork for considering these issues has been laid. Collaboration between tests theorists and experts in learning along the lines suggested by the National Research Council (2001) using models of cognition and learning in the test development process (Chudowsky and Pellegrino, 2003) is a start. Recent explorations of integrating the idea of learning progressions into classroom assessment (e.g., National Research Council, 2006; Smith, Wiser, Anderson, and Krajcik, 2006) are an example of how cognitive learning theory can be brought into the realm of assessment. Mislevy and colleagues’ evidence-centered design approach (e.g., Mislevy and Haertel, 2006) provides another example. The investigation of different ways of modeling test performance is represented by Mislevy and others’ use of Bayesian inferences networks (e.g., Levy and Mislevy, 2004). And Dorans, Pommerich, and Holland’s (2007) acknowledgement of and warnings about the “the descent of linking” (p. 355) shows that there is at least a recognition that we want to connect results of test variations in meaningful ways using appropriate techniques.

References

- Abedi, J. (2008, draft). Impact of language complexity on the assessment of ELL students: A focus on plain language assessment. Washington, DC: Council of Chief State School Officers.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Psychological Association.
- Barton, K. (February 2009). Alternative formats. Presentation at the February 2009 meeting of the State Collaborative on Assessment and Student Standards Technical Issues in Large-Scale Assessment Consortium, Orlando.
- Chudowsky, N., & Pellegrino, J. (2003). Large-scale assessments that support learning: What will it take? *Theory into Practice, 42*, 75-83.
- DePascale, C.A. (2009). Linguistic modifications and modified tests: Project summaries. Presentation at the February 2009 meeting of the State Collaborative on Assessment and Student Standards Technical Issues in Large-Scale Assessment Consortium, Orlando.
- DePascale, C.A. (March, 2009a, draft). Evaluating linguistic modifications: An examination of the comparability of a plain English mathematics assessment. Washington, DC: Council of Chief State School Officers.
- 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>
- Dorans, N.J., Pommerich, M., & Holland, P.W. (eds.) (2007). *Linking and aligning scores and scales*. New York: Springer Science + Business Media.
- Levy, R., & Mislavy, R.J. (2004). Specifying and refining a measurement model for a computer-based interactive assessment. *International Journal of Testing, 4*, 333-369.
- Linn, R.L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*, 83-102.
- Lottridge, S., Nicewander, W.A., & Mitzel, H.C. (February, 2009, draft). Summary of the online comparability studies for one state's end of course program. Washington, DC: Council of Chief State School Officers.
- Lottridge, S., Nicewander, W.A., & Schulz, E.M.. (February, 2009). Summary of online comparability studies. Presentation at the February 2009 meeting of the State Collaborative on Assessment and Student Standards Technical Issues in Large-Scale Assessment Consortium, Orlando.
- Mislavy, R.J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: Educational Testing Service.
- Mislavy, R.J., & Haertel, G. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practices, 25*, 6-20.

- National Research Council (1999). *Uncommon measures: equivalence and linkage among educational tests*. Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.
- National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. Pellegrino, J. Chudowsky, N., & Glaser, R. editors. Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.
- National Research Council. (2006). *Systems for state science assessment*. Committee on Test Design for K–12 Science Achievement. M.R. Wilson and M.W. Bertenthal, eds. Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.
- Phillips, G.W. (ed.) (1996). *Technical issues in large-scale performance assessment*. Washington, DC: National Center for Educational Statistics.
- Sireci, S. G., & Khaliq, S. N. (2002, April). *An analysis of the psychometric properties of dual language test forms*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Sireci, S.G. (December, 2007, draft). *Validity issues and empirical research on translating educational achievement tests: A review of the literature*. Washington, DC: Council of Chief State School Officers.
- Sireci, S.G., & Wells, C.S. (February, 2009). *An update on evaluating the comparability of video accommodations for ELLs*. Presentation at the February 2009 meeting of the State Collaborative on Assessment and Student Standards Technical Issues in Large-Scale Assessment Consortium, Orlando.
- Sireci, S.G., & Wells, C.S. (March, 2009, draft). *Evaluating the comparability of English and Spanish video accommodations for English language learners*. Washington, DC: Council of Chief State School Officers.
- C.L., Smith, Wiser, M., Anderson, C.W., & Krajcik, J. Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research & Perspectives*, 4, 1 – 98.
- Winter, P.C. (2006). *Strengthening the comparability and technical quality of test variations*. A proposal submitted to the US Department of Education.
- Winter, P.C. (February, 2009). *Evaluating the comparability of scores from test variations: Examples of possible evidence for an alternative format*. Presentation at the February 2009 meeting of the State Collaborative on Assessment and Student Standards Technical Issues in Large-Scale Assessment Consortium, Orlando.
- Wright, L.J. (March, 2009, draft). *A guide for developing translations of standardized assessments*. Washington, DC: Council of Chief State School Officers.

Appendix: List of Project Research Reports

- DePascale, C.A. (March, 2009a, draft). Evaluating linguistic modifications: An examination of the comparability of a plain English mathematics assessment. Washington, DC: Council of Chief State School Officers.
- DePascale, C.A. (March, 2009b, draft). Modified tests for modified achievement standards: Examining the comparability of scores to the general test. Washington, DC: Council of Chief State School Officers.
- Lottridge, S., Nicewander, W.A., & Mitzel, H.C. (February, 2009, draft). Summary of the online comparability studies for one state's end of course program. Washington, DC: Council of Chief State School Officers.
- Lottridge, S. & Nicewander, A. (2008a). Comparing computer-based and paper-based test scores in one state's End-of-Course biology program: Results using propensity score matching. Washington, DC: Council of Chief State School Officers.
- Lottridge, S., & Nicewander, A. (2008b). Comparing computer-based and paper-based test scores in one state's End-of-Course civics & economics program. Washington, DC: Council of Chief State School Officers.
- Lottridge, S., & Nicewander, A. (2008c). Comparing computer-based and paper-based test scores in one state's End-of-Course U.S. history program. Washington, DC: Council of Chief State School Officers.
- Lottridge, S., Nicewander, A., & Box, C. (2008). Comparing computer-based and paper-based test scores in one state's End-of-Course algebra I program: Results using propensity score matching with two approaches for missing data. Washington, DC: Council of Chief State School Officers.
- Lottridge, A., Nicewander, A., & Mitzel, H. (2008a). Comparing computer-based and paper-based test scores in one state's End-of-Course algebra I program. Washington, DC: Council of Chief State School Officers.
- Lottridge, S., Nicewander, A., & Mitzel, H. (2008b). Comparing computer-based and paper-based test scores in one state's End-of-Course English I program. Washington, DC: Council of Chief State School Officers.
- Lottridge, S., Nicewander, A., & Mitzel, H. (2008c). Comparing computer-based and paper-based test scores in one state's End-of-Course biology program. Washington, DC: Council of Chief State School Officers.
- Lottridge, S., Nicewander, A, & Schulz, M. (2008a). Comparing computer-based and paper-based test scores in one state's End-of-Course algebra I program: Results using propensity score matching. Washington, DC: Council of Chief State School Officers.
- Lottridge, S., Nicewander, A., & Schulz, M. (2008b). Comparing computer-based and paper-based test scores in one state's End-of-Course English I program: Results using propensity score matching. Washington, DC: Council of Chief State School Officers.
- Sireci, S.G., & Wells, C.S. (March, 2009, draft). Evaluating the comparability of English and Spanish video accommodations for English language learners'. Washington, DC: Council of Chief State School Officers.

Sireci, S.G., Wells, C.S., and Dunn, J. (March, 2008) Evaluating the comparability of video read-aloud accommodations for English language learners. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Sireci, S.G., & Wells, C.S. (January, 2009, draft).Evaluating video read-aloud accommodations for English language learners on a 5th-grade science test. Washington, DC: Council of Chief State School Officers.