

Evaluating the Comparability of Scores from Achievement Test Variations

Phoebe C. Winter, Editor

THE COUNCIL OF CHIEF STATE SCHOOL OFFICERS

The Council of Chief State School Officers (CCSSO) is a nonpartisan, nationwide, nonprofit organization of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five U.S. extra-state jurisdictions. CCSSO provides leadership, advocacy, and technical assistance on major educational issues. The Council seeks member consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public.

Evaluating the Comparability of Scores from Achievement Test Variation

Based on research funded by an Enhanced Assessment Grant from the U.S. 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 U.S. Department of Education, the North Carolina Department of Public Instruction, or the Council of Chief State School Officers.

COUNCIL OF CHIEF STATE SCHOOL OFFICERS

Steven Paine (West Virginia), President

Gene Wilhoit, Executive Director

Phoebe C. Winter, Editor

Council of Chief State School Officers
One Massachusetts Avenue, NW, Suite 700
Washington, DC 20001-1431
Phone (202) 336-7000
Fax (202) 408-8072
www.ccsso.org

ISBN: 1-884037-28-3

Copyright © 2010 by the Council of Chief State School Officers, Washington, DC

All rights reserved.

Acknowledgments

This research project benefited from the support and advice of members of two of CCSSO's State Collaboratives on Assessment and Student Standards: the Technical Issues in Large-scale Assessment consortium, led by Doug Rindone; and the Comprehensive Assessment Systems for Title I consortium, led by Carole White. The editor wants to thank particularly the state department staff members of the project research team for their work and thoughtful participation in the project:

Mildred Bazemore, North Carolina Department of Public Instruction
Jane Dalton, North Carolina Department of Public Instruction
Tammy Howard, North Carolina Department of Public Instruction
Laura Kramer, Mississippi Department of Education
Joseph Martineau, Michigan Department of Education
Nadine McBride, North Carolina Department of Public Instruction
Marcia Perry, Virginia Department of Education
Liru Zhang, Delaware Department of Education

Contents

Section 1: Introduction

Chapter 1: Comparability and Test Variations.....	1
---	---

Section 2: Studies of Comparability Methods

Chapter 2: Summary of the Online Comparability Studies for One State’s End-of-course Program	13
Chapter 3: Evaluating the Comparability of English and Spanish Video Accommodations for English Language Learners.....	33
Chapter 4: Evaluating Linguistic Modifications: An Examination of the Comparability of a Plain English Mathematics Assessment.....	69
Chapter 5: Evaluating the Comparability of Scores from an Alternative Format.....	95
Chapter 6: Modified Tests for Modified Achievement Standards: Examining the Comparability of Scores to the General Test	105

Section 3: Literature Reviews Related to Comparability of Various Types of Test Variations

Chapter 7: Comparability of Paper-based and Computer-based Tests: A Review of the Methodology	119
Chapter 8: Validity Issues and Empirical Research on Translating Educational Achievement Tests	153
Chapter 9: Considerations for Developing and Implementing Translations of Standardized K–12 Assessments	185
Chapter 10: Impact of Language Complexity on the Assessment of ELL Students: A Focus on Linguistic Simplification of Assessment.....	205
Chapter 11: Alternative Formats: A Review of the Literature.....	223

Section 4: Summary

Chapter 12: Where Are We and Where Could We Go Next?.....	233
---	-----

Section 1: Introduction

Chapter 1: Comparability and Test Variations

Phoebe C. Winter
Pacific Metrics

In 2006, a consortium of state departments of education, led by the North Carolina Department of Public Instruction and the Council of Chief State School Officers, was awarded a grant from the U.S. Department of Education¹ to investigate methods of determining comparability of variations of states' assessments used to meet the requirements of the No Child Left Behind Act of 2001 (NCLB). NCLB peer review guidelines includes a requirement that states using variations of their assessments based on grade level academic achievement standards must provide evidence of comparability of proficiency from these variations to proficiency on the general assessment² (U.S. Department of Education, July 2007, revised January 2009). The consortium of states focused on two types of K–12 achievement test variations: computer-based tests that are built to mirror the state's paper-based tests and test formats that are designed to be more accessible to specific student populations than the general test is. In addition, the studies conducted included a study of the comparability of an alternate assessment based on modified achievement standards; the comparability issues related to those assessments are different from the ones discussed here.

Researchers are beginning to think and write about comparability in a slightly different way than they have traditionally. In the recent past, comparability of test scores has been thought of mostly in terms of scale scores. Here, comparability is discussed in terms of the score that is being interpreted, which may be a scale score but may also be a less fine-grained score such as an achievement level score. Comparability has been thought of in terms of linking and, in state testing programs, most often in terms of equating. Here, linking scores across test forms is discussed, but the issues are expanded to include the comparability of achievement level scores from tests of very different formats. The issues include, and in the case of alternative formats go beyond, those related to what is referred to as “extreme linking” (Dorans, Pommerich, and Holland, 2007, p. 356; Middleton and Dorans, May, 2009, draft).

The details of what is meant by score comparability are still being defined as the measurement community conducts research and explores theoretical bases for comparability. The language used to discuss comparability is evolving as well. For example, Dorans and Walker (2007) differentiate between “interchangeability” and “comparability,” reserving the first term for scores that have been equated, and Kolen and Brennan (2004) have used “comparability” as an outcome of linking (see, for example, p. 3). *Standards for Educational and Psychological*

¹ This work was supported by a Grant for Enhanced Assessment Instruments awarded by the U.S. Department of Education.

² We use general assessment or test to refer to the usual format of the test, typically administered as a paper-and-pencil test.

Testing (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999) uses “comparable” when discussing whether scores can be considered interchangeable (p. 57). Because the research described here is concerned primarily with the use of the scores, “interchangeable” is used to mean that a score can be used in the same way, with the same construct-based inferences, as another score.

In one sense, different test formats meant to measure the same construct can be seen as alternate forms. For a computer-based test, the idea that it is an alternate form of a paper-based test is not much of a stretch from what measurement specialists are used to thinking about. To think of a portfolio-based version of a paper-and-pencil test as an alternate form of it is more of a stretch. Alternatively, test variations (other than computer-based versions of the paper-and-pencil test) can be thought of as accommodations, but that raises the question of whether accommodated tests can be thought of as alternate forms of the general, non-accommodated test. It can be seen, then, that the idea of comparability is just a part of larger issues of validity, the question of standardization and what part of the test, from administration to inference, is standardized, and how valid inferences about student achievement can be made across a large, diverse student population.

What Are Test Variations?

In the context of K–12 achievement testing, a test variation is an adaptation of the test that is used for most students (the general test). A test variation may be a computer-based version of the general test or it may be a format developed for students who cannot take the general test but are being taught to grade level content and achievement standards. Test variations

- address the *same content standards* as the general assessment
- address content standards with the *same rigor* as the general assessment
- yield scores that are interpreted in reference to the *same achievement (or performance) standards* as those used for the general assessment

These variations are not easier forms of the general test. Test variations are built to measure the same constructs (the knowledge and skills covered by the content standards) as the general test.³ States use test variations to measure individual student achievement, and these scores are aggregated with scores from the general test for purposes of program improvement and accountability.

In many cases (computer-based versions of the paper-based test are a notable exception), test variations are designed to increase accessibility for students who are receiving the general education curriculum but, for reasons of language barriers or disabilities, cannot demonstrate

³ While many test variations may have the same purpose as an accommodation (that is, to allow students to access the test), they are different in that they are different forms of the test rather than conditions added to the administration of the general test.

their knowledge and skills using the general state test. Such variations include translations⁴ into students' home languages for English language learners, linguistically simplified versions of the general test for English language learners and students with disabilities, and alternative formats referenced to grade level standards for both groups, such as portfolio assessments and checklists.

In other cases, states offer versions of their paper-and-pencil assessments via computer, in hopes of providing more cost-effective testing and timely reporting or to allow for more flexibility in administration. At this writing, most states allow schools and districts to choose the mode of administration for their students.

What Is Score Comparability?

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 variations of the same test. How comparability is defined depends 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 the that level. This means that the test and its variations must

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

Why Is Score Comparability Important?

Test variations are intended to provide scores that can be used in the same way as scores from the general test. 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 an achievement level score 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, achievement level scores can be used interchangeably regardless of the test variation taken.

⁴ Test variations in another language are not necessarily (indeed, are not usually) word-for-word translations of the general test. The more proper terms for such variations are “trans-adaptations,” “adaptations into another language,” or “test translation and adaptation.” We use translation to mean translating the test so that the meaning of the test items and the knowledge and skills required to respond are the same in the new language version as in the original English version.

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.

What Is the Relationship between Comparability and Validity?

It is unlikely that some groups of students will be able to demonstrate their knowledge and skills on the general version of a large-scale test; this means that the general version will not support valid inferences about these students' achievement in the targeted content area.⁵

An obvious example is when a student who does not read English is given a verbally loaded mathematics problem-solving test that is written in English. In this case, teachers and other test score users will get little or no information about the student's mathematics achievement. The inference made from the student's test score would likely be that the student has extremely low math problem-solving skills. This would be an invalid inference—the test did not measure the student's problem-solving skills at all.

A less obvious example is giving the mathematics test to a student who is not a fluent reader. It is likely that this student's math achievement will also be underestimated by the test. An oral English administration of the test might provide a more valid inference about the student's math knowledge and skills. For other students with poor reading skills, a test variation consisting of language accessible items, such as a linguistically simplified test,⁶ might be more appropriate.

In all three cases described above, the goal is the same: to allow the student to demonstrate his or her level of knowledge and skills in the state grade level content standards, against the state grade level achievement standards, so a valid inference, comparable to that from the general test, can be made.

In other cases, a paper-and-pencil test form, regardless of accommodations offered or revisions made to the test form, may not allow a student to demonstrate achievement. This might be the case for an English language learner with very low English comprehension skills in a state where neither written nor oral translations into the student's language are available. This may also be the case for certain students with disabilities, such as those who have been recently blinded or some students with autism.

To obtain valid inferences about these students, states may use a test variation in which the administration and data collection format is different from the standard form. For example,

⁵ Of course, a single test does not provide complete information about a student's achievement. However, some degree of inference is made about student achievement from large-scale tests, whether at the individual or aggregate level.

⁶ Linguistic simplification is a complex process. See Kopriva (2008) and Abedi (2006) for more information about creating tests that reduce the linguistic load on students.

teachers might collect student work that demonstrates student achievement on a content standard, or students may be given performance tasks aligned to the content standards. Again, the goal of these techniques is the same as described above: to provide more valid measures of students' knowledge and skills, relative to the same content and achievement standards measured by the general test, than can be obtained using the standard test format.

How Comparable Is Comparable Enough?

The comparability of test scores is a matter of degree. How comparable scores need to be for a specific test variation depends on how the test scores will be interpreted and used. We can think of the degree of comparability along two related dimensions, content and score level, as shown in Figure 1. Here content refers to both the subject matter and the degree of construct-relevant cognitive demand.

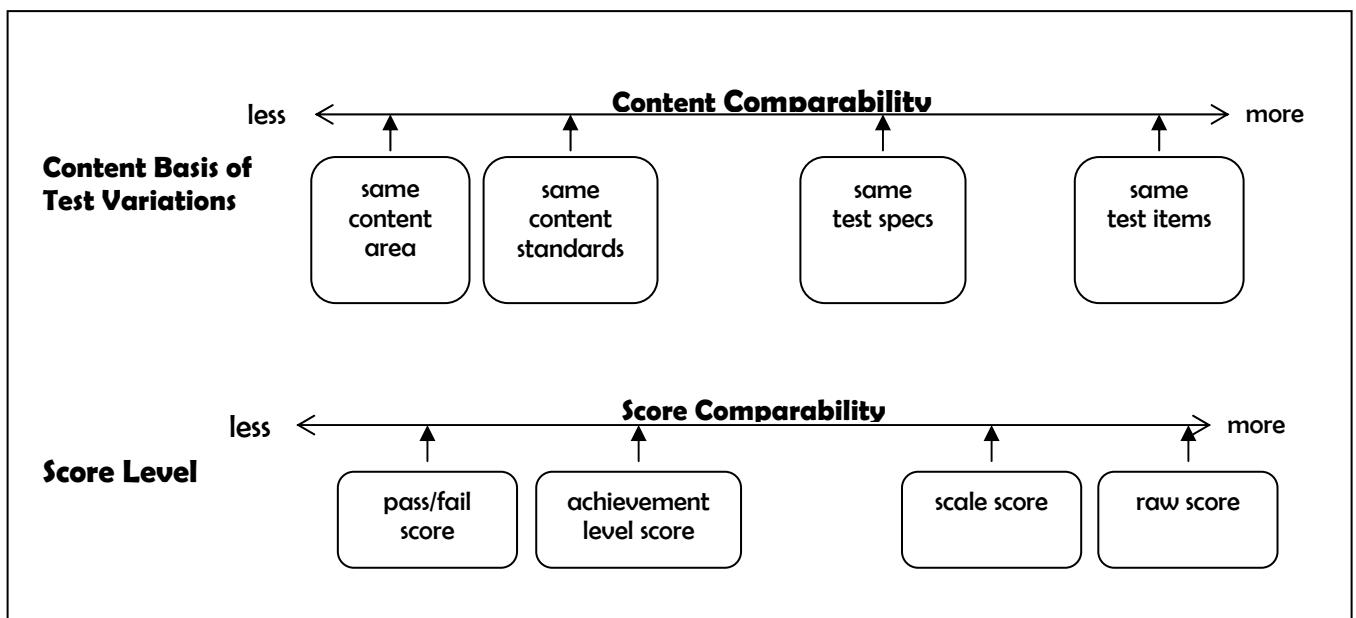


Figure 1. Comparability Continuum

These two dimensions combined determine the extent to which the same construct is measured, and the nature, or grain size, of construct equivalence. Where a test and its variation fit along one dimension is related to where they are on the other dimension. For example, computer-based and linguistically simplified versions of tests are both based on the *same test items* as the source paper-and-pencil tests; each item is intended to measure the exact same content, and the two tests are expected to yield *scale scores* (or, sometimes, raw scores) that are comparable to those from the paper-and-pencil tests. On the other hand, portfolios are based on the *same content standards*, but do not have the same items, as the general test; portfolios are expected to measure the same broad content (for example, defined at the objective, rather than item, level) and they are expected to yield *achievement level scores* (e.g., basic, meets, exceeds) that are comparable to those from the general assessments.

There is a fundamental assumption underlying the NCLB peer review requirement of comparability in reference to the general test (U.S. Department of Education, July 2007, revised January, 2009). The general test is presumed to support valid inferences about student achievement for the student population for whom the test was designed. It follows that a comparable test will support valid inferences for its target population.

So, How Comparable Is Comparable Enough?

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.⁷ To return to 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 test.⁸

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 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.

⁷ 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 valid.

⁸ 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.

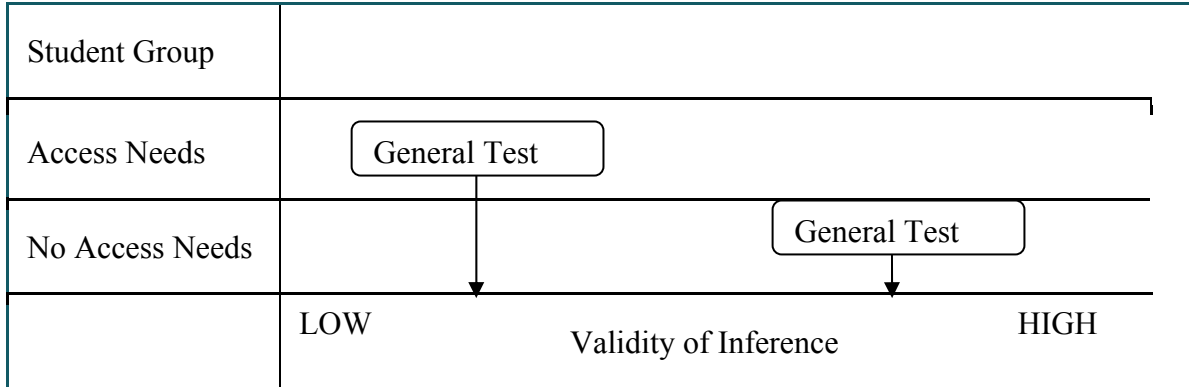


Figure 2. Illustration of Uneven Validity for Two Groups

Research and experience in teaching and assessing these students provide 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.

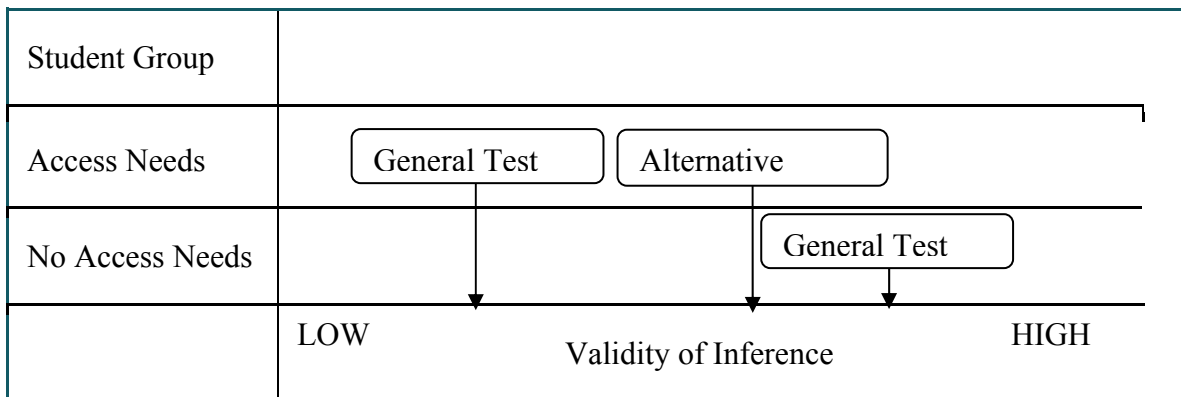


Figure 3. Illustration of Uneven Validity for Two Groups, with Improvements for the Group with Access Needs

The alternative format clearly measures the targeted construct for the group of students with access needs better than the general test does. Requiring these students to take the general test yields a less valid and less comparable (by our definition, comparable to the score on the general test for the general population) 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.

How Is Comparability Evaluated?

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999) includes the following requirement for using test scores interchangeably:

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 test variations: “Score equivalence is easiest to establish when different forms are established following identical procedures (p. 57),” which is close to the case with a computerized version of a paper-and-pencil test. “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),” reflects situations in which portfolios aligned to grade level standards or linguistically simplified English-language forms are used.

Four critical questions/criteria, based on the features defined earlier, can be used to evaluate the comparability of achievement test scores, in the framework of NCLB peer review requirements:

1. Does the test variation support more valid inferences about the target student population’s achievement than the general test does?
2. Is the test variation aligned to content standards at least as well as the general test is?
3. Does the test variation classify students into achievement levels based on the same degree of knowledge and skills as the general test does?
4. Does the test variation provide equally reliable results for the target population as the general test does for the general population?

Types of test variations will differ in the degree to which they satisfy each criterion. As in other work related to validity, a determination of “comparable enough” should be based on a reasoned appraisal of the strength of evidence for comparability and against comparability. The degree of comparability of test scores is evaluated through reviews of test content and student performance, analysis of student performance on the test variations and the general test, and review of test development procedures and the uses of test scores. Methods for evaluating

comparability are the focus of Sections 2 and 3. Section 2 contains the results of a series of studies of methods for evaluating comparability. Section 3 contains literature reviews related to comparability issues for several types of test variations.

The Need for Additional Research

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, questions of what is meant by “comparability,” what is desired in terms of score comparability, and how comparability can be evaluated become even more complex and even more central to good measurement. Researchers need to more carefully consider the constructs targeted in educational measures and the similarities and differences across test variations. Scoring and linking models that reflect the expanding research findings about student knowledge and skills and how students learn need to be developed and investigated.

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). Dorans, Pommerich, and Holland’s (2007) acknowledgement of and warnings about the “the descent of linking” (p. 355) shows that there is a recognition of the need to connect results of test variations in meaningful ways using appropriate techniques (see Middleton and Dorans, May 2009, draft, for an example of research in this area).

References

- Abedi, J. (2006). Psychometric issues in the ELL assessment and special education eligibility. *Teachers College Record*, 11, 2282–2303
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Chudowsky, N., & Pellegrino, J. (2003). Large-scale assessments that support learning: What will it take? *Theory into Practice*, 42, 75–83.
- Dorans, N.J., Pommerich, M., & Holland, P.W. (Eds.) (2007). *Linking and aligning scores and scales*. New York, NY: Springer Science+Business Media.
- Dorans, N.J., & Walker, M.E. (2007). Sizing up linkages. In Dorans, N.J., Pommerich, M., & Holland, P.W. (Eds.) (2007). *Linking and aligning scores and scales*. New York, NY: Springer Science+Business Media.
- Kolen & Brennan (2004) *Test equating, scaling, and linking: Methods and practices*. New York, NY: Springer Science+Business Media.
- Kopriva, R. (2008). *Improving testing for English language learners*. Mahwah, NJ: Laurence Erlbaum Associates.
- Levy, R., & Mislevy, R.J. (2004). Specifying and refining a measurement model for a computer-based interactive assessment. *International Journal of Testing*, 4, 333–369.
- Middleton, K., & Dorans, N.J. (May, 2009, draft). Assessing the falsibility of extreme linkages. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA, revised.
- Mislevy, R.J., & Haertel, G. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practices*, 25, 6–20.
- 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.

Smith, C.L., 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.

Section 2: Studies of Comparability Methods

Chapter 2: Summary of the Online Comparability Studies for One State's End-of-course Program

Susan M. Lottridge
W. Alan Nicewander
Howard C. Mitzel
Pacific Metrics

Acknowledgements

The authors thank the research and state members of the Enhanced Assessment Grant for their guidance and feedback on this study. In particular, the authors thank Phoebe Winter for her continued support and direction and Laura Kramer for working so closely with them and providing detailed feedback throughout this process.

Introduction

This report presents a summary of online comparability studies conducted on behalf of North Carolina's Enhanced Assessment Grant. The primary focus of this summary and of the grant was on methods for evaluating comparability of test variations. Thus, this summary will focus primarily on the methodological approaches and issues encountered in the comparability studies. In particular, the use of a new method (propensity score matching) for conducting the comparability studies was evaluated as a possible alternative to a within-subjects design. Additionally, the results are presented in light of the methodologies used.

Five subject areas of one state's end-of-course program were examined in a series of comparability studies. The subject areas were Algebra I, English I, biology, civics & economics, and U.S. history. Courses were taken mostly by high school students, although 8th grade students took the Algebra I course. The reports for these studies are available and the references are listed in Table 1. As stated in *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2004), comparability studies must be conducted if tests are given in different modes, such as on paper and online.

Scores from the end-of-course tests are used to provide feedback to students, teachers, parents, and administrators on what students know and can do relative to the achievement level standards and relative to other students taking the test. Scores are reported in terms of scale score and whether the student meets one of four hierarchical achievement level standards. The test scores count a minimum of 25 percent of a student's grade in the course. In addition, students must score as proficient in a series of end-of-course tests to satisfy the state's high school exit standards.

Method

Two types of designs were used to evaluate the comparability of scores from the online and paper versions of the end-of-course tests.

1. Within-subjects design. Each study used a within-subjects design. In the within-subjects design, examinees took both an online and paper version of the test. Counterbalancing was designed into the study to control for order effects—schools were assigned to a test order (online first or paper first). For two subject areas (Algebra I and English I), counterbalancing worked successfully; the number of examinees was evenly distributed across test order conditions, and the samples appeared to come from the same population. For three subject areas (biology, civics & economics, and U.S. history), counterbalancing was not implemented in practice. For these subjects, the number of examinees was not evenly distributed across test order conditions, presumably due to school self-selection into test order conditions.
2. Between-subjects design. Between-subjects follow-up studies using propensity score matching (PSM) were conducted in three subject areas: Algebra I, English I, and biology. PSM identifies a sample taking the paper test that is comparable to the sample taking the online test. As a result, it does not require that students take a test in both modes. The purpose of the follow-up studies was to explore the feasibility of using propensity score matching as an alternative to requiring examinees to take two tests. For these three subject areas, results from comparability analyses from PSM were compared to results from analyses based on the more burdensome double-testing (within-subjects) design. In civics & economics, propensity scoring was used as a complement to the within-subjects design because of the problems encountered in counterbalancing. A propensity score matching was not used in U.S. history because the results of that matching (as well as an exact match method) did not produce reasonable results when compared to the within-subjects results. The poor matching results suggest that the covariates used were not capturing key differences between the paper and online groups.

The PSM data were used to analyze most aspects of comparability. The relationship between mode-based scores (such as correlations or agreement in achievement levels) was not examined with PSM data because these values could not be computed with between-subjects data. Rather, within-subjects data were used. In addition, further propensity score studies were conducted in Algebra I in an attempt to improve upon the method. Table 1 lists the types of comparability studies conducted for this grant and references to the reports written. More detail on each of the studies can be found in the reports.

Table 1. Types of Comparability Studies Conducted on the End-of-course Data

Subject Area	Within-subjects	Propensity Score	Counter-Bal. successful?	Reports
Algebra I	yes	yes	yes	Lottridge, Nicewander, and Mitzel (2008a) Lottridge, Nicewander, and Schulz (2008a) Lottridge, Nicewander, and Box (2008)
English I	yes	yes	yes	Lottridge, Nicewander, and Mitzel (2008b) Lottridge, Nicewander, and Schulz (2008b)
Biology	yes	yes	no	Lottridge, Nicewander, and Mitzel (2008c) Lottridge and Nicewander (2008a)
Civics & Economics	yes	yes	no	Lottridge and Nicewander (2008b)
U.S. History	yes	no	no	Lottridge and Nicewander (2008c)

Examinees

Examinees were students who were enrolled in one of the designated courses and who were required to take the state's end-of-course test in spring 2007 in one of the five subject areas. The study sample was limited to students who

- commonly took that course (e.g., 10th and 11th graders)
- had tests with no administration issues
- attempted at least one item on the test
- had an online test with a valid administration date
- did not attend an alternative school

The full sample for the within-subjects study consisted of all students with two valid scores, one on the paper test and one on the online test. The sample for the PSM (between-subjects) study consisted of students who took the online test first (the treatment group) and a matched group of students, extracted from the population of students who took only the paper test (the control group).

For each subject area, Table 2 displays the number of schools that participated in the study, the number of examinees overall, the number of examinees taking the paper test first, the number of examinees taking the online test first, and the number of examinees in the propensity score matching study. In Algebra I, three propensity score studies were conducted:

- Study 1 conducted the matching of the paper and online samples across grades eight and nine, and did not impute missing data.
- Study 2 conducted matching separately for grades eight and nine and did not impute missing data.
- Study 3 conducted matching separately for grades and used mean imputation on missing data.

Study 2 produced results most similar to the within-subjects results, and these results were used throughout the remainder of this report.

Table 2. Sample Sizes and Number of Schools of Comparability Studies Conducted on the End-of-course Data

Subject Area	Total Schools	Within-subjects			Propensity Score Online First/Matched Paper Examinees ¹
		Total Examinees	Paper-first Examinees	Online-first Examinees	
Algebra I	49	2101	1202	899	Study 1: 741 *Study 2: 741 Study 3: 899
English I	25	1527	702	825	536
Biology	20	1004	194	810	764
Civics & Economics	19	951	272	679	579
U.S. History	17	938	317	621	n.a.

Note: n.a. = not applicable. * = Study results used in this report. ¹Students who took the online test first were matched to students who had only taken the paper-based test.

Procedures

The tests were administered in spring 2007 as part of operational testing for the end-of-course program. The state department of education enlisted volunteer schools to participate in the study. The department made a concerted effort to obtain a representative sample of the state. Schools' motivations for participating in the study were

- the opportunity for students to be tested twice with the highest test score counting
- preparation for the eventual use of statewide online testing
- a sufficiently long testing window
- entreaties by the department

Presumably, schools that chose not to participate were not comfortable with online testing or did not want the additional testing burden required by the within-subjects design. The state used random assignment at the school level to counterbalance test administration order. Schools with limited computer access were given the option of dividing their students into the two test administration orders.

A maximum of two weeks was allowed between test administrations. Students were told that the higher of the two test scores counted as the official test score in order to help foster similar motivation levels across administrations.

Test Instruments

Prior to conducting the comparability studies, the state had created multiple test forms to assess student achievement in each of the five subject areas. These forms were designed according to blueprint specifications aligned to state-specified goals for that course. All forms consisted of multiple-choice items.

The test forms were originally created for paper-based administration, and were later adapted for online administration. The online testing software presented items on the screen one at a time—a few items, with accompanying stimulus material, required scrolling. The examinee had the option of increasing or decreasing the font size, navigating around the test at will, and clicking on an answer choice or typing in a letter. Forms were spiraled in the paper-based and online administrations.

Table 3 displays the number of items on each test, the number of paper forms, the number of online forms, the number of forms administered in both modes, and whether forms had overlapping items. Because forms were spiraled in both administrations, some examinees took the same form in both modes. Because some forms shared items, some examinees took forms that shared items. The extent of form overlap was 20, 40, and 60 items, depending upon the subject area.

Table 3. Number of Test Forms in Each Mode and Form Overlap across Modes, for Each Subject Area

Subject Area	Test Length	Number of Paper Forms	Number of Online forms	Number of Forms Administered in Both Modes	Form Overlap?
Algebra I	64	5	5	5	yes
English I	56	6	6	6	no
Biology	88	3	1	1	no
Civics & Economics	80	5	5	4	yes
U.S. History	80	5	5	3	yes

Counterbalancing Analyses

The analyses of test order suggested that some schools did not adhere to their assigned test order. Because random counterbalancing is critical to control for a test order effect, it was important to investigate whether the schools that did not adhere to their assigned order differed from those who did. The order in which schools tested was verified by both a data-based analysis and by a survey of district or school test coordinators. Schools with an unverified or “mixed” test order were removed from the sample.

In Algebra I and English I, the previous year’s end-of-grade test scores in math and reading, respectively, were used to determine whether there were differences among the samples of schools that adhered to test orders, did not adhere to test orders, and schools not assigned a test order. Scaled score and standardized scaled score mean differences were analyzed across the three groups. If the differences varied significantly across the three groups, then it was assumed that the samples did not come from the same population. Based on these analyses, only schools that adhered to their assigned test order were retained for analyses in Algebra I. Schools that adhered to their assigned test order or were not assigned a test order were retained for analyses in English I. In biology, civics & economics, and U.S. history, examinee participation rates in the test administration orders suggested that counterbalancing was not implemented as expected. For these subject areas, further counterbalancing analyses were not conducted.

Analysis of Form Status

Examinees in the within-subjects design took either the same form, forms that overlapped on some items, or a completely different test form. For each subject area, within-subjects (mode) and between subjects (test order, form overlap status) ANOVAs were conducted to determine if the mean end-of-course scaled scores differed across these factors. These analyses were undertaken primarily to examine whether there was a practice effect due to the degree of form overlap. In all subject areas, there was either a three-way interaction (English I, Algebra I) between mode, test order, and form overlap status or a two-way interaction (biology, civics & economics, U.S. history) between form overlap status and test order. The statistical significance indicated that form overlap was an important effect and that larger mean differences between modes were associated with higher degrees of overlap. These results are consistent with a practice/memory effect. As a result, the analyses were disaggregated by the extent of form overlap.

Propensity Score Matching

Propensity score matching (PSM) was investigated as an alternative method of evaluating score comparability of the online and paper tests. Such a procedure allows a state to use the data from the regular online and paper administrations rather than administering the two types of tests to the same sample of students. The results from the online administration are compared to results from a matched sample from the regular paper administration.

PSM (Rosenbaum & Rubin, 1983) attempts to predict group membership using logistic regression, with the covariates as predictors. A one-zero variable, indicating whether a person is in the treatment (online) or control (paper) group, is regressed on the covariates using logistic regression. The so-called propensity score for each person is the probability of being in the treatment group. Each person in the treatment group is matched to an accompanying person in the control group using a “nearest neighbor” (in terms of the propensity score) from the control group. Once the PSM procedure is complete, the two matched groups can be compared on one or more independent variables.

PSM represents an improvement over exact matching methods because it

- contains exact matching information in the propensity score (e.g., equal propensity scores suggest an exact match)
- allows for matching on a single value, the propensity score
- uses statistical significance to identify characteristics that predict group membership
- allows the statistical model to determine closest matches when exact matches are not possible

As with any other matching method, the quality of the match is determined by the availability and quality of matching variables. Weaknesses specific to PSM are that there are several steps in the matching procedure, and it can be difficult to determine the quality of the match without comparing the matched groups on key variables.

Propensity score matching was used as a follow-up to the Algebra I, English I, and biology within-subjects studies. It was used as a complement to the within-subjects design in civics & economics because counterbalancing was not implemented as anticipated. In Algebra I, PSM was studied more in depth by separating the matching by grade and by imputing missing data. More details on PSM as applied to these studies can be found in Lottridge, Nicewander, and Schulz (2008a), Lottridge, Nicewander, and Schulz (2008b), Lottridge, Nicewander, and Box (2008), Lottridge and Nicewander (2008a), and Lottridge and Nicewander (2008b).

Sample Characteristics

Characteristics of the within-subjects sample were compared with characteristics of the examinees who did not participate in the study and took only the paper test. Individual, school, and test administration characteristics were compared on a large number (14 to 17) of characteristics. Individual characteristics compared were sex, ethnicity, grade level, free lunch status, LEP status, exceptionality, and related test scores (if available). School characteristics compared were school type, Title I status, region of state, and wealth rank of the county. Test administration characteristics compared were testing cycle, amount of make-up testing, and accommodations. The purpose of these comparisons was to provide evidence that the sample results and conclusions can be generalized to the entire population.

The within-subjects samples were similar to the population on most variables. The differences are outlined in Table 4. The major differences across subject areas were that examinees in the study sample came from different regions of the state and generally poorer counties. Additionally, the online sample tended to include more charter school examinees than the paper sample.

Table 4. Characteristics (of 14–17 Included in the Comparison) in Which the Sample Differed from the Population in the Five Subject Areas

Subject Area	Characteristics
Algebra I	<ul style="list-style-type: none"> • Eighth graders, academically/intellectually gifted, and students enrolled in charter schools were over-represented in the sample. • Ninth graders, students receiving free lunch, students not identified as exceptional, students enrolled in 'regular' schools and in non-Title I schools were under-represented in the sample. • Students in sample had a moderately higher mean computer skills score. • Students in the sample came from proportionally different areas of the state and came from slightly poorer counties.
English I	<ul style="list-style-type: none"> • Males, blacks, students receiving free lunch, students enrolled in charter schools and in schools with schoolwide Title I status were over-represented in the sample. • Females, whites, students paying full price for lunch, and students enrolled in "regular" schools were under-represented in the sample. • Students in the sample had slightly lower mean reading, computer skills, and math scores. • Students in the sample came from proportionally different areas of the state and came from poorer counties.

Subject Area	Characteristics
Biology	<ul style="list-style-type: none"> • Tenth graders, students receiving temporary free lunch, “non-exceptional” students, and students enrolled in charter schools were over-represented in the sample. • Eleventh graders, students paying full price for lunch, and students enrolled in “regular” schools were under-represented in the sample. • Students in the sample came from proportionally different areas of the state and came from poorer counties.
Civics & Economics*	<ul style="list-style-type: none"> • Tenth graders, students receiving free lunch, non-LEP students, “non-exceptional” students, students enrolled in charter schools, and non-Title I schools were over-represented in the sample. • Students paying full price for lunch and students enrolled in “regular” schools were under-represented in the sample. • Students in the sample came from proportionally different areas of the state and came from poorer counties.
U.S. History	<ul style="list-style-type: none"> • Females, 11th graders, whites, “non-exceptional” students, and students enrolled in charter schools and in non-Title I schools were over-represented in the sample. • Students paying full price for lunch and students enrolled in “regular” schools were under-represented in the sample. • Students in the sample had slightly lower mean reading and math scores. • Students in the sample came from proportionally different areas of the state and came from poorer counties.

*The sample taking the online test first was used for this set of comparisons.

Overall Methodology

Tables 5 and 6 outline the methods used in the comparability study designs, as well as the impact of those methods on interpretation of results. Table 5 displays methods and impacts relative to external validity evaluations, and Table 6 displays this information relative to internal validity evaluations. Sampling, instrumentation, administration, and scoring were used as organizers in considering the impact of design on validity issues.

The major external validity issues related to the volunteer sample. Volunteer samples are the only feasible method for conducting studies in educational research, but self-selection in participation weakens the external validity argument. The comparison of the sample and general population yielded some differences, particularly related to the region of the state and the wealth of the counties. The major internal validity issues related to the lack of successful counterbalancing in three subject areas, and the small sample size for each form, which precluded IRT analyses. IRT analyses would have enabled the comparison of test characteristic curves and number correct to scaled score tables. These comparisons provide the most detailed evidence of score comparability. However, the comparability studies, as conducted, adequately addressed each of the issues outlined in Tables 5 and 6.

Table 5. Potential Effects on the External Validity of the Mode Comparability Studies

Category	Design	Potential Impact on External Validity
<i>Sampling</i>		
Identifying the Population	Sample restricted to “typical” students taking the course and test in spring 2007.	Results may not generalize to “non-typical” students.
Obtaining the Sample	Schools volunteered to participate in the study. <ul style="list-style-type: none"> • Reasons to participate: <ul style="list-style-type: none"> ○ higher test score used ○ practice for future online testing ○ entreaties by state department • Reasons not to participate: <ul style="list-style-type: none"> ○ resources to test students twice ○ lack of technology resources for online testing 	Because schools were not chosen randomly, schools that participated may differ from the population. For example, schools confident in their technology use may be more likely to participate than schools that are not confident.
Sampling Unit	Sampling was at school level. Schools were randomly assigned test order.	Sampling at the school level may create non-equivalent groups because it can be difficult to obtain school-based samples that exactly match on key characteristics.
Sample Participation	Most schools that were assigned a test order chose to participate. Some schools did not adhere to assigned test order. Reasons for non-participation were not gathered. Schools with unverified test order were removed from the sample. In some cases, schools that did not adhere to their assigned test order were removed.	Self-selection in participation and adherence to assigned test order may result in a sample unlike the population. For instance, schools may choose not to participate if they were assigned a test order that they felt disadvantaged their students. Or, schools may ignore their assigned test order and choose their own test order.
Similarity to Population	The sample and population were compared on a large range of student, test, administration, and school characteristics.	Differences between the sample and population may reduce generalizability. Overall, the samples were reasonably close to the population. Primary differences were distribution among regions of the state and mean county wealth rank.
<i>Instrumentation</i>		
Instruments Used	Instruments studied were operational test forms.	None.
<i>Administration</i>		
Administration Conditions	Tests were administered as part of operational testing. The higher of two test scores counted for the final score. A maximum of two weeks were allowed between paper and online administrations.	Impact of taking two tests is unknown. Counterbalancing should average out the order effect.
Online Experience	Students in the study did not receive training to take the online test. However, many students had the opportunity to take a computer skills test in 8 th grade.	It is possible that the state may offer training in the future, and this may affect the generalizability of the study results.
<i>Scoring</i>		
Scores Used	The analyses examined raw scores, scaled scores, achievement levels, and proficiency categories.	None.
Disaggregation	Mean scaled scores were disaggregated by subgroups such as sex and ethnicity to determine whether differential mode effects existed.	If differential mode effects are identified for subgroups, this may affect generalizability to these subgroups.

Table 6. Potential Effects on the Internal Validity of the Mode Comparability Studies

Category	Design	Potential Impact on Internal Validity
<i>Sampling</i>		
Group Assignment	Some schools did not adhere to their assigned test order. For some subject areas, the schools that were assigned to administer the paper first participated in lower numbers than schools that were assigned to administer the online test first.	For subjects in which counterbalancing was not successful in practice, the results were disaggregated by test order. This confounded the order and mode effects.
Sample Size	The samples were edited to remove unverified schools. In some subject areas, the analyses were disaggregated by test order, form, and degree of form overlap (see Instrumentation).	The removal of examinees and disaggregation reduced the sample size sufficiently that some analyses (such as IRT-based analyses) could not be conducted.
Matching	Propensity score matching was used to identify a matched sample in the population to the online sample. The technique is new in comparability studies, and its feasibility is yet unknown.	Propensity score matching can only produce matches on covariates entered into the model. If these covariates do not adequately represent the sample and population then the two samples may differ in key ways, influencing the study results. The results of these studies suggest that PSM produces comparable samples.
<i>Instrumentation</i>		
Forms Used	Forms were spiraled within the online and paper conditions, and as a result some examinees took the same form in both modes. In addition, some forms shared items, and as a result some examinees took a subset of the same items in both modes.	The practice/memory effect for students taking the same form or forms with overlapping items is likely to be stronger than that for students taking different forms across modes. This effect can be controlled by disaggregating results across degrees of form overlap.
Online Experience	The online forms were paper forms adapted into the online environment. The online environment was developed to work similarly to the paper environment.	None.
<i>Administration</i>		
Administration Conditions	Schools were allowed a maximum of two weeks to administer the tests in both modes. Data were not collected to determine the time elapsed between the two administrations for each school and student.	Differences in the time elapsed between the test administrations may influence the extent of the order effect. For these studies, it was assumed to be random error.
Study Conditions	No major issues were reported with the online testing.	Issues related to computer-based testing (e.g., crashes) can be considered random error.
<i>Scoring</i>		
Scores Used	The paper-based number correct to scaled score tables were used to assign online responses a scaled score. This method assumes that the IRT item parameters and TCC of the online test are (within measurement error) the same as that of the paper test.	The use of the paper-based number correct to scaled score tables may obscure mode effects. However, if the mean and standard deviation of the scaled scores are the same across the paper and online groups, then it may be assumed that no mode effect has occurred at the scaled score level. If the mean and standard deviation of the scaled scores differ both statistically and practically, then a mode effect has occurred.
Score Comparisons	When no paper data were available, the online results were compared to expected (from IRT parameters) paper results in order to better understand the mode effect.	This approach helped to better understand the direction and magnitude (if any) of the mode effect.

Analyses

Chapter 1 describes four criteria that can be used to judge whether test variations are comparable: the extent to which the online test produces more valid inferences, addresses the same content standards, is equally as reliable as the paper test, and classifies examinees in the same manner. The first criterion (producing more valid inferences) is not relevant to the studies presented in this report; the online tests were simply computer-based versions of the paper tests. One could imagine that this criterion is relevant for other types of online tests. In particular, online tests that use innovative items in order to more accurately assess content standards may be examined against this criterion.

Using the figure from Chapter 1 regarding the content and score dimensions of comparability, the comparability examined in these studies represents the highest expectation of comparability. Because the online forms in this series of studies are online versions of paper forms, we expect to be able to judge comparability at the item level and at the raw score level.

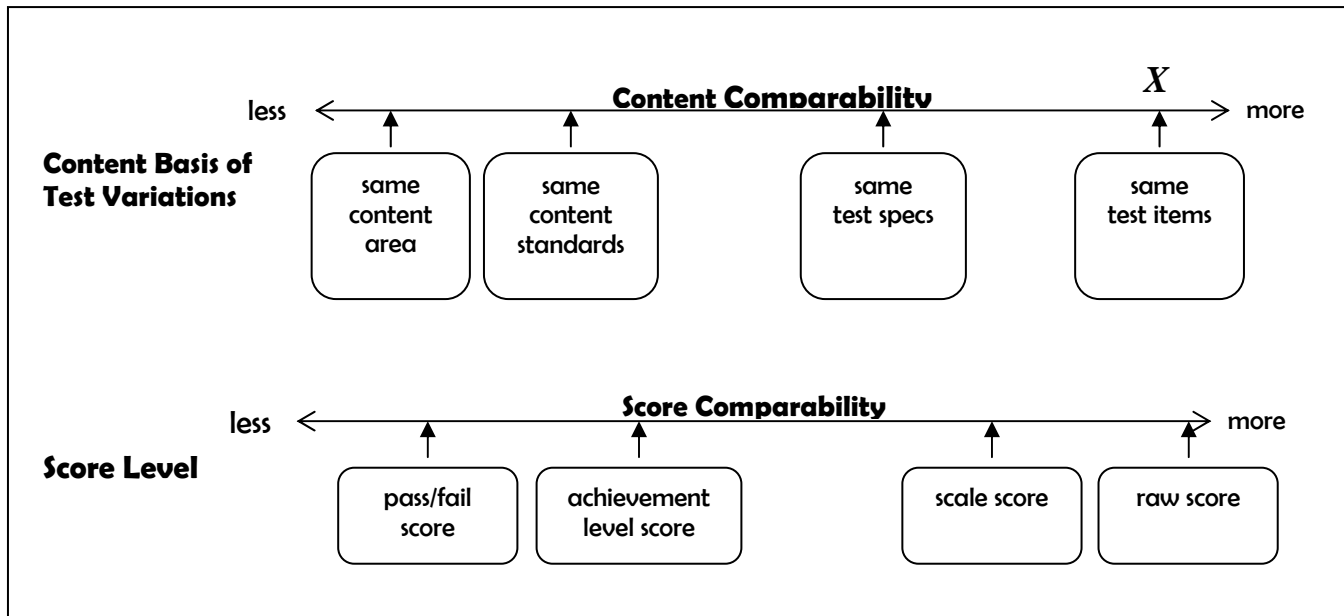


Figure 1. Comparability Continuum

Table 7 presents the three criteria, suggested analyses, and the analyses conducted in the five studies. Not all of the suggested analyses were conducted in these studies. Because the sample sizes across conditions (*i.e.*, across forms) were small, IRT analyses could not be conducted for four of the five subject areas. This meant that conditional standard errors of measurement, test characteristic curves, number correct to scaled score tables, and item level dimensionality tests could not be conducted. A large enough sample was available in the biology PSM study to conduct some IRT analysis; the test characteristic curves and mean item parameter estimates were compared across the modes.

Table 7. Study Design and Analyses by Three Comparability Criteria

Comparability Criterion	Suggested Analyses	Analyses Conducted
Assess Same Content Standards	test blueprint comparisons	yes
	expert review of item online modifications	no
	comparisons of different test formats	no
	tests of dimensionality at item level	no
	tests of dimensionality at test level	Algebra I, English I
Comparable Reliability	item level comparisons/DIF	yes
	overall reliability values	yes
Comparable Classification	conditional SEM/Test information	no
	frequency distribution of scores	yes
	measures of central tendency and dispersion	yes
	cumulative frequency distribution	no
	test characteristic curves	biology
	number correct to scaled score tables	no
	performance level distributions	yes
	correlation (corrected and uncorrected for unreliability)	yes
	agreement of assignment into performance levels	yes
	correlation (corrected and uncorrected for reliability) with variables	yes
ANOVA by subgroups (e.g., gender and race)	yes	

Results

The three relevant comparability criteria were used to organize the study results. In general, the results across the five studies suggest that the online and paper tests appear to be measuring similar content standards with the same level of reliability and are slightly more difficult than their paper counterparts. The difference in difficulty suggests that there is some construct irrelevance variance associated with the online version; however, the effect of that variance is small and does not appear in analyses aside from mean score differences. In essence, the results suggest that once equating is conducted to remove the difference in difficulty, the online and paper versions of the tests are comparable. The follow-up propensity score matching results mirrored the within-subjects results in Algebra I, English I, and biology, suggesting that PSM is a feasible method for identifying matched samples in comparability studies. A brief overview of the study results appears below for the three criteria.

The online and paper versions should address the same content standards.

The online test forms were paper forms adapted to the online environment. As a result, the online forms and paper forms had the same test blueprint specifications and were constituted of the same item text. The online environment was created to be similar to the paper environment, as described in the methods section.

The overall test factor structure was examined only for Algebra I and English I. These subjects were chosen because data from a previous year's subject test could be used to satisfy the model constraints. Additionally, the factor structure could only be compared in a within-subjects study. A confirmatory factor analysis (CFA) was conducted with increasing levels of constraints to test

whether the paper and online test were parallel, tau-equivalent, or congeneric. The results of the CFA supported the parallel test model.

Differential item functioning (DIF) analyses were also conducted in all studies to determine whether, controlling for ability, examinees were likely to perform better or worse on test items in one or the other mode of administration. The Mantel-Haenszel test was used, along with a classification of effect size used by the Educational Testing Service (ETS; Zwick and Ercikan, 1989). Very few items (less than 4.4 percent across all subject areas) were identified as exhibiting DIF. Follow-up PSM DIF studies produced the same result, although different DIF items were identified between the within-subjects and follow-up DIF study.

The online and paper scores should be equally reliable.

The internal coefficient of reliability (Cronbach's alpha) was calculated based on raw score data for each online and paper form in each subject area. In the within-subjects analyses for Algebra I, English I, biology, and U.S. history, and the PSM analysis conducted for civics & economics, the reliability values were very close, within .03 of one another. In the follow-up PSM analyses for Algebra I, English I, and biology, the reliability values were also very close, within .02 of one another.

The online and paper versions should classify students similarly.

The classification criterion covers a range of statistical analyses, including mean comparisons, cross-mode correlations, decision consistency analyses, and subgroup analyses. The results are presented for each of these analyses.

Mean Differences

Table 8 presents the mean scores in each mode for the five subject areas. In Algebra I and English I, the paper-first and online-first groups were combined because counterbalancing was implemented as intended. In biology, civics & economics, and U.S. history, counterbalancing did not work as intended. For biology and U.S. history, the data are presented separately for the online-first and paper-first students. Only propensity score matching was conducted for civics & economics. Only data from students taking different forms of the test across modes are included in the within-subjects analysis. Follow-up propensity score matching results are presented for Algebra I, English I, and biology.

Generally, mean differences across all forms in each subject area were small to moderate—less than 2 scale score points. The standardized mean differences ranged from -.25 to +.16. In Algebra I, English I, biology, and civics & economics, the online test was slightly harder than the paper test. The follow-up PSM mean differences were similar to the within-subjects results for Algebra I and English I. The follow-up PSM differences in biology were larger than were seen in the within-subjects mean differences, and suggested that the online test was harder than the paper test. In U.S. history, the mean online test scores were 1.4 points higher than the paper test scores when the paper test was administered first, and almost the same when the online test was administered first. Both the biology and U.S. history within-subjects results are consistent with a practice effect; the score of the second test was higher than the score of the test taken first.

Table 8. Scaled Score Mean and Standard Deviations and Differences by Subject Area and Study Design

Subject Area	Study Design	Paper Mean (SD)	Online Mean (SD)	Mean Difference (Online–Paper)	Standardized Mean Difference
Algebra I	WS:	154.61 (10.37)	153.34 (10.53)	-1.27	-0.12
	PSM:	155.05 (10.84)	153.99 (10.97)	-1.06	-0.10
English I	WS:	149.44 (8.46)	147.72 (8.40)	-1.72	-0.20
	PSM:	150.59 (8.64)	148.44 (8.09)	-2.15	-0.25
Biology	WS PF:	57.44 (6.53)	57.89 (6.31)	+0.45	+0.07
	WS OF:	56.47 (6.73)	55.95 (6.52)	-0.52	-0.08
	PSM:	57.34 (7.04)	55.77 (6.67)	-1.57	-0.22
Civics & Economics	PSM:	151.91 (8.98)	150.82 (8.09)	-1.06	-0.12
U.S. History	WS PF:	147.56 (8.80)	148.96 (9.19)	+1.40	+0.16
	WS OF:	149.92 (9.31)	149.85 (8.93)	-0.07	<-0.01

Note: WS=within-subjects; PSM=propensity score matching; PF=paper-first administration; OF=online-first administration

Cross-mode Correlations

The correlations between paper and online scaled scores were computed in the five studies, as were the agreements in achievement level assignment. The correlations of overall paper and online scaled scores ranged between .80 and .90 for all subject areas. The corrected correlations ranged from .87 to .97. Table 9 presents the correlation results for the five studies. Correlations were corrected by the average alpha reliability across forms in each mode.

Table 9. Corrected and Uncorrected Correlations of Paper and Online Scaled Scores

Subject Area	Design	Correlation	Corrected Correlation
Algebra I	test orders combined	.90	.97
English I	test orders combined	.82	.91
Biology	paper first	.84	.89
	online first	.84	.89
Civics & Economics	paper first	.80	.87
	online first	.88	.96
U.S. History	paper first	.89	.97
	online first	.87	.94

Decision Consistency

Decision consistency between the online and paper-assigned achievement levels was also evaluated and compared to expected decision consistency rates of two paper tests administrations. Table 10 presents the observed exact agreement of achievement levels assigned using the online and paper scaled scores. In Algebra I, the observed exact agreement rate was almost identical to the expected agreement rates. In English, the observed exact agreement rate was 9.4 percent lower than the expected rate. This difference was presumably due to mean differences between the online and paper scores. The biology observed exact agreement rates for the paper- and online-first samples were both similar to the expected agreement rate, suggesting that the mean differences were small. The civics & economics and U.S. history rates were much lower than the expected rates, and these differences were likely due to a combined mode-practice effect.

Table 10. Observed and Expected Exact Agreement of Achievement Level Assignment

Subject Area	Design	Observed	Expected
Algebra I	test orders combined	70.1%	70.8%
English I	test orders combined	65.4%	74.0%
Biology	paper first online first	71.5% 71.4%	74.3%
Civics & Economics	paper first online first	65.2% 68.4%	75.0%
U.S. History	paper first online first	66.5% 67.2%	73.5%

Table 10 presents the exact agreement of proficiency level classification (i.e., below proficient vs. proficient and above) using the online and paper scaled scores. In Algebra I, the observed exact agreement rate was almost identical to the expected agreement rates. In English, the observed exact agreement rate was 5.5 percent lower than the expected rate. Again, this difference was presumably due to the mean differences between the online and paper scores. Unlike the achievement level decision consistency rates, the biology observed exact agreement rates were lower than the expected rates. The civics & economics and U.S. history rates were much lower than the expected rates, and these differences were likely due to a combined mode-practice effect.

Table 11. Observed and Expected Exact Agreement of Proficiency Level Classification

Subject Area	Design	Observed	Expected
Algebra I	test orders combined	88.3%	87.7%
English I	test orders combined	84.5%	90.0%
Biology	paper first online first	82.1% 84.1%	90.2%
Civics & Economics	paper first online first	82.6% 84.7%	89.7%
U.S. History	paper first online first	84.9% 86.8%	88.9%

Relationship to External Measures

Correlations with various external measures were computed and compared for the paper and online scaled scores. If the online and paper versions are testing the same content standards, then their correlations with other measures should be similar. In Algebra I and English I, the external measures used were a computer skills test administered in 8th grade, an 8th-grade math test, and an 8th-grade reading test. In biology and U.S. history, the external measures used were a 10th-grade algebra test, a 10th-grade English test, and the student's anticipated grade in the course. In civics & economics, an 8th-grade reading test and the student's anticipated grade in the course were used. Table 12 presents the correlations. None of the differences in correlations across mode were significant at the .05 level.

Table 12. Similarity of Paper (Online) Correlations with External Measures

Subject Area	Study Design	Computer Skills Test	Math Test	Reading/English Test	Anticipated Course Grade
Algebra I	test orders combined PSM	.68 (.68)	.84 (.84)	.69 (.70)	
		.72 (.69)	.86 (.86)	.72 (.70)	
English I	test orders combined PSM	.59 (.61)	.71 (.71)	.77 (.79)	
		.66 (.64)	.76 (.74)	.83 (.79)	
Biology	paper first online first PSM		.55 (.58)	.69 (.69)	.61 (.64)
			.61 (.59)	.73 (.67)	.57 (.54)
			.65 (.60)	.67 (.68)	.48 (.52)
Civics & Economics	PSM			.74 (.71)	.57 (.56)
U.S. History	paper first online first		.55 (.56)	.66 (.60)	.49 (.45)
			.48 (.48)	.59 (.66)	.54 (.54)

Note. Number in parenthesis is online correlation. * $p < .05$.

Subgroup Differences

Analyses of variance were conducted in each study to determine whether there was a differential mode effect for members of subgroups. The subgroups examined were sex, ethnicity, grade, free lunch status, LEP status, and exceptionality status. Analyses were conducted only for subgroups with two or more levels with more than 30 examinees. Within-subjects ANOVAs were conducted in the within-subjects studies and between-subjects ANOVAs were conducted in the propensity score studies. Within-subjects ANOVAs are more sensitive to differences, and thus are likely to produce statistically significant results. The between-subjects ANOVAs were conducted on a smaller sample (online first examinees) and so significance was less likely to be detected and fewer groups were examined due to samples sizes being below 30 in subgroup levels.

Table 13 displays the results of the ANOVAs for each subject area, study design, and subgroup. “Yes” in a cell means that statistical significance was detected for that subgroup and indicates a possible differential mode effect. “No” means that statistical significance was not detected and indicates no mode effect. Very few differential mode effects were detected overall. A differential effect for free lunch status was detected in Algebra I and civics & economics. A gender effect was detected for biology, and a differential effect for ethnicity was detected for U.S. history. The propensity score results were mostly similar for Algebra I, English I, and biology. In algebra, the within-subjects ANOVA detected a differential mode effect for free lunch status, but the PSM ANOVA did not. In biology, within subjects ANOVA detected a differential mode effect for free lunch status, but the PSM ANOVA did not.

Table 13. Existence of Mode Effects by Subgroup

Subject Area	Study Design	Sex	Ethnicity	Grade	Free Lunch Status	LEP	Exceptionality
Algebra I	WS	no	no	no	yes	no	no
	PSM	no	no	no	no	no	no
English I	WS	no	no	no	no	no	no
	PSM	no	no	n.a.	no	n.a.	no
Biology	CBT first	yes	no	no	no	n.a.	no
	PSM	no	no	no	no	n.a.	no
Civics & Economics	PSM	no	no	no	yes	n.a.	no
U.S. History	CBT first	no	yes	n.a.	no	n.a.	no

Note: yes = statistical significance detected at the .05 level; no = statistical significance not detected at the .05 level; n.a. = analyses not conducted due to small sample size

Limitations of the Studies

In general, there were four limitations to the comparability studies. First, the study used volunteer samples. While the use of volunteer samples is unavoidable in applied research, it is unclear whether schools that participated were different in some important way from schools that did not. A comparison of characteristics at the student, test, and school levels suggested that the study samples came from different regions of the state and from poorer counties. In addition, charter schools had a slightly higher representation in the sample than in the population. Second, the within-subjects study changes the nature of the test administration process because examinees take the test twice. The effect of taking the tests twice on the scores is unknown. Counterbalancing deals with the order effect by placing the order effect as an error occurring equally for both test orders. In the case of biology, civics & economics, and U.S. history, counterbalancing did not work effectively, and so these analyses were disaggregated by test order. Disaggregating the test results confounds the mode and order effect. Third, while the overall sample size was large, the sample size for each form was relatively small. The sample size was further reduced by schools not adhering to the assigned test order and by the fact that some examinees were assigned the same form. The small sample size meant that some analyses could not be conducted, such as analyses involving item response theory (e.g., calibration of online item parameters, computation of conditional standard errors), equating, or item level factor analyses. Fourth, the propensity score matching procedure did not produce completely matched paper and online samples. Differences may have been due to the matching variables employed. In addition, the propensity score matching produced unreasonable results in U.S. history. A follow-up analysis showed that exact matching also produced unreasonable results, suggesting that the covariates used in the matching were not adequate for the matching procedure. The covariate set in U.S. history included the 10th-grade English score as the achievement variable, and this variable had a fairly low correlation with the U.S. history scaled score (.59). It is possible that this achievement variable was not a sufficient proxy for U.S. history knowledge and skills. All other subject areas used an achievement variable with at least a correlation of .70 with the subject-area scaled score.

Conclusions

Overall, these analyses suggest that the paper and online scores are comparable using the three relevant comparability criteria. Scores in both modes appear to be measuring the same content standards, with the same level of reliability, and appear to be classifying examinees in mostly the same way. The mean online test scores were lower than the paper test scores, and more online examinees were placed in lower achievement categories. However, this mode effect appears to be pervasive across all test items, and does not appear to be influencing any particular subgroup. The difference in mean scores suggests that equating might be required to ensure that the scores are truly interchangeable.

The analyses also indicate that propensity score matching produced similar results and interpretations as the within-subjects studies for Algebra I, English I, and biology. These results suggest that propensity score matching is a viable procedure for identifying matched samples to be used in comparability studies when a proper covariate set (particularly, a strong achievement variable) is available.

References

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2004). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Cronbach, L.J. & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 283–302.
- Lottridge, S.M. & Nicewander, W.A. (2008a). Comparing computer-based and paper-based test scores in one state's End-of-Course biology program: Results using propensity score matching. Monterey, CA: Pacific Metrics Corporation.
- Lottridge, S.M., & Nicewander, W.A. (2008b). Comparing computer-based and paper-based test scores in one state's End-of-Course civics & economics program. Monterey, CA: Pacific Metrics Corporation.
- Lottridge, S.M., & Nicewander, W.A. (2008c). Comparing computer-based and paper-based test scores in one state's End-of-Course U.S. history program. Monterey, CA: Pacific Metrics Corporation.
- Lottridge, S.M., Nicewander, W.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. Monterey, CA: Pacific Metrics Corporation.
- Lottridge, S.M., Nicewander, W.A., & Mitzel, H.C. (2008a). *Comparing computer-based and paper-based test scores in one state's End-of-Course algebra I program*. Monterey, CA: Pacific Metrics Corporation.
- Lottridge, S.M., Nicewander, W.A., & Mitzel, H.C. (2008b). *Comparing computer-based and paper-based test scores in one state's End-of-Course English I program*. Monterey, CA: Pacific Metrics Corporation.
- Lottridge, S.M., Nicewander, W.A., & Mitzel, H.C. (2008c). *Comparing computer-based and paper-based test scores in one state's End-of-Course biology program*. Monterey, CA: Pacific Metrics Corporation.
- Lottridge, S.M., Nicewander, W.A., & Schulz, E.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. Monterey, CA: Pacific Metrics Corporation.
- Lottridge, S.M., Nicewander, W.A., & Schulz, E.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. Monterey, CA: Pacific Metrics Corporation.

Lottridge, S.M., Nicewander, W.A., Schulz, E.M., & Mitzel, H.C. (2008). *Comparability of paper-based and computer-based Tests: A review of the methodology*. Monterey, CA: Pacific Metrics Corporation.

Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.

Winter, P.C. (2009, in preparation). Introduction. In Winter, P.C. (Ed.), *Evaluating the comparability of scores from achievement test variations*. Washington, DC: Council of Chief State School Officers.

Zwick, R & Ercikan, K. (1989). Analysis of Differential Item Functioning in the NAEP History Assessment. *Journal of Educational Measurement*, 26(1), 55–66

Chapter 3: Evaluating the Comparability of English and Spanish Video Accommodations for English Language Learners

Stephen G. Sireci
Craig S. Wells
University of Massachusetts Amherst

Summary

A large Midwestern state developed an accommodation for English language learners that involved a video presentation of the test material on a large screen while a narrator read the test material aloud. Alternate versions of the video were developed where the narrator read in English, Spanish, or Arabic. The focus of our study was whether scores from standard and video accommodation conditions could be interpreted similarly. We analyzed data from three different tests in three grades: a grade three math test, a grade five science test, and a grade six social studies test. In each grade/subject area, the data we analyzed came from three test administration conditions: standard administration, English video accommodation, and Spanish video accommodation. The sample size for the Arabic video condition was too small for analysis.

Two types of statistical analyses were conducted. First, we analyzed the structure, or “dimensionality,” of the data from each administration condition. For scores from different test administration conditions to be “comparable” (i.e., interpreted similarly), the dimensionality of the data from each administration condition should be similar. Our analyses of structural similarity used an exploratory approach based on multidimensional scaling, and a confirmatory approach based on structural equation modeling. The second type of analysis we conducted was an analysis of differential item functioning (DIF). This analysis evaluated the difficulty of each item across administration conditions, after controlling for overall differences on the test across students who took the test under the different administration conditions. The statistical procedure we used to evaluate DIF was logistic regression (Zumbo, 1999).

Our research also used two different sampling strategies to separate differences due to the groups taking the test under each condition from differences due to the test administration condition itself (standard versus video). The first sampling strategy pulled multiple random samples from the standard group to gauge the variability in test structure and DIF due only to sampling error. The second sampling strategy pulled multiple random samples that were selected in such a way that the samples had the exact same score distribution as the video groups. The logic of this sampling strategy was to account for the lack of overlap in the total test score distributions from the standard and accommodated groups.

The results found that differences in the proficiency distributions across the standard and accommodated groups did affect findings regarding the similarity of test structure and item functioning. After accounting for such differences, the dimensionality of the tests from all grades was similar across the standard and video accommodation conditions. At the item level, we found some items that functioned differentially across administration conditions, even after matched samples were used; however, the magnitude of the difference tended to be small, and in

some cases, the items flagged for DIF were extremely easy under all administration conditions. Nevertheless, we noted that the largest numbers of items flagged for DIF (four in grade five science and six in grade six social studies) were flagged in the standard versus Spanish video comparisons. It is recommended that linguists and bilingual education specialists be recruited to help determine reasons why these items functioned differentially across administration conditions.

The full results of our report are presented later in this chapter. Table 1, below, provides a very brief summary of our interpretation of the results across all methods and sampling conditions.

Table 1: Summary of Invariance and DIF Conclusions

Grade, Subject	Comparison	Sampling Strategy						Common DIF Items
		Unmatched			Matched			
		MDS	CFA	# DIF Items	MDS	CFA	# DIF Items	
3 Math	Standard vs. Eng. Video	=	=	2	≈	=	2	1
	Standard vs. Span. Video	=	≠	3	=	=	2	1
5 Science	Standard vs. Eng. Video	≈	=	2	=	=	2	2
	Standard vs. Span. Video	≈	=	5	=	=	6	5
6 Social Studies	Standard vs. Eng. Video	≠	=	0	=	=	0	n.a.
	Standard vs. Span. Video	≠	≠	3	=	≈	4	3

Note: = signifies a conclusion of invariant test structure, ≈ signifies approximate invariance, and ≠ signifies a lack of structural invariance. See later in this chapter for full results.

Design and Implementation Recommendations

We identify four questions regarding the comparability of test scores from standard and video read-aloud accommodation conditions. In this section we provide brief responses to each of these questions. Readers are referred to the full report following this section for more details on our study and the results.

1. Does the test variation support more valid inferences about the target student population’s achievement than the general test does?

The video read-aloud accommodation was designed to make these tests more accessible to English language learners. Given that our analyses showed similar test structure and item functioning, we do not see the accommodation as a threat to valid test score interpretation. If students feel more comfortable taking the test with the video accommodation, it is likely to lead to more valid inferences with respect to their performance in math, science, and social studies. Of course, our analyses are limited to only one specific subject area in each grade in one state. However, the results suggest this accommodation leads to comparable scores that can be

interpreted in the same way that scores from the standard administration condition are interpreted.

2. Is the test variation aligned to content standards at least as well as the general test is?

The read-aloud video accommodation used the exact same test that was used in the standard administration, ; therefore the content standards-assessment alignment results would generalize perfectly to this administration condition.

3. Does the test variation classify students into achievement levels based on the same degree of knowledge and skills as the general test does?

Given that our analyses supported invariance with respect to test structure and item functioning, it is likely the video read-aloud accommodation does not have an effect on classification of students into performance levels. However, such classifications can be affected by a single score point, so any items flagged for DIF should be further inspected to ensure the DIF is not caused by the test administration condition.

4. Does the test variation provide equally reliable results for the target population as the general test does for the general population?

Our study did not focus on measurement precision, decision consistency, or decision accuracy, so we cannot speak directly to this question. However, we did note smaller factor loadings for some content areas for the Spanish video group, as well as smaller proportions of variance accounted for in the data for this group, and for all matched groups, in the multidimensional scaling analyses. These findings suggest more random error in the data for student groups that score very low on state assessments. If educational tests are too difficult for certain groups of students, it will be difficult to get an accurate and reliable indication of their true proficiency. Thus, reliability issues in testing ELLs are complex. It may be that when an ELL is classified as performing in a lower achievement level than a peer, that classification is reliable in the sense that if the student were tested again, s/he would end up in the same (i.e., lowest) achievement level. However, the low score could be due to the test being too hard for the student and the student's response reflecting guessing and other behaviors that would not replicate if a test of more appropriate difficulty were administered.

Evaluating the Comparability of English and Spanish Video Accommodations for English Language Learners

Standardized educational tests are used throughout the U.S. to assess the competencies of students and to meet the accountability demands of the No Child Left Behind (NCLB) legislation. Standardized tests ensure uniform content, test administration, and scoring procedures. However, in many cases not all students have equal access to an assessment due to limited proficiency in the English language. The *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999) acknowledged the problem of language proficiency interfering with proper measurement of students' knowledge, skills, and abilities, as illustrated in the following excerpt:

...any test that employs language is, in part, a measure of their language skills... This is of particular concern for test takers whose first language is not the language of the test... In such instances, test results may not reflect accurately the qualities and competencies intended to be measured. (AERA, et al., 1999, p. 91)

To address this concern, many states have proposed accommodations such as translated versions of a test into one or more languages (Sireci, 2007). However, translated tests cannot be assumed to be equivalent to the original versions, since translation may change the meaning of what is being measured and make the test, or some of its components, differentially easier or harder (Angoff & Cook, 1988; Hambleton, 2005; van de Vijver & Poortinga, 1997, 2005). Recently, a large Midwestern state developed an accommodation for students with limited English proficiency that did not involve translation of test booklets. This accommodation, described here as a read-aloud video, involved giving these students the standard test booklet, but the students watched a video that projected the booklet on a screen and read the test material aloud. The video focused on the test booklet and an arrow pointed to what was being read aloud. There were three versions of this video read-aloud accommodation. One version had the audio in English, one version had the audio in Spanish, and the other version had the audio in Arabic.

Our investigation involved analyzing data from the standard (non-accommodated) English, English read-aloud video, and Spanish-read aloud video administration conditions. The Arabic read-aloud video condition did not have enough students to justify statistical analyses. Our analyses involved three different grades and three different subject areas—grade three mathematics, grade five science, and grade six social studies. We focused on comparing the dimensionality, or structure, of the data across these three conditions, and on evaluating the statistical functioning of the items across the three conditions. Our primary purpose was to ascertain whether the structure of the data and the functioning of the items were relatively consistent across the administration conditions to support aggregation of the scores for accountability purposes. Key questions motivating the analyses were whether the construct was altered due to the accommodation (i.e., structural non-equivalence) and whether the difficulties of the items were altered by the accommodation (i.e., differential item functioning). A secondary purpose of the study was to investigate data analysis methods that can be used to evaluate test comparability issues that arise when tests are translated or test administration conditions are altered.

Method

Data

The data from this study came from a third-grade mathematics test, a fifth-grade science test, and a sixth-grade social studies test, from a large Midwestern state. Although each test contained a small number of constructed-response items, we focused only on the multiple-choice items. All tests were administered during the fall of 2005. The sample sizes for all groups of students are presented in Table 1.

Table 1. Sample Sizes for Analysis

Test Administration Condition	Grade/Subject		
	3/Math	5/Science	6/Social Studies
Standard Administration	119,008	122,251	125,482
English Read-Aloud Video	448	322	322
Spanish Read-Aloud Video	189	165	173
Total	119,645	122,738	125,977

As is evident in Table 1, the differences across the sample sizes for the three test administration conditions were huge, with the read-aloud video conditions representing less than one-half of 1 percent of the standard administration conditions. In considering our statistical analyses, we were concerned that these differences in sample size would have an effect on the results because statistics derived from the video conditions are likely to suffer from more estimation error due to the smaller sample sizes. Given that we were interested in the effect of the test accommodation, rather than the effect of sample size, we decided to select multiple random samples from the standard administration of equal size to the other groups.

A second concern we had about differences across the standard administration and read-aloud video groups was the large differences in overall test performance across the groups. Table 2 presents descriptive statistics for each group for the items analyzed in this study. Across all grade/subject areas, the standard administration group scored noticeably higher (on average) than the English video group, which scored higher than the Spanish video group. The magnitudes of the mean differences were greater than one standard deviation for the standard group/Spanish video comparisons.

Table 2. Descriptive Statistics for Standard and Video Read-aloud Groups

Grade/Subject	Group					
	Standard Administration		English Video		Spanish Video	
	\bar{X}	SD	\bar{X}	SD	\bar{X}	SD
3/Math	44.36	9.35	40.97	9.42	33.17	8.79
5/Science	22.68	6.86	18.53	6.06	14.45	4.37
6/Soc. Studies	27.70	9.08	21.43	6.55	17.13	5.78

Figures 1–3 present boxplots representing the proficiency distribution for the standard administration and English and Spanish video groups for each of the grades/subjects. As is obvious in these figures, the distributions were not the same, and there was very little overlap between groups of students who took the test under Standard conditions and those who took the test in the Spanish video administration condition. Our concern was that these large differences in proficiency could affect our invariance analyses due to statistical or response characteristics such as restriction of range or guessing.

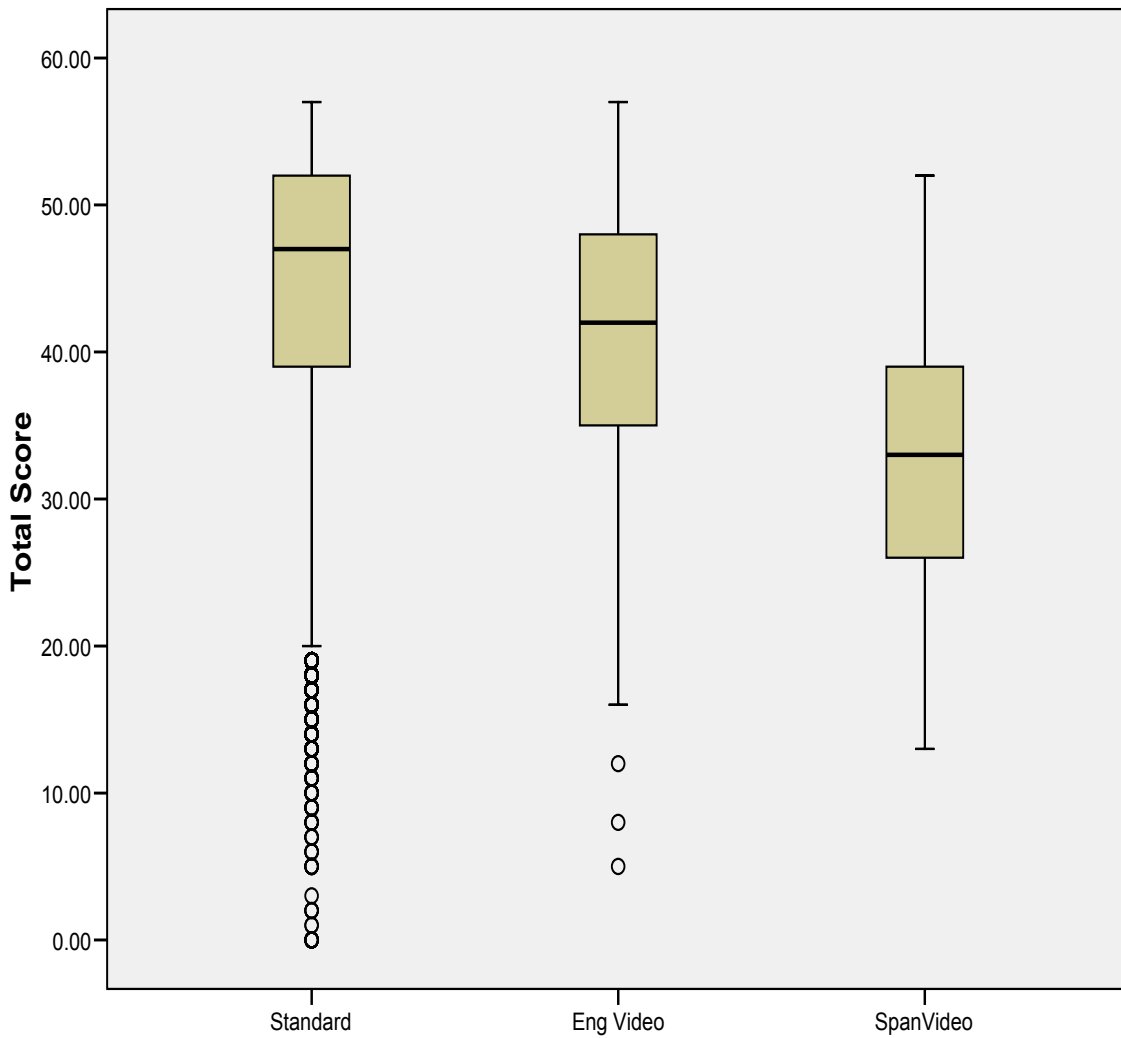


Figure 1. Boxplot Representing the Proficiency Distribution for the Grade Three Math Items

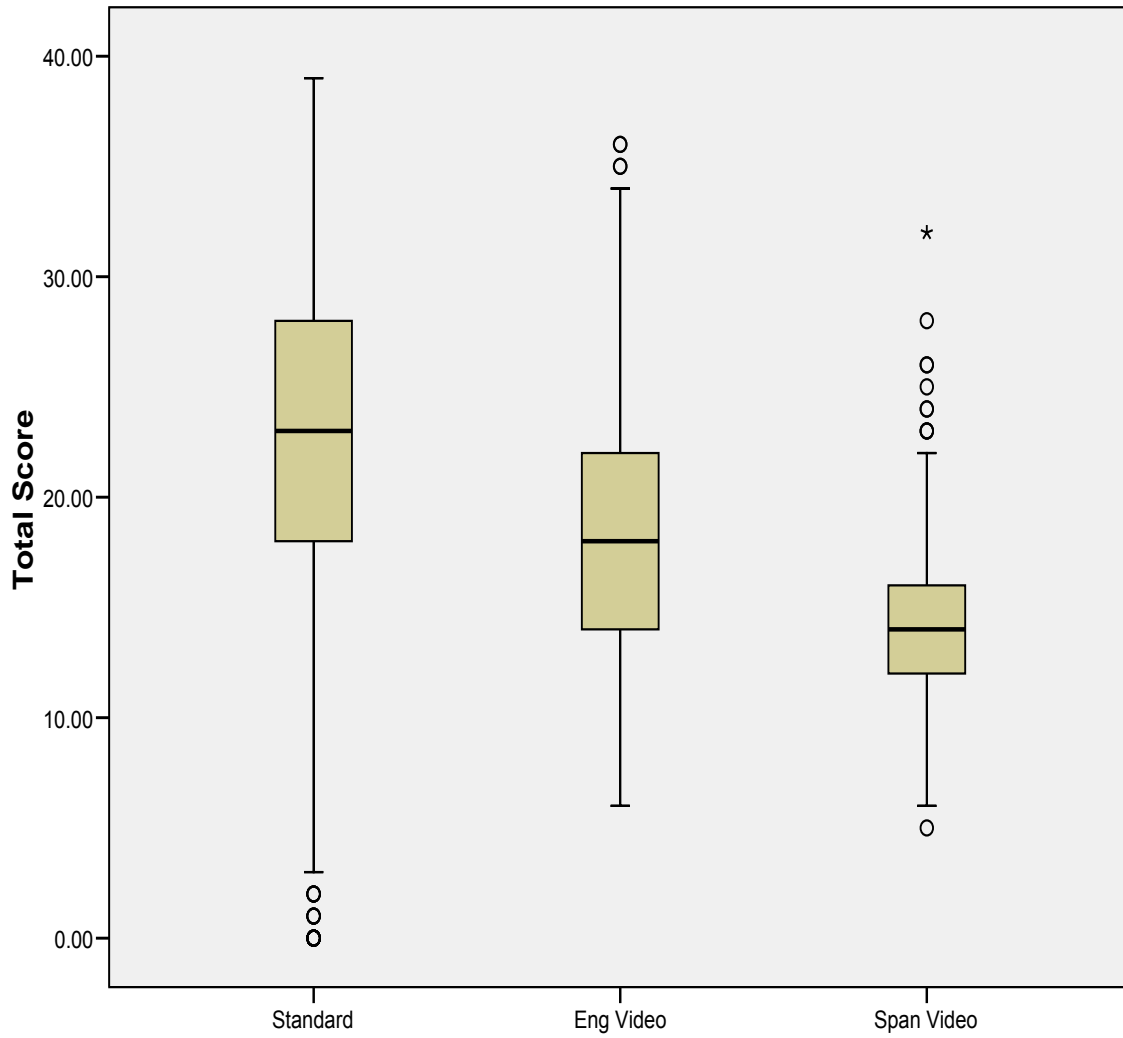


Figure 2. Boxplot Representing the Proficiency Distribution for the Grade Five Science Items

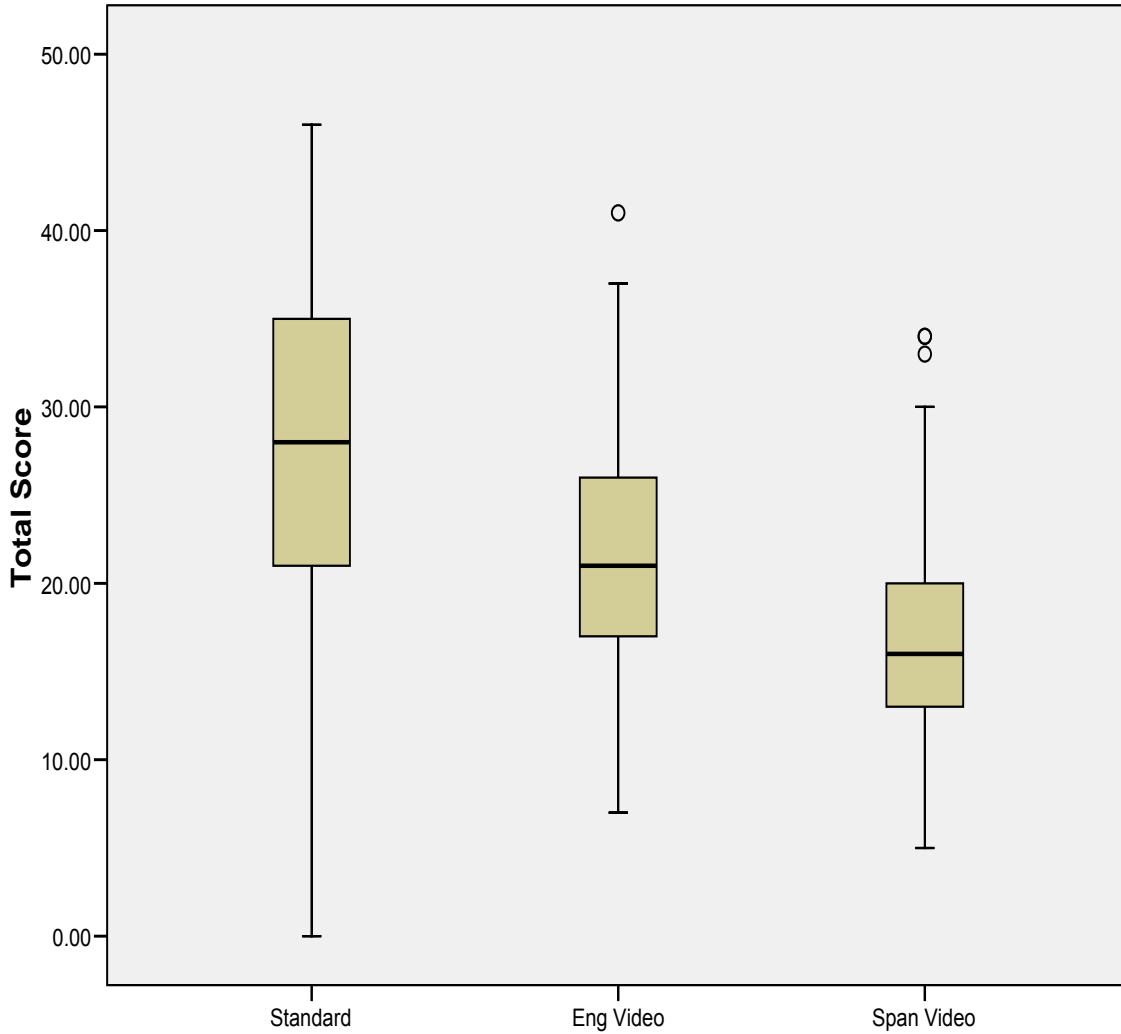


Figure 3. Boxplot Representing the Proficiency Distribution for the Grade Six Social Studies Items

Data Analyses

To evaluate the consistency of the test structure across the three different test administration conditions, we applied a multi-group exploratory dimensionality analysis and a multi-group confirmatory analysis. The exploratory analysis used weighted multidimensional scaling, while the confirmatory analysis used confirmatory factor analysis.

Weighted Multidimensional Scaling

The purpose of a multidimensional scaling (MDS) analysis is to find a spatial representation of data so that the distances between points in the multidimensional space correspond as closely as possible to the measurement properties of the original data. With respect to the evaluation of test

structure, MDS can be used to portray test items in a multidimensional space that best represents the similarities among the items as determined by students' responses to the items. Specifically, distances can be computed among items by using the Euclidean distance formula,

$$\delta_{jj'} = \sqrt{\sum_{i=1}^N (u_{ij} - u_{ij'})^2} \quad (1)$$

where j and j' are two different test items and $u_{ij} = 1$ if the response of examinee i to item j is correct and 0 if the response is incorrect, and N is the number of candidates in the dataset.

MDS is a nonlinear, exploratory procedure that can be used for single- or multi-group analyses. When applied to a single matrix (classical MDS), the matrix of observed item dissimilarity is modeled as accurately as possible in 1, 2, ..., or R -dimensional space. Here, the matrix of observed dissimilarity (distance) between item j and j' , was obtained from the student response data using the distance formula in Equation 1. The observed dissimilarity $\delta_{jj'}$ is modeled as

$$d_{jj'} = \sqrt{\sum_{r=1}^R (x_{jr} - x_{j'r})^2} \quad (2)$$

where R indicates the maximum dimensionality of the model, and x_{jr} is the coordinate for item j on dimension r .

In multigroup (weighted) MDS there is more than one inter-item distance matrix, which in our case refers to the matrices derived separately for the standard, English video, and Spanish video groups. These distance matrices are simultaneously fit using the INDSCAL model (Carroll & Chang, 1970), which incorporates a weight on each dimension for each matrix, as follows:

$$d_{jj'}^k = \sqrt{\sum_{r=1}^R w_r^k (x_{jr} - x_{j'r})^2} \quad (3)$$

where w_r^k corresponds to the weight associated with dimension r for group k .

The end result of a weighted MDS analysis is a multidimensional configuration that best fits the data for all groups when considered simultaneously and a matrix of group weights (with elements w_r^k) that represent how the group stimulus space should be adjusted to best fit the data for a particular group (k). The weights on each dimension for each group can be used to "stretch" or "shrink" a dimension from the simultaneous solution to create a multidimensional solution that best fits the data for a particular group. Thus, the weights (w_r^k) contain the information regarding structural differences across groups. If a similar pattern of weights is observed across groups, the same dimensions can be used to account for the test structure in each group. That is, similar weights across all language groups indicate structural equivalence, while differences

between group weights indicate structural differences across groups. Using simulated data, Sireci, Bastari, and Allalouf (1998) found that when structural differences existed across groups, one or more groups have weights near zero on one or more dimensions relevant to at least one other group.

Confirmatory Factor Analysis

Confirmatory factor analysis (CFA) was used to assess the measurement invariance at the content strand level between the standard group and the English and Spanish video groups, separately. Although we initially considered conducting these analyses at the item level, convergence could not be obtained. Problems associated with factor analysis of dichotomous item data are well known (Marsh, Hau, Balla, & Grayson, 1998; Yuan, Bentler, & Kano, 1997), so non-convergence of the item level data was not surprising. Drawing from the literature in this area (Bagozzi & Heatherton, 1994; Kishton & Widaman, 1994; MacCallum, Widaman, Zhang & Wong, 1999), we decided to parcel (group together) the items by content strand to investigate the consistency of a unidimensional model based on content strands across the groups. Parceling the items by content strand was the most logical way to parcel items from a construct perspective. To parcel the items we simply summed together the scores for items within each content area. The content areas and number of items within each area are presented in Table 3 for each grade level. Our use of MDS at the item level and CFA at the parcel level allowed us to investigate the consistency of the test structure at both levels of analysis. Since MDS implements a nonlinear model, it is better equipped to handle dichotomous data (Meara, Robin, & Sireci, 2000).

An example of a one-factor CFA model fit to the data is presented in Figure 4. This figure illustrates the model applied to the sixth-grade science test, which involved five indicators, each based on the subscores pertaining to the relevant content strand: history, geography, civics, economics, and inquiry. In this figure, λ represents the factor loading, δ represents measurement error, and ϕ indicates the variance for the latent construct.

Table 3. Content Strands and Number of Items per Strand

Grade/Subject	Content Strand	# Items
3/Math	number and operations	30
	measurement	18
	geometry	9
	total	57
5/Science	construct scientific knowledge	8
	reflect scientific knowledge	2
	use life science knowledge	11
	use physical knowledge	10
	use earth science knowledge	8
	total	39
6/Social Studies	history	10
	geography	10
	civics	10
	economics	10
	inquiry	6
	total	46

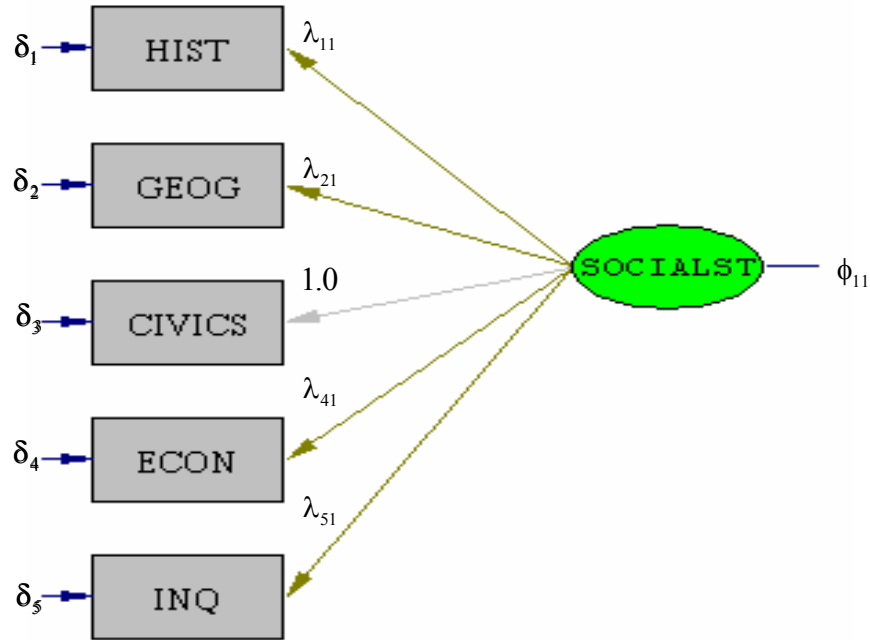


Figure 4. One-factor CFA Model to Assess Accommodated-related Measurement Invariance

The CFA model may be expressed as follows:

$$\mathbf{X} = \mathbf{\Lambda}_x \xi + \delta \quad (4)$$

where \mathbf{X} represents the vector of observed variables based on the content strand scores; $\mathbf{\Lambda}_x$ represents the vector containing the factor loadings; ξ represents the latent construct; and δ represents the vector of residuals. In our analyses, we were interested in determining if the factor loadings (λ 's) differed between the groups.

To test invariance of test structure at the content strand level between the respective groups, we compared the fit of a compact and augmented model. The compact model constrained the unstandardized factor loadings (λ 's) to be equal between the groups of interest (i.e., assuming measurement invariance). All but one of the unstandardized factor loadings in the augmented model were unconstrained (i.e., freely estimated in each group). One factor loading was fixed to 1.0 to define the scale of the latent variable in each group. Measurement invariance was assessed by comparing the overall fit of both models. The overall fit of the respective model was provided by the χ^2 statistic. Since the compact model is hierarchically nested within the augmented model, the difference between the χ^2 statistics ($\chi_C^2 - \chi_A^2$) is distributed as a χ^2 with the degrees of freedom equal to the number of parameters being compared. The software package LISREL (Joreskog & Sorbom, 2001) was used to perform the CFA based on the covariance matrix. The parameters were estimated using maximum likelihood.

Drawing random samples for the structural analyses. Given that disparate sample sizes and large differences in proficiency distributions across the video and standard groups can affect the results of structural analyses, we used two different sampling strategies. First, we drew 10 random samples from the standard administration group to mimic the sample size for the English video group and 10 random samples to mimic the sample size for the Spanish video group. We refer to these samples as random samples or “unmatched” random samples. Next, we drew another set of randomly drawn samples that were also matched to the English video or Spanish video group such that the proficiency distributions based on the raw scores were identical. The matched random samples were drawn from the standard administration group without replacement such that the raw score distribution was identical to that of the respective video group. For grade five science and grade six social studies, we were able to draw 10 matched samples for both the English video and Spanish video groups. However, for grade three math, the distribution of raw scores for the standard administration group was much more negatively skewed (i.e., there were fewer low-scoring students in this group relative to the video groups), so we were only able to extract four samples matched to the Spanish video group and five samples matched to the English video group.

Our first sampling strategy (random, but not matched) was used so that we could gauge the variability in results due only to sampling error. The second sampling strategy (random, matched) was used to gauge the effect that differences in the proficiency distribution could have on the analyses. Both strategies evaluate a potential factor that could affect the analysis that is *not* due to the test administration condition. After describing the analyses for detecting differential item functioning (DIF), we present a summary of the sampling conditions.

DIF detection. Logistic regression (Swaminathan & Rogers, 1990) was used to detect DIF between the standard and accommodated (video) groups. Logistic regression is a non-model based technique in that it does not require estimation of parameters from an item response theory (IRT) model, nor does it require an IRT model to fit the data. It is simple to implement, tests uniform and non-uniform DIF, and provides an effect size for the magnitude of DIF via a pseudo- R^2 statistic (Zumbo, 1999).

Logistic regression provides the probability of a correct response given an examinee’s proficiency level. When testing for DIF between a reference and focal group, the formulation of the logistic regression model may be specified as follows:

$$P(u_{ij} = 1 | \theta_j) = \frac{e^{[\tau_0 + \tau_1 \theta_j + \tau_2 g_j + \tau_3 (\theta_j g_j)]}}{1 + e^{[\tau_0 + \tau_1 \theta_j + \tau_2 g_j + \tau_3 (\theta_j g_j)]}}, \quad (4)$$

where $P(u_{ij} = 1 | \theta_j)$ represents the probability of correctly answering item i given examinee j ’s proficiency, denoted θ_j . The term g_j is a dummy code used to represent whether examinee j is in the reference ($g=0$) or focal ($g=1$) group. τ_2 represents the difference between the reference and focal group, controlling for proficiency level. τ_3 corresponds to the interaction between group

and proficiency, denoted $\theta_j g_j$. Uniform DIF is indicated by $\tau_2 \neq 0$ and $\tau_3 = 0$ while non-uniform DIF is represented by $\tau_3 \neq 0$, whether or not $\tau_2 = 0$.

Drawing random samples for DIF analyses. The purpose of the DIF analyses was to determine if the accommodation resulted in differential performance on the science items. To address the problem of comparing groups with widely disparate sample sizes and proficiency distributions, we again used two sampling strategies. Like the structural analyses, the first sampling strategy selected random samples of equal size to the English or Spanish video conditions, and the second sampling strategy selected random samples that were matched to the total score distributions for these groups. However, unlike the structural analyses, more than 10 samples were drawn for most conditions. Instead, the maximum number of samples that could be drawn without replacement was used. This strategy provided a much larger number of replications, and each item on the assessment was tested for DIF between the non-accommodated and the respective accommodated group over multiple replications (see Table 4 for specifics). We drew the maximum number of samples (without replacement) for these analyses, rather than only the 10 we used for the structural analyses, because it was less laborious to run the DIF analyses relative to the structural analyses and the DIF analyses involved hundreds of statistical tests. Thus, control over type I error rate was particularly important.

Criteria for flagging items for DIF. Logistic regression was conducted to test each item for DIF between the respective accommodation group and the corresponding non-accommodated sample using code written in the software package R. In addition to evaluating statistical significance, we also applied the logistic regression effect size guidelines proposed by Jodoin and Gierl (2001). An item was flagged as DIF in the presence of a statistically significant result and when the R^2 -change (i.e., R^2 due to DIF) surpassed .035, which signifies medium (the lowest category of non-negligible DIF) using the Jodoin-Gierl criteria. For each item tested, the proportion of replications in which the item was flagged as non-negligible DIF was recorded. The average R^2 -change for each item over replications was used to judge the magnitude of DIF.

Summary of Sampling Conditions

Given the complex sampling conditions and the different types of analyses conducted, a summary of the analyses and sampling conditions is in order. This summary is presented in Table 4. For the structural analyses, two statistical procedures were used, one based on an exploratory model fit to the item level data (MDS), the other based on a confirmatory model fit to the content strand level data (CFA). For grades five and six, 10 matched and 10 unmatched random samples pertaining to each of the English and Spanish video conditions were used for these analyses. For grade three, fewer matched samples were selected because there was less overlap of the distributions across the standard and video administration conditions.

Table 4. Summary of Analyses and Sampling Conditions

Grade/Subject	Analysis	Sampling Strategy	Number of Samples (Size of Sample)		
			Standard Administration	English Video	Spanish Video
3/Math	MDS	random, unmatched	20	1 (n=448)	1 (n=189)
	CFA		(10 n=448; 10 n=189)		
	MDS	random, matched	9		
	CFA		(5 n=448; 4 n=189)		
	DIF	random, unmatched	894		
		random, matched	(265 n=448; 629 n=189)		
5/Science	MDS	both*	20	1 (n=322)	1 (n=165)
	CFA		(10 n=352; 10 n=226)		
	DIF	random, unmatched	1,119		
		random, matched	(379 n=322; 740 n=165)		
	DIF	random, unmatched	162		
		random, matched	(109 n=322; 53 n=165)		
6/Social Studies	MDS	both*	20	1 (n=322)	1 (n=173)
	CFA		(10 n=322; 10 n=173)		
	DIF	random, unmatched	1,114		
		random, matched	(389 n=322; 725 n=173)		
	DIF	random, unmatched	185		
		random, matched	(152 n=322; 33 n=173)		

*10 samples were chosen for the unmatched condition, and 10 separate samples were chosen for the matched condition.

For the DIF analyses, the maximum number of samples that could be drawn to represent the sample size of the English or Spanish video group was drawn. Again, both matched and unmatched random samples were drawn. We went beyond drawing ten random samples where we could, but the number of matched samples drawn was limited by the numbers of students from the standard condition that overlapped with the video conditions. That limitation accounts for the fewer samples drawn in the matched condition relative to the unmatched condition. Our sampling strategies not only helped investigate specific factors irrelevant to the test accommodation condition (i.e., sample size and group proficiency differences), they also provided multiple replications so we could evaluate sampling error in general.

Results

Our analyses involved three different statistical procedures applied to three different subject area tests, under two different sampling conditions. Thus, there are a lot of results to summarize. We organize the presentation of the results by data analysis procedure, starting with grade three and ending with grade six. The results for the unmatched random samples are provided first, followed by the results of the matched analyses.

Multidimensional Scaling Results

For each grade and sampling condition, one- through six-dimension MDS models were fitted to the data⁹. As described earlier, for each sampling condition except the grade three matched condition, the MDS analyses involved 22 inter-item distance matrices—one computed from the data for the English video condition, one computed from the data for the Spanish video condition, 10 based on random samples from the standard condition with a sample size equal to that of the English video condition, and 10 based on random samples from the standard condition with a sample size equal to that of the Spanish video condition (see Table 4).

The first decision to be made in interpreting MDS results is selection of the most appropriate dimensionality of the data. There are two descriptive data-model fit statistics that can be used to help determine the most appropriate dimensionality—STRESS and R^2 . STRESS is a badness of fit measure with a value of zero indicating perfect fit. The R^2 index reflects the proportion of variance in the (transformed) item distance data accounted for by the model. Of course fit improves as dimensions are added to the model, and so the goal is to select the solution that appears to account for most of the variation in the data and yields an interpretable solution.

A summary of the fit results across all analyses is presented in Table 5. For all three grades, there was very little improvement in fit beyond two dimensions¹⁰. The two-dimensional solutions accounted for at least 42 percent of the variance in the data in all conditions, and adding additional dimensions yielded relatively small increments in improvement in variance accounted for. Therefore, we focus our results on the two-dimensional solution for each grade level and sampling condition. It is interesting to note that the fit for the unmatched samples was somewhat better than the fit for the matched samples, which may be a consequence of the relative restriction of range in that sampling strategy or simply less salient dimensionality in the data for relatively lower-performing students.

Analysis of Group Weights

In this section, we present the weights for the English video, Spanish video, and samples from the standard test administration condition for the two-dimensional solution for each condition. The results are presented separately for each grade. The weights are presented graphically and so a few words about how to interpret these graphs are in order.

⁹ The one-dimensional model was a replicated MDS analysis (Schiffman, Reynolds, & Young, 1981), whereas the other models were weighted MDS analyses.

¹⁰ We must qualify this conclusion for grade three in the matched-samples condition because convergence could not be reached for the 2D solution (unless negative dimension weights were permitted). Therefore, the subsequent discussion of the group weights uses a subspace from the 3D solution.

Table 5. Fit Results for MDS Solutions (Decimals Omitted)

Grade/Subject	Condition	Dimension	STRESS	R ²
3/Math	Unmatched	6	16	90
		5	18	89
		4	20	88
		3	23	86
		2	28	83
		1	36	77
	Matched	6	16	81
		5	18	80
		4	21	78
		3	25	76
		2	NC	NC
		1	41	59
5/Science	Unmatched	6	17	58
		5	19	58
		4	21	56
		3	25	56
		2	31	55
		1	40	55
	Matched	6	17	49
		5	19	49
		4	22	49
		3	27	47
		2	33	48
		1	42	52
6/Social Studies	Unmatched	6	18	59
		5	20	59
		4	NC	NC
		3	NC	NC
		2	35	54
		1	46	50
	Matched	6	18	43
		5	20	43
		4	23	43
		3	27	43
		2	35	42
		1	44	48

Note: 1D solutions are based on a replicated MDS analysis; NC = nonconvergence

As discussed in the method section, the weights for each group indicate the relative salience or importance of each dimension for accounting for the item similarity data in each group. When a group has a weight near zero on a dimension, that dimension does not account for much variation in the data for that group. What we inspect when looking at these graphs is the relative weighting of the dimensions, which can be thought of as the angle between the origin in each figure and the point represented by the group weights along each dimension. The distance of that point from the origin is less relevant, because it merely reflects the amount of variance accounted for in the data for the group when considering *both* dimensions. Weight points further from the origin indicate groups whose data are better fit by the two dimensions.

Grade Three Math: Unmatched Samples. Figure 5 presents the group weights for the grade three math test for the unmatched random samples. The weights (points) for each of the 22 matrices from the two-dimensional MDS analysis are portrayed. The weight point for the English video and Spanish video groups are represented by capital letters (E and S, respectively), while the lower-case letters represent the random samples taken from the standard group, with the former representing random samples of n=448 (e), and the latter n=189 (s).

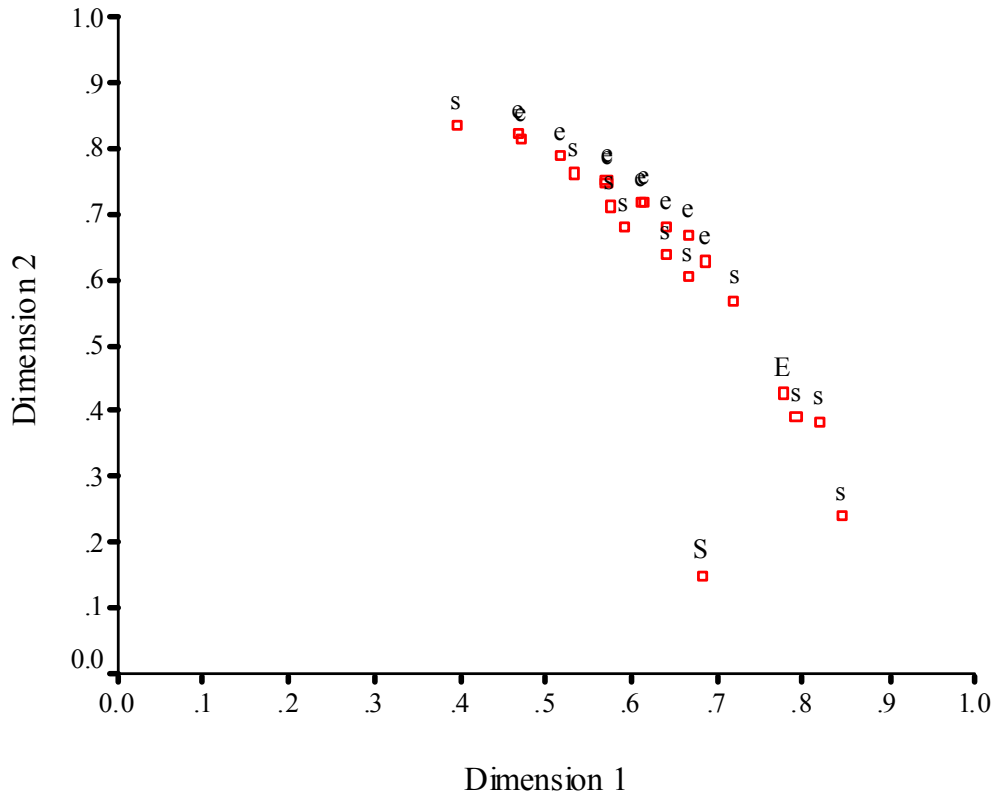


Figure 5. MDS Weights for English Video (E), Spanish Video (S) and Random Samples (e, s): Grade Three Math

As can be seen in Figure 5, all matrices had a large weight on the first dimension and were differentiated with respect to the second, minor dimension. However, it is clear there is a lot of sampling error along this second dimension, particularly for the smaller random samples based on the sample size for the Spanish video group. The English video group, has roughly equal weights on each dimension. The Spanish video group has the lowest weight on Dimension 2, but its weight is close to that of one of the random samples. These results suggest there is a good deal of sampling error operating, but all samples had a relatively large weight on the first dimension.

Grade Three Math: Matched Samples. Figure 6 presents the weight space for the matched samples for the Grade 3 Math test¹¹. As noted earlier, fewer matched samples could be drawn, which makes it more difficult to judge the sampling error. The random samples matched to the English video condition had large weights on Dimension 1 and small weights on Dimension 2. The random samples matched to the Spanish video condition, and the Spanish video group, exhibited the reverse pattern. The English video group had approximately equal weights on both dimensions. The fact that the Spanish video group had similar weights to the random samples matched to its distribution suggests structural equivalence across the standard and Spanish video conditions. It is unclear why the random samples matched to the English video group and those matched to the Spanish video group exhibited a different pattern of dimension weights, after drawing the matched samples.

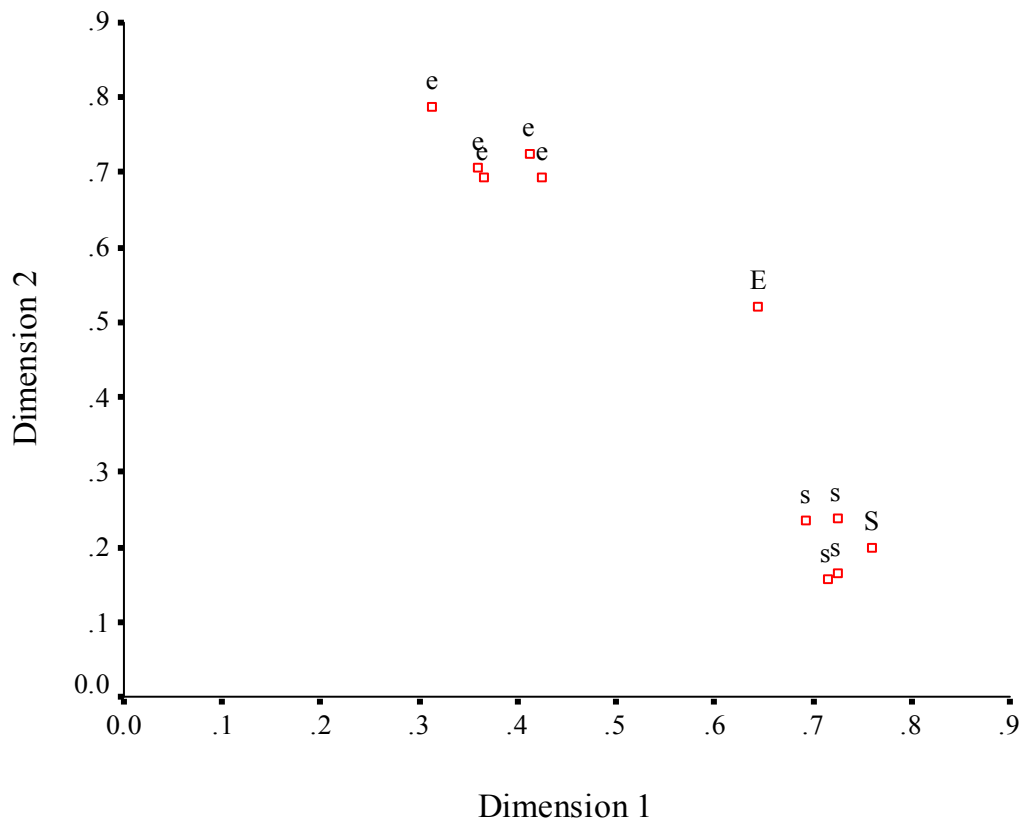


Figure 6. MDS Weights for English Video (E), Spanish Video (S) and Matched Random Samples (e, s): Grade Three Math

¹¹ This weight space represents the first two dimensions from the three-dimensional solution because the two-dimensional solution failed to reach convergence.

Grade Five Science: Unmatched Samples. Figure 7 presents the endpoints of the weight vectors for the 22 samples from the two-dimensional solution for the grade five science test using the unmatched random samples. The Spanish video group (S) is the closest to the origin, which indicates relatively poor fit compared to the other samples. In fact, the solution accounted for only 8 percent of the variation in the data for the Spanish video group, compared with 31 percent for the English video group. For the other 20 samples, the lowest variance accounted for was 48 percent. However, with respect to the dimension weights, both the English and Spanish video groups had roughly equal weights on both dimensions. The random samples had large weights on both dimensions, but although there was some variation in the weights for these samples, all had a higher weight on Dimension 1. These results suggest a dominant dimension underlying the data, but the dimension did not account as well for the variation in the data for the two video groups.

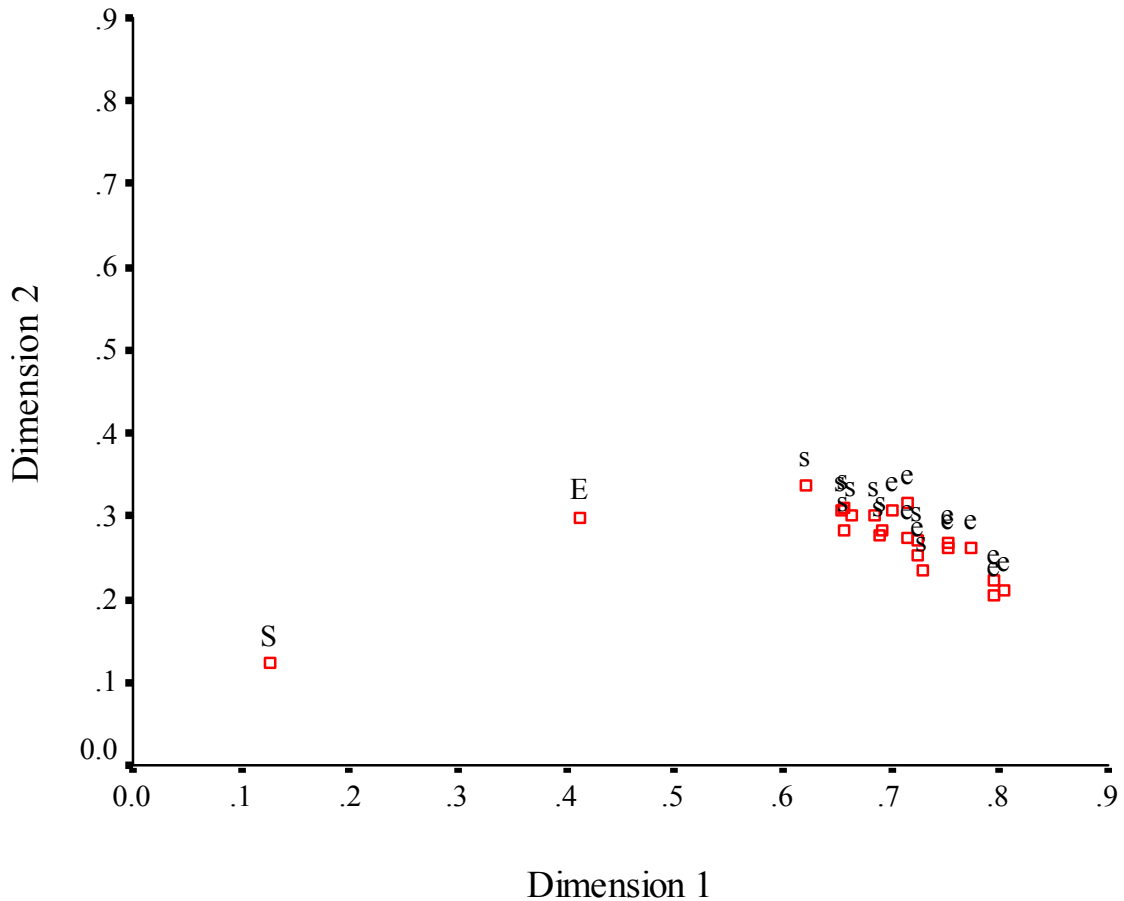


Figure 7. MDS Weights for English Video (E), Spanish Video (S) and Random Samples (e, s): Grade Five Science

Grade Five Science: Matched Samples. The weight space for the matched random samples for grade five science is displayed in Figure 8. After matching the proficiency distributions, all of the groups had a higher weight on the first dimension and the relative weighting of the dimensions for the Spanish video and English video groups appeared similar to those from the random samples. Again, the Spanish video group weight was closest to the origin, which indicates it was the group that was fit least well by the two-dimensional solution.

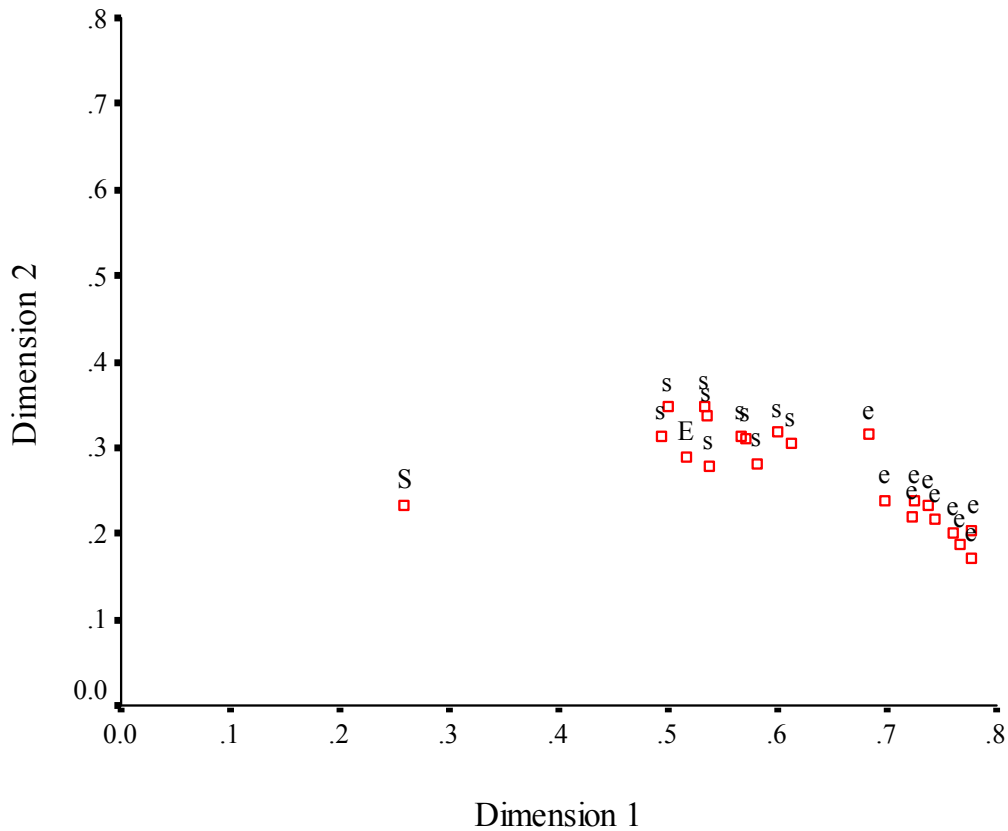


Figure 8. MDS Weights for English Video (E), Spanish Video (S) and Matched Samples (e, s): Grade Five Science

Grade Six Social Studies: Unmatched Samples. The weight space for grade six social studies for the unmatched random samples is presented in Figure 9. The Spanish video group has a weight near zero on the first dimension, and the English video group also has a small weight on this dimension. In contrast, the unmatched random samples all have a large weight on Dimension 1. This analysis suggests totally different dimensions are needed to characterize the variation in the data from the standard administration and video administration groups.

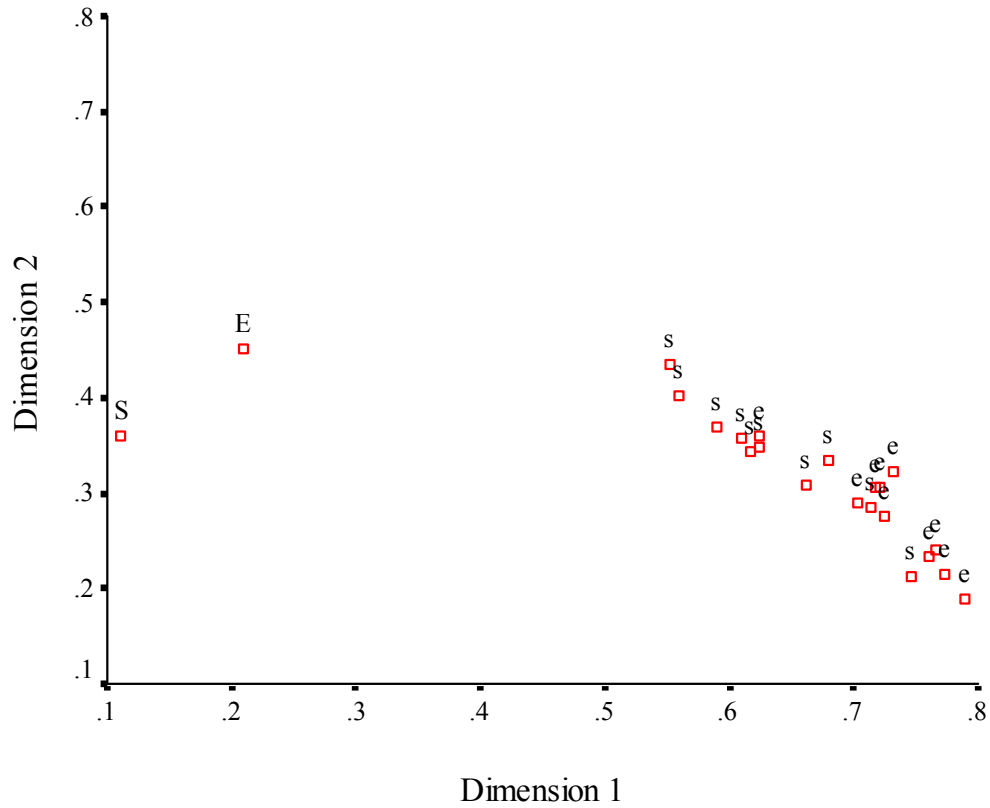


Figure 9. MDS Weights for English Video (E), Spanish Video (S) and Random Samples (e, s): Grade Six Social Studies

Grade Six Social Studies: Matched Samples. The weight space for grade six social studies based on the matched random samples is presented in Figure 10. The result is dramatically different than that observed in the unmatched analyses. All groups had approximately equal weights on both dimensions. The weight point for the English video group is indistinguishable from the random samples matched to its proficiency distribution. The weight point for the Spanish video group is similar to those of the random samples matched to its proficiency distribution. It appears that the lack of structural equivalence across test administration noted in the unmatched analysis is due not to the test administration condition, but rather to the large proficiency differences across the groups.

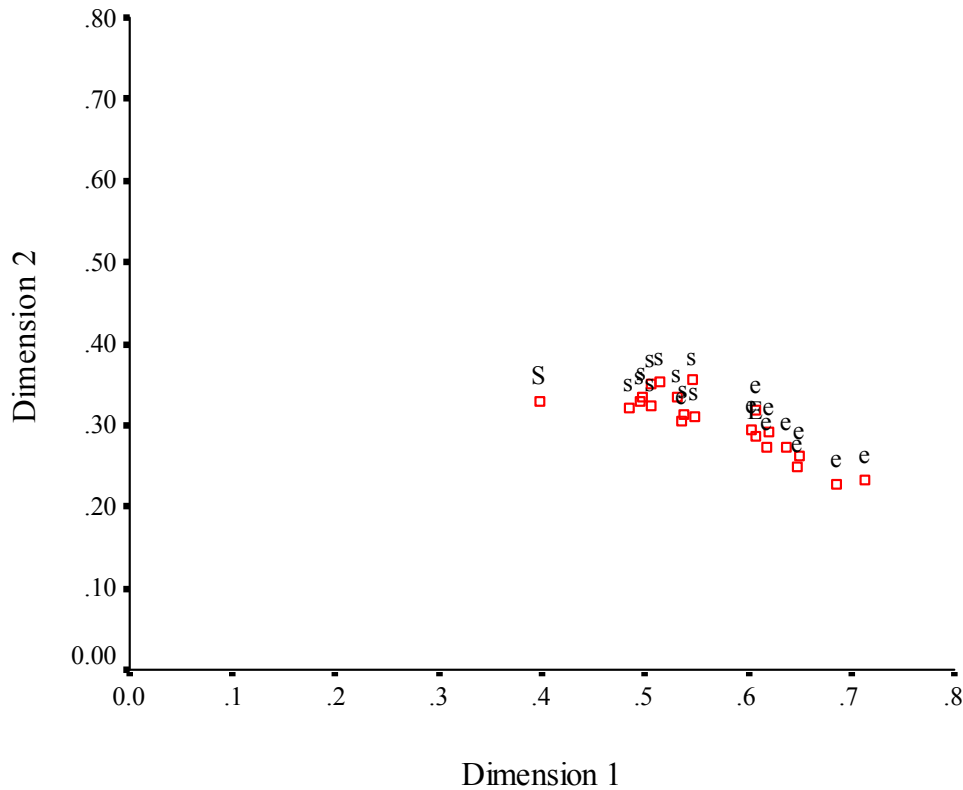


Figure 10. MDS Weights for English Video (E), Spanish Video (S) and Matched Samples (e, s): Grade Six Social Studies

Confirmatory Factor Analysis Results
 Grade Three: Mathematics
 English video vs. standard group

When testing the measurement invariance between the English video group and each of the 10 replicated samples from the standard group that were not matched on proficiency, the augmented model fit better than the compact model according to the difference in χ^2 statistics in only one of the 10 replications. Furthermore, the average RMSEA value (0.04) for the compact model over the 10 replications was indicative of good fit (RMSEA values less than 0.06 are an indication of acceptable fit (Hu & Bentler, 1999)). The factor loadings for the respective content strand were similar between the two groups.

Table 6 reports the average standardized factor loadings per strand for the English video and non-accommodated group across the 10 comparisons for both the unmatched and matched random samples. The average factor loadings for the number and operations content strand are not reported in Table 6 because the factor loading was fixed to 1.00 in both groups to establish the scale of the latent variable.

Table 6. Average Standardized Factor Loading across 10 Replications English Video/Standard Comparison for Grade Three Math

Group	Content Strand			
	Measurement		Geometry	
	Unmatched	Matched	Unmatched	Matched
Standard Admin.	0.89	0.85	0.70	0.70
English Video	0.81	0.81	0.70	0.69

The CFA analyses using the matched samples produced similar results in that the augmented model did not significantly fit better than the compact model for any matched sample. Furthermore, the average RMSEA value for the compact model across replications was essentially zero, indicating that placing constraints on the factor loadings did not detrimentally influence model fit. Therefore, the excellent fit of the compact model is an indication that the factor loadings between the standard administration and English video group were not different.

Spanish video vs. non-accommodated group

When testing the measurement invariance between the Spanish video group and each of the 10 replicated samples from the non-accommodated group that were not matched on proficiency, 8 of the 10 comparisons indicated that the augmented model fit better than the compact model according to the difference in χ^2 statistics. Furthermore, the average RMSEA (0.15) for the compact model was indicative of a poorly fitting model. The average factor loadings for the respective content strands were different between the two groups. Table 7 reports the average standardized factor loadings per strand for the Spanish video and non-accommodated group across the 10 replications for both the unmatched and matched analyses.

Table 7. Average Standardized Factor Loading across 10 Replications: Spanish Video/Standard Comparison for Grade Three Math

Group	Content Strand			
	Measurement		Geometry	
	Unmatched	Matched	Unmatched	Matched
Standard Admin.	0.90	0.82	0.70	0.74
Spanish Video	0.70	0.79	0.54	0.61

It is clear from Table 7 that the factor loadings were larger for the non-accommodated group compared to the group that used the Spanish video in the unmatched analyses. Interestingly, the CFA analyses using the samples from the standard administration group that were matched on proficiency were not consistent with the analyses using the unmatched samples. When using the matched samples, the augmented model did not exhibit significantly better fit than the compact model, according to the difference in χ^2 statistics. Furthermore, in only one of the replications was the RMSEA value for the compact model greater than 0.06 indicating poor fit. Finally, the differences between the average standardized factor loading for measurement and geometry between the Spanish video group and standard administration group were less severe compared to those obtained by analyzing the unmatched samples. Given these results and the fact that the factor loadings for the Spanish video group were large, the CFA results suggest invariant test structure (at the content strand level) across the standard and Spanish video conditions.

Grade Five: Science
English video vs. standard group

When testing the measurement invariance between the English video group and each of the 10 replicated samples from the standard group that were not matched on proficiency, the augmented model did not fit better than the compact model according to the difference in χ^2 statistics in any of the 10 replications. Furthermore, the average RMSEA (0.02) for the compact model over the 10 replications was indicative of good fit. The factor loadings for the respective content strands were similar between the two groups. Table 8 reports the average standardized factor loadings across content strands for the English video and non-accommodated group across the ten comparisons. (The factor loading corresponding to content strand “Use Life Science Knowledge” was fixed to 1.0 to establish the scale of the latent variable.)

The analysis using the randomly drawn samples from the standard administration group matched on proficiency was very similar to the analyses using the unmatched samples. The augmented model did not fit better than the compact model in any of the 10 replicated samples. The average RMSEA value (0.03) was indicative of good fit, and the average standardized factor loadings were similar to those obtained in the unmatched analysis.

Table 8. Average Standardized Factor Loading across 10 Replications: English Video/Standard Comparison for Grade Five Science

Group	Content Strand							
	Construct Scientific Knowledge		Reflect Scientific Knowledge		Use Physical Knowledge		Use Earth Science Knowledge	
	Unmatched	Matched	Unmatched	Matched	Unmatched	Matched	Unmatched	Matched
Standard Admin.	0.66	0.65	0.50	0.49	0.64	0.59	0.63	0.56
English Video	0.69	0.64	0.49	0.44	0.80	0.75	0.61	0.59

Spanish video vs. standard group

When testing the measurement invariance between the Spanish video group and each of the 10 replicated samples from the standard group that were not matched on proficiency, the augmented model did not fit better than the compact model according to the difference in χ^2 statistics in any of the 10 replications. The average RMSEA (0.01) for the compact model across the 10 replications was indicative of good fit. However, the factor loadings for the respective content strands were consistently smaller for the Spanish video group.

Table 9 reports the average standardized factor loadings across the content strands for the Spanish video and non-accommodated group across the ten comparisons for both the unmatched and matched samples. The analysis using the randomly drawn samples from the standard administration group matched on proficiency were very similar to the analyses using the unmatched samples. Similarly, the augmented model did not fit better than the compact model in any of the ten replicated samples; the average RMSEA value (0.01) was indicative of good fit.

Table 9. Average Standardized Factor Loading across 10 Replications: Spanish Video /Standard Comparison for Grade Five Science

Group	Content Strand							
	Construct Scientific Knowledge		Reflect Scientific Knowledge		Use Physical Knowledge		Use Earth Science Knowledge	
	Unmatched	Matched	Unmatched	Matched	Unmatched	Matched	Unmatched	Matched
Standard Admin.	0.69	0.70	0.48	0.47	0.65	0.45	0.48	0.48
Spanish Video	0.50	0.50	0.30	0.30	0.39	0.39	0.27	0.27

Grade Six: Social Studies

English video vs. standard group

When testing the measurement invariance between the English video group and each of the ten replicated samples from the standard group that were *not* matched on proficiency, the augmented model did not fit better than the compact model according to the difference in χ^2 statistics in any of the 10 replications. Furthermore, the average RMSEA (0.02) for the compact model across the 10 replications was indicative of good fit. The factor loadings for the respective content strands

were similar between the two groups. The analysis using the randomly drawn samples from the standard administration group that were matched on proficiency was very similar to the analyses using the unmatched samples. The augmented model did not fit better than the compact model in any of the 10 replicated samples. The average RMSEA value (0.01) was indicative of good fit and the average standardized factor loadings were similar for the standard administration and English video group. A summary of the standardized factor loadings is presented in Table 10.

Table 10. Average Standardized Factor Loading across 10 Replications: English Video /Standard Comparison for Grade Six Social Studies

Group	Content Strand							
	Construct Scientific Knowledge		Reflect Scientific Knowledge		Use Physical Knowledge		Use Earth Science Knowledge	
	Unmatched	Matched	Unmatched	Matched	Unmatched	Matched	Unmatched	Matched
Standard Admin.	0.66	0.66	0.49	0.49	0.64	0.59	0.63	0.56
English Video	0.66	0.64	0.49	0.44	0.80	0.75	0.61	0.57

Spanish video vs. standard group

When testing the measurement invariance between the Spanish video group and each of the 10 replicated samples from the standard group that were *not* matched on proficiency, the augmented model fit better than the compact model in all 10 replications. The average RMSEA (0.08) for the compact model across the 10 replications was indicative of poor fit. Furthermore, the factor loadings for the respective content strands were consistently smaller for the Spanish video group for content three of the strands (history, geography, and economics), while the factor loading was larger for the inquiry content strand.

The analysis using the randomly drawn samples from the standard administration group that were matched on proficiency provided a slightly different picture regarding measurement invariance. The augmented model fit better than the compact model in only 3 out of the 10 matched replications and the average RMSEA value (0.05) was (barely) indicative of good fit. Table 11 presents the average standardized factor loadings across the content strands for the Spanish video and standard administration group across the 10 comparisons. (The factor loading for content strand “civics” was fixed to 1.0 to establish the scale of the latent variable.)

The average standardized factor loadings exhibited a similar pattern to those from the unmatched analyses. That is, the average standardized factor loadings for content strands history, geography, and economics were larger for the standard administration while the average standardized factor loading for the inquiry content strand was larger for the Spanish video group.

Table 11. Average Standardized Factor Loading across 10 Replications: Spanish Video /Standard Comparison for Grade Six Social Studies

Group	Content Strand							
	History		Geography		Economics		Inquiry	
	Unmatched	Matched	Unmatched	Matched	Unmatched	Matched	Unmatched	Matched
Standard Admin.	0.76	0.70	0.74	0.67	0.78	0.85	0.56	0.38
Spanish Video	0.68	0.54	0.48	0.38	0.46	0.37	0.80	0.57

Differential Item Functioning Results
 Grade Three: Mathematics
 English video vs. standard group

When using the randomly drawn samples from the standard administration that were *not* matched on proficiency, two items were flagged as exhibiting non-ignorable DIF. Table 12 reports the proportion of replications in which the item was flagged as DIF and the average R^2 -change value for the respective item. Although both items were identified as exhibiting non-ignorable DIF, the magnitude of DIF was not large for either item.

Table 12. Items Flagged for DIF: Grade Three Math: Standard vs. English Video Comparisons (Unmatched Samples)

Item	Proportion of Replications Flagged as DIF ^a	Average R^2 -change
33	0.42	0.034
50	0.80	0.044

^aAn item was flagged as DIF if it was statistically significant and the R^2 -change was greater than or equal to 0.035.

Two items were flagged as exhibiting non-ignorable DIF when using the randomly drawn samples from the standard administration that were matched on proficiency, one of which (item 50) was also identified using the unmatched samples. Table 13 reports the proportion of replications in which the item was flagged as exhibiting non-ignorable DIF, the average R^2 -change for the respective item, and the classical item difficulty index (i.e., p -values) for the standard administration and English video groups. We report the item difficulties because we believe they help explain the reason why one item was flagged for DIF.

Item 16 was extremely easy for both groups, but was slightly easier for the English video group. This item was a graphical item that required students to select among three geometric shapes. Approximately 98 percent of the examinees in the English video answered the item correctly while roughly 95 percent of the examinees in the non-accommodated (standard) group responded correctly. Since the standard group outperformed the English video group on virtually all the other items, this item popped out as an interaction, and hence was flagged for DIF. Clearly, given the p -values, the “advantage” to the English video group is not substantive. The other flagged item (item 50) was easier for the standard administration group, and although the magnitude of

DIF was only medium, qualitative analyses should be conducted to identify the cause of the DIF, since it was not obvious upon our inspection of the item. Given that this item was flagged consistently in both the matched and unmatched analyses, the DIF appears to be substantive.

Table 13. Items Flagged for DIF: Grade Three Math: English Video vs. Standard Administration (Matched Samples)

Item	Proportion of Replications Flagged as DIF ^a	Mean R^2 -Change	p -value for Standard Administration	p -value for English Video
16	1.00	0.065	0.958	0.984
50	1.00	0.054	0.679	0.545

^aAn item was flagged as DIF if it was statistically significant and the R^2 -change was greater than or equal to 0.035.

Spanish video vs. standard group

Table 14 presents a summary of the items flagged for DIF using the Spanish video and unmatched random samples from the standard administration group. The proportion of replications in which the item was flagged as DIF and the average R^2 -change value for the respective item are reported. Three items were flagged as exhibiting non-ignorable DIF. One item was the very easy item mentioned earlier (item 16, where the p -value for the Spanish video group was 0.97). For all three items, the average DIF effect size was medium.

Table 14. Items Flagged for DIF: Grade Three Math: Standard vs. Spanish Video Comparisons (Unmatched Samples)

Item	Proportion of Replications Flagged as DIF ^a	Average R^2 -change
16	0.33	0.053
19	0.37	0.041
33	0.96	0.052

^aAn item was flagged as DIF if it was statistically significant and the R^2 -change was greater than or equal to 0.035.

When using the randomly drawn samples from the standard administration that were matched on proficiency, two items were flagged as exhibiting non-ignorable DIF, one of which (item 19) was also identified using the unmatched samples (see Table 15). In considering the p -values for the item for each group, item 14 was easier for the standard administration group while item 19 was easier for the Spanish video group. The average effect size noted for each item was medium. The cause of DIF was not apparent from inspecting these items (although we note item 19 is very easy for both groups) and so qualitative analysis by bilingual math experts is recommended.

Table 15. Items Flagged for DIF: Spanish Video vs. Standard Administration (Matched Samples)

Item	Proportion of Replications Flagged as DIF ^a	Mean R^2 -Change	p -value for Standard Administration	p -value for Spanish Video
14	1.00	0.035	0.434	0.291
19	1.00	0.042	0.878	0.952

^aAn item was flagged as DIF if it was statistically significant and the R^2 -change was greater than or equal to 0.035.

Grade Five: Science
English video vs. standard group

When using the randomly drawn samples from the standard administration that were not matched on proficiency, two items were flagged as exhibiting non-ignorable DIF. Table 16 summarizes the results for the flagged items. Table 17 summarizes the results for the items flagged using the matched samples. The same two items were flagged using both sampling strategies. Both items were easier for the English video group compared to the standard administration group. Although the cause of DIF was not apparent to us, qualitative analyses should be conducted to see if the video administration somehow made these items easier.

Table 16. Items Flagged for DIF: Grade Five Science: English Video vs. Standard Administration (Unmatched Samples)

Item	Proportion of Replications Flagged as DIF ^a	Mean R^2 -Change
37	0.73	0.043
39	0.73	0.042

^aAn item was flagged as DIF if it was statistically significant and the R^2 -change was greater than or equal to 0.035.

Table 17. Items Flagged for DIF: Grade Five Science: English Video vs. Standard Administration (Matched Samples)

Item	Proportion of Replications Flagged as DIF ^a	Mean R^2 -Change	p -value for Standard Administration	p -value for English Video
37	0.61	0.039	0.454	0.596
39	0.49	0.035	0.316	0.457

^aAn item was flagged as DIF if it was statistically significant and the R^2 -change was greater than or equal to 0.035.

Spanish video vs. standard group

When using the randomly drawn samples from the standard administration that were not matched on proficiency, five items were flagged as exhibiting non-ignorable DIF (see Table 18). The average effect size for all five items was medium.

Table 18. Items Flagged for DIF: Grade Five Science: Spanish Video vs. Standard Administration (Unmatched Samples)

Item	Proportion of Replications Flagged as DIF ^a	Mean R^2 -Change
3	0.93	0.062
13	0.65	0.047
22	0.47	0.036
38	0.56	0.039
39	0.62	0.044

^aAn item was flagged as DIF if it was statistically significant and the R^2 -change was greater than or equal to 0.035.

When using the matched samples, six items were flagged as exhibiting non-ignorable DIF, five of which were the same items identified in the unmatched analyses, and one of which (item 15) was not identified using the unmatched samples (see Table 19). Two items exhibited large DIF—items 3 and 13. Item 3 was a difficult item that was much more difficult for the Spanish video group. In fact, the proportion correct for the Spanish video group on this item was below chance level (p -value=0.206 < $\frac{1}{4}$). The DIF observed in item 3 may be more related to cultural factors associated with the population that received the Spanish video accommodation rather than the accommodation itself. Item 3 asked the student to indicate which theory most geologists support with respect to the formation of the Great Lakes. Students receiving the Spanish video accommodation are probably less likely to be natives of the Midwest and thus would be at a disadvantage when answering items regarding knowledge related to the geographic area. Item 13, on the other hand, was of moderate difficulty for the standard group (p -value=.67), but much easier for the Spanish video group (p -value=.86).

Table 19. Items Flagged for DIF: Grade Five Science: Spanish Video vs. Standard Administration (Matched Samples)

Item	Proportion of Replications Flagged as DIF ^a	Mean R^2 -Change	p -value for Standard Administration	p -value for Spanish Video
3	1.00	0.117	0.482	0.206
13	1.00	0.082	0.670	0.855
16	0.92	0.057	0.602	0.400
22	0.68	0.041	0.228	0.376
38	0.79	0.053	0.285	0.467
39	0.94	0.066	0.263	0.461

^aAn item was flagged as DIF if it was statistically significant and the R^2 -change was greater than or equal to 0.035.

Of the other four items flagged for DIF, one was easier for the standard group and the other three were easier for the Spanish video group. Potential causes of DIF were not apparent to us for

these items and so follow-up qualitative analyses by qualified educators (e.g., English-Spanish bilingual science teachers) are recommended.

Grade Six: Social Studies
English video vs. standard group

No items were flagged as exhibiting non-ignorable DIF when using either the randomly drawn samples from the standard administration group that were not matched on proficiency or when using the randomly drawn samples from the standard administration group that were matched on proficiency.

Spanish video vs. standard group

When using the randomly drawn samples from the standard administration that were not matched on proficiency, three items were flagged as exhibiting non-ignorable DIF, all of which had an average effect size of medium (see Table 20). When using the matched samples, four items were flagged as DIF—the same three items flagged using the unmatched samples and one other (see Table 21).

Table 20. Items Flagged for DIF: Spanish Video vs. Standard Administration (Unmatched Samples)

Item	Proportion of Replications Flagged as DIF ^a	Mean R^2 -Change
2	1.00	0.084
9	0.62	0.044
26	0.71	0.046

^aAn item was flagged as DIF if it was statistically significant and the R^2 -change was greater than or equal to 0.035.

Item 2 was one of the items flagged in both the matched and unmatched analyses. As the p-values in Table 21 illustrate, the item was difficult for the standard group (p-value=.33), but extremely difficult for the Spanish video group (p-value=.10). Similar to item 3 on the grade five science test, the DIF observed in item 2 may be more related to the population that received the Spanish video accommodation than the accommodation itself. Item 2 requested information regarding the travel via the Erie Canal. Since students who received the Spanish video accommodation may be less likely to be natives of the Midwest, they may be at a disadvantage when answering items regarding knowledge related to the geographic area. This possibility is congruent with the finding that the geography content strand had a relatively lower factor loading for the Spanish video group in comparison to the standard group in the CFA analysis. As for the other three items flagged for DIF, all were difficult, with the Spanish video group having higher p-values on two of the three items. Interpretations of the cause of the DIF in these items were not apparent to us and so, again, follow-up analyses by qualified personnel are recommended.

Table 21. Items Flagged for DIF: Spanish Video vs. Standard Administration (Matched Samples)

Item	Proportion of Replications Flagged as DIF ^a	Mean R^2 -Change	p -value for Standard Administration	p -value for Spanish Video
2	1.00	0.125	0.326	0.098
9	0.79	0.052	0.292	0.474
26	0.91	0.059	0.293	0.486
27	0.79	0.048	0.366	0.208

^aAn item was flagged as DIF if it was statistically significant and the R^2 -change was greater than or equal to 0.035.

Discussion

In this study, we evaluated invariance of test structure and item scores across three test administration conditions using three different statistical procedures and sampling strategies. The results provide information regarding the comparability of test structure and item functioning across these different test administration conditions, as well as information regarding the utility of the different data analysis methods and sampling strategies we used.

Given the many different analyses we conducted, a summary of the results is helpful. Table 22 presents a summary of the conclusions we drew for each analysis for each grade/subject under each sampling condition. The symbols used in the table ($=$, \neq , \approx) are simplifications of our conclusions, but they represent our final judgment based on our interpretations of the results. With respect to the similarities across the MDS and CFA analyses of test structure, when using the unmatched random samples, there were more differences noted across the methods. Only two of the six analyses (grade three English video/standard comparison for math and grade six Spanish video/standard comparison for social studies) led to the same conclusions regarding test structure when the samples were not matched on proficiency.

Table 22. Summary of Invariance and DIF Conclusions

Grade, Subject	Comparison	Sampling Strategy						Common DIF Items
		Unmatched			Matched			
		MDS	CFA	# DIF Items	MDS	CFA	# DIF Items	
3 Math	Standard vs. Eng. Video	=	=	2	\approx	=	2	1
	Standard vs. Span. Video	=	\neq	3	=	=	2	1
5 Science	Standard vs. Eng. Video	\approx	=	2	=	=	2	2
	Standard vs. Span. Video	\approx	=	5	=	=	6	5
6 Social Studies	Standard vs. Eng. Video	\neq	=	0	=	=	0	n.a.
	Standard vs. Span. Video	\neq	\neq	3	=	\approx	4	3

Note: = signifies a conclusion of invariant test structure, \approx signifies approximate invariance, and \neq signifies a lack of structural invariance.

When using the random samples matched on proficiency, MDS and CFA led to similar conclusions in four of the six comparisons. However, unlike two cases for the unmatched samples, the different conclusions were “approximate invariance” versus “invariance,” rather than “invariance” versus “lack of invariance.” Thus, the MDS and CFA analyses led to similar conclusions when using the matched samples. The general conclusion to be drawn from the results using the matched samples is that the test structure was invariant across the standard and video test administration conditions. This conclusion supports aggregating the results from the standard and video conditions for accountability purposes.

For some grade/subject areas, the conclusions regarding test structure differ across the matched and unmatched samples. For grade three math and grade six social studies, the conclusion of invariant test structure is less justified when considering only the results from the *unmatched* samples. The CFA indicated a much less dominant factor for the grade three Spanish video sample, relative to the random samples from the standard condition, and the structure for the Spanish and standard groups for the grade six science test looked different based on both the MDS and CFA results. In addition, the structure of the grade six English video data appeared different from the standard condition when using MDS. Interestingly, in all three cases, the structure appeared to be invariant, or approximately invariant, across groups when the random samples were matched on proficiency. This is an important finding because it suggests the differences in test structure, noted in the analyses using the unmatched samples, are probably due to artifacts such as restriction of range or proficiency differences, rather than due to use of the read-aloud video.

The DIF results are less clear regarding which sampling strategy is better, which makes sense because the DIF approach already conditions the analysis based on total test score. Although there were differences in some of the items identified, there was also considerable overlap (50–100 percent). One glaring limitation of the study is that we did not have bilingual content experts available to help us interpret the DIF results. Such interpretations might not only help identify causes for the items flagged for DIF, they might suggest whether drawing matched random samples is helpful for identifying substantive DIF. For example, if the causes of DIF for items flagged using the matched samples could be identified more than for items flagged using the unmatched samples, we would recommend drawing matched random samples for the DIF analyses also. Without such information we cannot make such a recommendation and so, begrudgingly, we leave the qualitative analyses of DIF for future research.

In general, we believe the results of this study support the use of the video accommodations, at least from the perspective of structural invariance. However, it is clear the items flagged for DIF need to be investigated by qualified educators to see if the read-aloud video erroneously affected the difficulty of particular items. In addition, it will be important to determine whether the students found the video accommodation helpful or distracting.

Moving from the implications of our study of the video accommodation to the implications with respect to research methods for evaluating invariance, we think there are several features of the study that have implications for future practice. When researchers are faced with a situation where measurement and item invariance need to be inspected across groups that differ with respect to sample size and distribution of proficiency, the results suggest that drawing multiple

samples matched on proficiency provides a better picture of structural invariance. Drawing multiple samples allows researchers to gauge variation in test structure or item function due to only random variation. For the data we analyzed, it was clear that differences in the proficiency distribution affected the analysis of test structure. If we had not drawn the matched samples, we may have concluded that the video administration resulted in a change in the construct measured. By drawing the random samples so that the proficiency distributions overlapped completely, we see that video condition had little to no effect on test structure.

While that conclusion is good news from a methodological perspective (i.e., we can draw matched samples to avoid making incorrect conclusions about the effect of an accommodation) some of our findings are bad news with respect to valid measurement of students who are linguistic minorities. Some of these students may be insufficiently proficient in the language of the test to understand what is being asked. The translated audio in the read-aloud video accommodation is supposed to address that problem, but the very low performances of the majority of students in the Spanish video condition suggest the test was still too difficult for them. That conclusion may explain the weaker dominant dimension and factor loadings for these groups, and for the random samples matched to them. Clearly, we need to have tests that are more appropriately targeted to students with proficiencies that are noticeably lower than their peers.

In closing, it should also be pointed out that although we found the strategy of matching samples helpful, not all researchers will be able to do such matching in all conditions. In some cases, such as we found for grade three, the overlap among the standard and accommodation groups may be so small, it will be difficult to draw matched samples. In other cases, the sample sizes may be far smaller than the 100,000 we had here in the standard group. Nevertheless, where sample sizes permit, we recommend drawing random samples matched on proficiency to provide an additional lens for evaluating test and item invariance. Other features of this study, such as using both exploratory and confirmatory multigroup analyses of test structure and looking at both macro (parcel level) and micro (item level) units of analysis, are also likely to be helpful for obtaining a comprehensive look at measurement invariance.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.
- Angoff, W. H., & Cook, L. L. (1988). *Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test (Report No. 88-2)*. New York, NY: College Entrance Examination Board.
- Bagozzi, R. P., & Heatherton, T. F. (1994). A general approach to representing multifaceted personality constructs: Application to state self-esteem. *Structural Equation Modeling, 1*, 35–67.
- Carroll, J.D. and Chang, J.J. (1970). An analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika, 35*, 238–319.
- Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R.K. Hambleton, P. Merenda, & C. Spielberger (Eds.) *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3–38). Hillsdale, NJ: Lawrence Erlbaum.
- Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care, 44* (11), 1–7.
- Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, *6*, 1–55.
- Jodoin, M. G., & Gierl, M. J. (2001). Evaluating power and Type I error rates using an effect size with the Logistic Regression procedure for DIF. *Applied Measurement in Education, 14*, 329–349.
- Joreskog, K. & Sorbom, D. (2001). LISREL (Version 8.50) [Computer program]. Chicago: Scientific Software International.
- Kishton, J. M., & Widaman, K. F. (1994). Unidimensional versus domain representative parceling of questionnaire items: An empirical example. *Educational and Psychological Measurement, 54*, 757–765.
- MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*, 84–99.
- Marsh, H.W., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33*, 181–220.

- Meara, K. P., Robin, F., & Sireci, S. G. (2000). Using multidimensional scaling to assess the dimensionality of dichotomous item data. *Multivariate Behavioral Research*, 35 (2), 229–259.
- Schiffman, S.S., Reynolds, M.L., & Young, F.W. (1981). *Introduction to multidimensional scaling*. New York: Academic Press.
- Sireci, S. G. (2007). 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., Bastari, B., & Allalouf, A. (1998, August). *Evaluating construct equivalence across adapted tests*. Invited paper presented at the annual meeting of the American Psychological Association (Division 5), San Francisco, CA.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361–370.
- van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, 13, 29–37.
- van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R.K. Hambleton, P. Merenda, & C. Spielberger (Eds.) *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39–63). Hillsdale, NJ: Lawrence Erlbaum.
- Yuan, K.-H., Bentler, P. M., & Kano, Y. (1997). On average variables in a confirmatory factor analysis model. *Behaviormetrika*, 24(1), 71–83.
- Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Chapter 4: Evaluating Linguistic Modifications: An Examination of the Comparability of a Plain English Mathematics Assessment

Charles A. DePascale

National Center for the Improvement of Educational Assessment

In this report, we summarize lessons learned from an examination of one state’s efforts to develop and administer a comparable plain English version of its mathematics assessments at grades 3–8 and high school. The report includes analyses evaluating the comparability of the results from the state’s Plain English Mathematics (PEM) test forms and its standard test forms, but the focus of the analysis is not on the comparability of the particular test forms examined. Rather, those results are used to support the primary purposes of this study, which are evaluating the extent to which the processes and procedures used by the state to develop its PEM are successful in producing test forms that are comparable to its standard test forms, and identifying and evaluating methods that can be helpful in evaluating the comparability of test forms and increasing understanding of issues that can impact the comparability of results.

The Plain English Mathematics tests are offered as an accommodation to students with disabilities in language processing skills and English languages learners meeting specified participation requirements. The study included a review of tests across grades 3–8, but the analyses conducted in this study focus on the sixth-grade mathematics test. The conclusions from analyses conducted on the sixth-grade test are representative of the rest of the program; the sixth grade test is no longer a secure document, making it possible to include sample items in the report.

Background

The tests examined in this study are mathematics tests administered at grades 3–8 as part of a statewide K–12 large-scale assessment program. The state administers the tests each spring to public school students in compliance with the requirements of No Child Left Behind (NCLB). Results from the tests are used in the state’s school and district accountability program.

Test Design

At each grade level, the mathematics test consists of 50 multiple-choice items measuring the state’s mathematics content standards. Items are distributed across five reporting categories: Number and Number Sense; Computation and Estimation; Measurement and Geometry; Probability and Statistics; and Patterns, Functions, and Algebra. The distribution of items across reporting categories follows a traditional pattern with an emphasis on Number and Number Sense and Computation at grade three (48 percent of the test), shifting across grades to an emphasis on Patterns, Functions, and Algebra (32 percent of the test) by grade eight.

The primary scores of interest for this study are total mathematics scores based on performance across the 50 items on the test. Total student scores on each test are expressed in terms of

number of items correct (0–50), scaled scores (0–600), and achievement level classifications (400 Level, 500 Level). Individual student achievement level classifications are aggregated to the school and district level and reported as the percentage of students meeting the achievement level standards—a common reporting metric under NCLB. Test forms at each grade level are linked within and across years using item response theory (IRT) techniques based on the Rasch model.

Plain English Mathematics Test

A corresponding Plain English Mathematics (PEM) test form is constructed for each of the grade 3–8 standard tests. The test blueprint for the design of each PEM is identical to the blueprint for the corresponding standard test form. As will be described in more detail in the following section, linguistic modifications to the standard test form are made on an item-by-item basis. The result is a test form that is intended to measure the same mathematics content and skills as the standard test form while reducing unnecessary linguistic complexity.

The PEM is offered as an accommodation to the state’s standard test form. It is not considered as an alternate assessment, either measuring students against alternate achievement standards or modified content standards. It is administered to students with disabilities in language processing skills and English language learners (ELLs) who meet the state’s participation requirements. Among ELLs, the PEM is administered primarily to students in their first year of enrollment in a United States school, and students whose level of proficiency in English is not sufficient to take the standard form of the tests. Students who are ELL may not take the PEM for more than three consecutive years.

Developing the Plain English Mathematics Form

The state’s goal in the development of the PEM is to produce a test form that mirrors the standard test form in terms of the mathematics concepts and skills measured, but eliminates (or significantly reduces) unnecessary linguistic complexity that may be interfering with students’ ability to demonstrate what they know and are able to do in mathematics. Consistent with this goal, the process for developing the PEM is based on an item-by-item review of the standard test form. In a multi-stage process involving personnel from the state department of education, development staff from the department’s assessment contractor, and a panel of local educators, each item is reviewed to determine what modifications, if any, are needed to make the item more accessible to ELLs and students with disabilities in language processing skills.

The review of the items focuses on three areas of linguistic complexity: reading load, syntax, and vocabulary. Within those three major areas, developers and panelists are presented with guidelines for reviewing and modifying items. Those guidelines contain many of the 14 linguistic features identified by Abedi (2007; also Handbook chapter) as “slowing reading speed, making misinterpretation more likely and adding to the reader’s cognitive load, thus interfering with concurrent tasks.” The development guidelines and a list of Abedi’s 14 features are presented in Table 1.

Table 1. Mapping of Plain English Mathematics Test Development Guidelines to Research-based List of Features Causing Linguistic Complexity

	PEM Review and Modification Guidelines	Abedi List of Linguistic Features Causing Complexity
Reading Load	<ul style="list-style-type: none"> include only what is essential use clear, direct wording use present tense when feasible 	<ul style="list-style-type: none"> word length
Syntax	<ul style="list-style-type: none"> make sentences short eliminate clauses and conditionals eliminate compound verbs, partial phrases, and passive voice break lengthy phrases into separate sentences avoid negative questions 	<ul style="list-style-type: none"> sentence length long question phrases conditional clauses relative clauses subordinate clauses passive voice constructions long noun phrases sentence and discourse structure negation comparative structures prepositional phrases
Vocabulary	<ul style="list-style-type: none"> use high-frequency words avoid idioms and colloquial usage use technical vocabulary only when necessary use concrete nouns, not abstract nouns break up compound nouns into parts reduce use of pronouns and synonyms 	<ul style="list-style-type: none"> word frequency/familiarity concrete versus abstract or impersonal presentations

Test Score Comparability

As defined in Chapter 1 of the Comparability Handbook, “test scores can be considered comparable if they can be used interchangeably.” Of particular importance is the level at which the test scores are considered interchangeable (e.g., raw score, scaled score, achievement level).

As described above, the development process for the PEM was designed to produce item level comparability. That is, when determined necessary, linguistic modifications were made on an item-by-item basis to enhance the accessibility of the item but not alter the particular mathematics skills measured by the item. Consequently, both the standard and PEM test forms were not only built to the same test specifications, but should measure the same mathematics knowledge and skills item for item. The ideal result from such a design is that scores from the standard and PEM forms of the test are comparable at the raw score level and can be used interchangeably. A single raw score to scaled score conversion would be used to assign scaled

scores, and ultimately make achievement level classifications, regardless of the form that a student completed.

Evaluating Comparability

Even though the process for modifying test items was carefully designed to produce comparable results, it is necessary to gather empirical evidence that the modified items and resulting test form do produce comparable results. In this case, the state used two methods for evaluating comparability:

1. administering the original and modified forms of the items to randomly equivalent groups of students
2. linking the standard and PEM forms through an anchor set of items

Method 1 examines student performance on individual items, and Method 2 focuses on the overall results on the test.

Method 1: Comparing Performance on Original and Modified Items

As part of the test design for its standard test form, the state uses a common-matrix design to embed field test items within the operational administration of the state assessments. This design allows the state to administer both the original and modified versions of test items within the field test portion of the standard assessment before the items are used operationally on the STAND and PEM, respectively. Under the assumption that the field test forms are administered to randomly equivalent groups of students throughout the state, it is possible to directly compare student performance on the original and modified versions of the test items. If the mathematics demand of the item has not changed, it is expected that student performance will be the same on both forms of the item. The equivalence of performance can be compared using classical test statistics (e.g., p-value, discrimination, point biserial correlation) and/or through item response theory methods (e.g., examination of item parameters, fit statistics).

However, there are also disadvantages to this method that must be considered. First, if used to evaluate all modified items, the method can be costly in terms of use of field test slots on the standard assessment. In many state assessment programs available field test slots are limited as states try to balance the multiple demands of minimizing testing time for individual students, ensuring a sufficient field test sample for each item, and containing costs associated with the development and production of multiple test forms. Second, this method is fundamentally focused on the comparability of item scores and not test scores. Evidence of the comparability of individual item scores is useful, of course, but is not a direct indicator of the comparability of scores from a test form comprising many items. Third, because the items are administered through the standard test forms it is likely that very few students for whom the modified items are designed will be included in the field test sample. Therefore, although this method can provide valuable information about the mathematics demands of the original and modified items for the general student population, it offers little information about how the intended population will interact with the modified items.

Method 2: Linking the Standard and PEM Forms through Anchor Items

Method 2 involves statistically linking the standard and PEM test forms through a set of anchor items (i.e., a set of items administered to students completing the standard test and students completing the PEM). In this case, the anchor items are those items on the mathematics test that were not modified during the review process. Those items are identical and administered in the same position on both forms of the test, two advantageous conditions for linking. Also, because the items are embedded within the operational test, the linking does not require additional testing time for students and concerns that might arise with an external anchor test (e.g., fatigue, motivation) are minimized. This state used an item response theory (IRT) approach involving the Rasch model to link the tests, but other linking methodologies could also be used.

Linking the standard and PEM test forms in this manner serves two purposes. First, it can provide evidence to evaluate the comparability of test results across the two test forms. If the two test forms are interchangeable, one would expect the two forms to produce identical test characteristic curves (through an IRT approach) and/or identical raw score to scaled score conversion tables. The extent to which the test forms are not interchangeable will be reflected in the TCC or conversion tables. Second, linking the standard and PEM test forms provides a means for reporting results on the same scale even if the two test forms are not strictly interchangeable. In the same way that standard test forms are linked across years and alternate test forms are linked within years, the PEM can be linked to the standard test form. Applying this method, separate raw score to scaled score conversion tables are produced for the standard and PEM test forms.

The linking approach, of course, does reflect a slight shift in perspective. There is still the expectation that the item review and modification process will have accomplished its goals:

1. identify all items that need to be modified to enhance accessibility
2. ensure that the mathematics concepts and skills being assessed are the same whether the item is in its original or modified format

If those goals have been met there is an assumption that unmodified items will interact in the same way with students completing either form, and that modified items measure the same construct. There is no longer an assumption, however, that test forms comprising original and modified test forms will produce results interchangeable at the raw score level (i.e., that a single raw score to scaled score conversion table will be appropriate for both tests). The linking process will account for any differences in difficulty between the test forms due to modifications to the items.

Examining the Results of the Linking Analyses

The results of the linking analyses demonstrate that although the standard and PEM test forms may not be interchangeable in the strictest sense they do produce highly consistent results. The plot in Figure 1 shows the raw score to scaled score conversion tables for the standard and PEM test forms at grade six. The solid line represents the PEM and the dashed line represents the standard test. Similar plots for grades three, four, five, seven, and eight are presented in Appendix A.

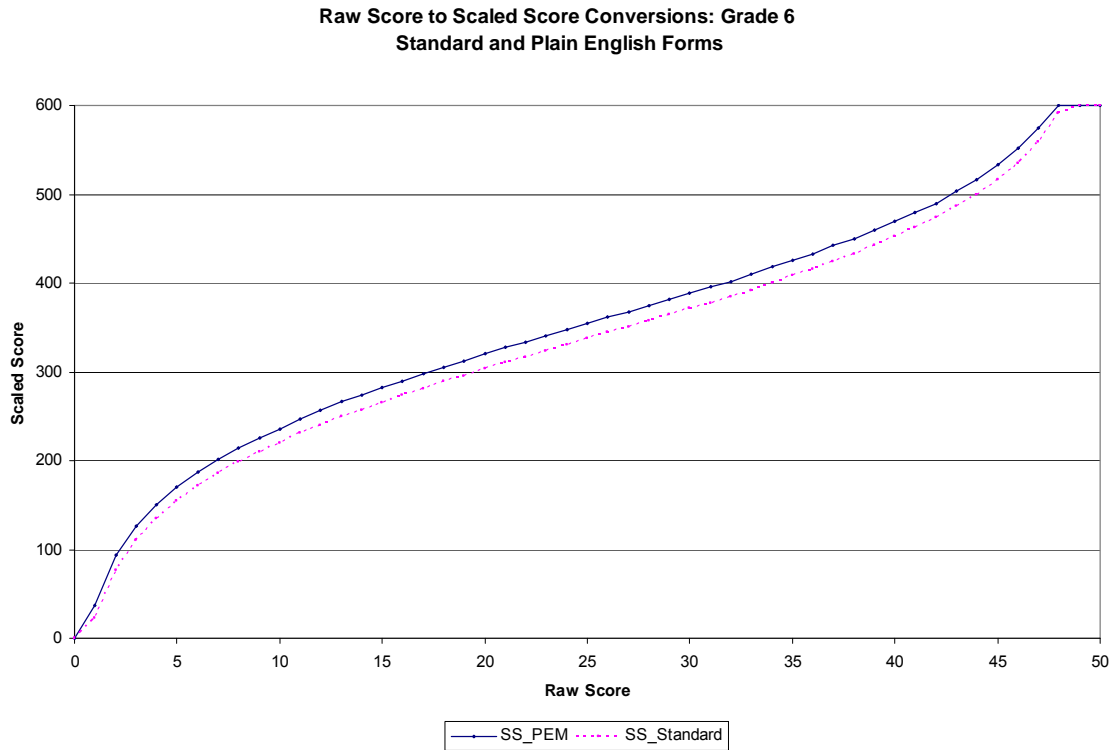


Figure 1. Raw Score to Scaled Score Conversions: Grade Six Standard and Plain English

The plot shows that for most raw scores there is a 15–20 point difference in scaled scores obtained on the standard and PEM test forms, with a higher scale score corresponding to each PEM raw score. For example, a raw score of 25 produces a scaled score of 337 on the standard test form and a scaled score of 354 on the PEM, a difference of 17 scaled score points. The 17-point difference must be evaluated, of course, within the context of the 600-point scale and the standard error of measurement. The classical standard error of measurement is approximately 30 scaled score points on each test. Conditional standard errors of measurement can range from approximately 20 points in the middle of the raw score range to greater than 45 points at the extremes, based on the particular test. Consequently, the 15–20 point difference in scaled scores between the standard and PEM test forms is within a single standard error of measurement, although the difference is all in one direction.

Impact on Achievement Level Classifications

One alternative to examining the differences in scaled scores for a given raw score is to examine the differences in raw scores needed to obtain the scaled score achievement level thresholds of 400 and 500. In Figure 1, for example, a raw score of 34 on the standard test form and a raw score of 32 on the PEM are needed to meet or exceed the 400 level. Raw scores of 44 and 43 on the standard and PEM test forms, respectively, are needed to meet or exceed the 500 level. The 2-point raw score difference at the 400 level on grade six was the largest difference found across

tests at grades 3–8. Table 2 presents the raw score differences at the 400 and 500 achievement level cuts on 12 pairs of standard and PEM tests forms at grades 3–8.

Table 2. Difference between Raw Scores Needed on Standard and PEM Test Forms to Attain Achievement Level Thresholds (Difference = Standard Test – PEM)

	Form A: Standard and PEM		Form B: Standard and PEM	
	400 level	500 level	400 level	500 level
Grade 3	1	0	1	0
Grade 4	1	0	1	1
Grade 5	1	0	2	1
Grade 6	2	1	0	0
Grade 7	1	0	1	0
Grade 8	0	0	1	0

Overall, the results in Table 2 show minimal differences between the PEM and standard test forms in raw scores needed to attain the achievement level thresholds. There are raw score differences of no more than 1 raw score point on 22 of the 24 thresholds examined across both achievement levels, and no differences in raw scores at the 500 level thresholds on 9 of the 12 test form pairs examined. Although no difference is greater than 2 raw score points, the finding of greater differences at the 400 level than the 500 level is not surprising given the increased precision of the test forms at the 400 level. As reflected in Figure 1, a change of a single raw score point near the 400 level translates to approximately 7 scaled score points, and a corresponding change at the 500 level translates to twice as many scaled score points.

Although the differences presented in Table 2 are small, it is clear that if the tests were considered to be completely interchangeable, and a single raw score to scaled score conversion table were used, some students would be assigned to a different achievement level depending on the test form that they completed. The impact of these achievement level classifications on the comparability of results from the two test forms depends, in part, on the intended uses of the test results. If the focus is on individual student results, there is as much imprecision in the score of an individual student on either test form as there is a difference between forms. That is, achievement level classifications of students scoring near the achievement level threshold should be interpreted cautiously regardless of the test form completed. If the focus is on school- or district level results (e.g., the percentage of students attaining the 400 achievement level) the impact would be dependent upon the number of students completing each test form and the distribution of their scores.

Comparison to Changes across Years

A second alternative to evaluating the extent of the comparability between test forms is to compare the differences found between the PEM and standard test forms within a year to the differences found between test forms across years. Every year, states attempt to develop a new standard test form that is *comparable* to standard test forms used in previous years. That is, the test form is built to the same test specifications and, in most cases, there is an attempt to keep the difficulty of the test constant from year to year. In practice, of course, given the improbability of creating strictly parallel test forms, states rely on various techniques to link their test forms across years. It can be informative, nevertheless, to compare the differences across the three forms.

In Figure 2, the standard test from the following year has been added to the plot comparing the raw score to scaled score conversion tables for the grade six standard and PEM test forms. (The additional standard form is represented by the bottom curve on the plot.) In this single case, it appears graphically that the comparability of the two standard forms across years in terms of raw score is greater than that of the standard and PEM forms within a year. Recall, that if the three test forms were interchangeable, one would expect the three curves to totally overlap. As stated, above, however, the significance of the observed differences must be evaluated within a particular context.

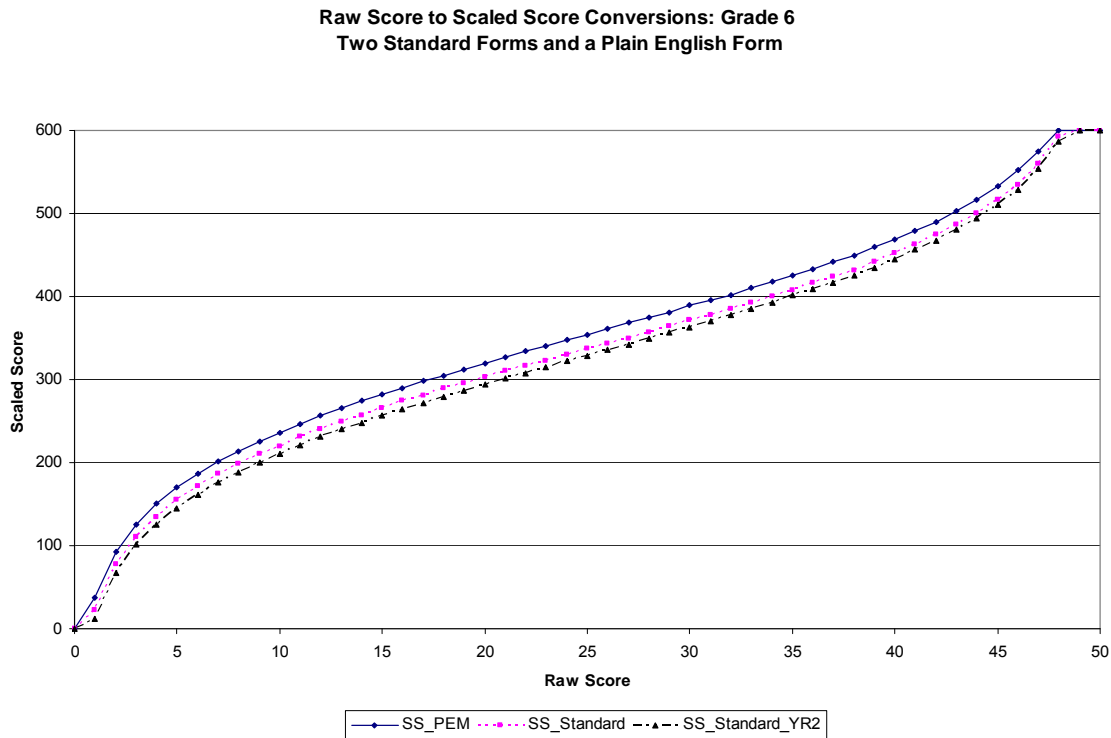


Figure 2. Raw Score to Scaled Score Conversions: Grade Six; Two Standard Forms and a Plain English Form

The PEM Paradox

The results of the IRT-based linking of the standard and PEM test forms presented above were useful in evaluating the comparability of the test forms. However, those results also presented a finding that was unexpected and perplexing. The information provided in Table 2, Figures 1 and 2, and all of the plots in Appendix A consistently pointed to the same paradoxical conclusion:

- The Plain English Mathematics test forms are more difficult than the standard test forms.

On every pair of standard and PEM test forms examined, the same raw score results in a higher scaled score for students completing the PEM than for students completing the standard test form. (The only exceptions are the extreme upper ends of the scales, where scaled scores are truncated at 600.) Because the standard and PEM test forms have been placed on the same scale

(i.e., because the mathematics ability required to reach each scaled score is the same on the two forms), the PEM test form is, by definition, more difficult than the standard test form.

Given that the intent in developing the PEM is to reduce unnecessary linguistic complexity in order to make the tests more accessible, and that all of the modified items appear to have been *simplified*, the outcome that the PEM test form is more difficult, no matter how slightly or lacking in practical or statistical significance, is counterintuitive. The more common concern throughout the development process is that the modifications will make the PEM *easier* than the standard test form. Note that, in theory, the relative difficulty of the two forms is a finding that applies to all students taking the test and not only to students with disabilities in language processing skills or students who are ELL. That is, based on the results of this linking analysis, any student taking both test forms would be expected to earn a higher raw score on the standard test form than the PEM. How can this finding be explained?

When puzzling results such as these occur, there are normal sequences of reactions and hypotheses to explain them or rationalize them. A first reaction is that the results are inaccurate—something was done incorrectly during the analysis. However, very similar results were found in another analysis of a linguistically modified state assessment (Shaftel et al., 2003). In that study, “ELL students taking the original items performed better than the ELL students taking the plain English versions of the test at grade 7.” As was the case in this analysis, the differences were present, but small. A second reaction is that the results are accurate—the changes made to the items do actually make the modified items more difficult than the original items. This often leads to post hoc rationalizations such as concluding that the linguistic modifications removed context that either made the item easier to answer or enhanced students’ engagement with the item leading to improved performance. As with the first reaction, however, there is little evidence to support the explanation of why the modified items are more difficult. The process for modifying the items is based on a solid body of research on linguistic complexity and includes the expert judgment of specialists in assessment, language, and local educators working in the classroom. Also, there was no evidence from the item level comparison described in Method 1 to support the conclusion that the PEM should be more difficult than the standard test form. Although there may be a tendency in psychometric research to value data over judgment, all else being equal, the preponderance of evidence in this case supports the qualitative (i.e., expert judgments) over the quantitative (i.e., linking results). Satisfied that the results are not due to a computational or programming error, but still unable to explain them, the next course of action is to examine more closely the process for linking the standard and PEM test forms.

From a simplistic perspective, the examination of the linking process can be divided into two distinct stages. One stage involves an examination of the selection of a linking/equating method, the resulting decisions related to that method (e.g., number of anchor items, necessary sample sizes), the IRT models used, and the application of the chosen methods and procedures. On these factors, the linking of two 50-item multiple-choice test forms through the Rasch model using an internal set of anchor items is about as straightforward and plain-vanilla as a linking problem and solution can be. Sample sizes are sufficient with more than 1,000 students taking each test form, and the size of the anchor test is adequate with approximately 12–15 linking items on each grade level test. The second stage involves an examination of the characteristics of the samples of

students completing the test forms and the items used in the anchor set. On these factors, issues surrounding the linking of the standard and PEM test forms are more complex.

In theory, the IRT procedures used to calibrate and link (or equate) the standard and PEM test forms should not be impacted *greatly* by differences in performance between the groups taking the standard and PEM test forms or the particular sample of items chosen to serve as the linking items on the anchor test. The extent of the impact of those factors on equating is difficult to isolate, as their impact is often confounded with other factors such as the calibration and equating methods selected. In practice, it is recommended that the groups taking the two forms to be equated are not “extremely different,” and the set of items in the anchor test is representative of the total test in content and statistical characteristics (Kolen and Brennan, 2004). Regarding the statistical characteristics of the anchor test, the similarity of the mean difficulty of the anchor test to the overall test form appears to be particularly relevant. Recent research suggests that the spread of item difficulties might not be as important as the mean difficulty (Sinharay and Holland, 2007).

In linking the standard and PEM test forms, we know that there are differences between the samples of students taking each test form, and there are also difference between the sample of items in the anchor test and the test as a whole. The group of students taking the PEM consists of students who are ELL and students with disabilities in language processing skills. The set of linking items in the anchor test are items that were not modified during the review process. By definition, those items differ from the other items on the test in terms of linguistic complexity. Whether they also differ in terms of test content and statistical characteristics is a question that can be answered. The manner and extent to which any of those differences impact the equating process is unknown. As an example of the process and the potential impact of various factors, the following section examines the process of linking the grade six standard and PEM test forms. The linking processes described in this report are for illustrative purposes and are not intended to fully reflect the operational procedures implemented by the state.

Linking the Grade-six Standard and PEM Test Forms

The grade six mathematics test was administered to nearly 40,000 students with approximately 36,500 completing the standard test form and 3,500 completing the PEM. A comparison of the two groups of students on some background and performance characteristics is presented in Table 3. It is clear from the information provided in Table 3 that there are significant differences in both background characteristics and performance between the two groups of students.

Table 3. Comparison of Samples of Students Completing the Grade Six Standard and PEM Test Forms

Background	Standard	PEM
Sex		
Female	50%	37%
Male	50%	63%
Students with Disabilities	8%	77%
English Language Learners	8%	35%
Performance		
Percentage of Students achieving Level 400	55%	20%
Scaled Score: Mean (sd)	411.50 (84.04)	347.35 (68.17)
Number Correct: Mean (sd)	33.86 (9.36)	23.92 (8.78)
Number Correct: Percent	67%	48%

Consistent with mathematics tests at other grade levels, the grade six mathematics test consisted of 50 multiple-choice items measuring content standards distributed across five major reporting categories. After the review process, 36 items were modified to reduce unnecessary linguistic complexity, leaving 14 items in their original format to serve as the anchor test or linking items between the standard and PEM test forms. Information on the distribution of items across reporting categories and a crude measure of overall performance on the items (i.e., percent correct) are presented in Table 4 and Table 5.

Table 4. Comparison of Content of the Linking Items to the Total Test on the Grade Six Standard and PEM Test Forms

Distribution across Reporting Categories (Number and Percentage of Items)		
Reporting Category	Total Test	Linking Items
Number and Number Sense	8 items (16%)	4 items (29%)
Computation and Estimation	10 (20%)	2 (14%)
Measurement and Geometry	12 (24%)	5 (36%)
Probability and Statistics	8 (16%)	0 (0%)
Patterns, Functions, and Algebra	12 (24%)	3 (21%)
Total	50 (100%)	14 (100%)

Table 5. Comparison of Performance on the Linking and Non-linking Items on the Grade Six Standard and PEM Test Forms

Mean Percent Correct (sd)		
	Standard	PEM
All Items (50 items)	67.7 % (12.1)	47.8 % (14.9)
Linking Items (14 items)	67.8 % (9.7)	50.8 % (14.3)
Non-linking Items (36 items)	67.7% (13.1)	46.7% (15.2)

In terms of content, the information provided in Table 4 indicates that there are some similarities and some clear differences between the sets of linking (original) and non-linking (modified) items. In both sets, Measurement and Geometry is the largest category represented, and the two categories Measurement and Geometry and Patterns, Functions, and Algebra account for roughly half of the items. The most striking difference between the sets is the total lack of Probability and

Statistics items in the linking set. Apparently, all items in this category were modified to some extent.

In terms of performance, the information provided in Table 5 shows little difference between student performance on the linking and non-linking items on the standard test. Students taking the standard test form answered approximately 68 percent of the items in each of the groups correctly. Students taking the PEM, however, showed slightly better performance on the linking items than the non-linking items. A closer examination of performance on the linking and non-linking items shows that the difference in performance within the PEM group can be largely attributed to two computation items included in the anchor set.

Figure 3 shows a scatterplot of the percent correct (i.e., p-values) on each of the 14 linking items for students completing the standard (x-axis) and PEM (y-axis) test forms. Also included on the scatterplot is an identity line ($y=x$). Items on which students taking the standard test form performed better appear below the identity line. Items on which students taking the PEM performed better appear above the identity line. Items on which the two groups performed the same would be plotted on the identity line.

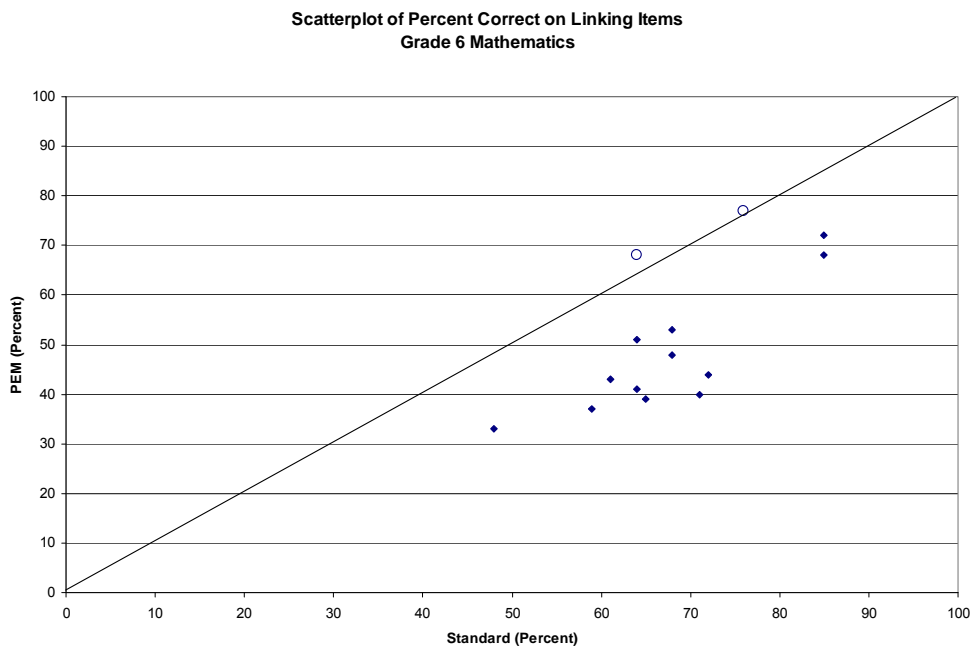


Figure 3. Scatterplot of Percent Correct on Linking Items: Grade Six Mathematics

The plot in Figure 3 clearly shows that the two computation items (designated by a transparent rather than a solid symbol) are behaving differently than all of the other linking items. For some reason, students taking the PEM outperformed students taking the standard test form on those two items. On the remaining 12 items, performance fairly consistently favored the students taking the standard test form. The two computation items are presented in Figure 4. Both items appear to measure the same general computational skill.

<p>4 0.084 ÷ 0.6 =</p> <p>F 7.14</p> <p>G 1.4</p> <p>H 0.714</p> <p>J 0.14</p>	<p>9 6.596 ÷ 0.04 =</p> <p>A 164.9</p> <p>B 16.49</p> <p>C 6.06</p> <p>D 0.61</p>
---	--

Figure 4. Computation and Estimation Linking Items

Within an equating context, the information provided in Figure 3 (whether presented as p-values, delta statistics, IRT item parameters, or any other relevant statistics) would raise a red flag. Based on some pre-established criteria, it would be necessary to decide whether to include those two items in the equating pool. Excluding the items from the anchor test produces a better statistical match between the linking and non-linking items, but results in an anchor test that does not include items that measure standards from two of the five major reporting categories and account for 36 percent of the items on the original test. Including the two items in the equating pool may improve the content match, but results in a slight differential in the difference between standard and PEM performance on the linking items (68 percent v. 50 percent) and non-linking items (68 percent v. 47 percent).

In terms of impact on equating, it appears that the differential in difference between performance on the linking and non-linking items does help explain the why the PEM appears to be more difficult than the standard test form. Conceptually, the *distance* between the performances of the two groups is established based on a comparison of the statistical characteristics of the linking items. Any additional differences found on the non-linking items will be attributed to differences in the difficulty of the items rather than to differences in the performance of the groups. The relationship between the crude measure of difficulty, percent correct, and the ultimate equating results will vary based on the type of IRT calibration model and equating methodology selected. In general, the strongest relationship probably exists when a 1-parameter IRT model is used with a linear equating design. For this study, a variety of equating analyses were conducted using item parameters generated from 1-parameter and 2-parameter IRT models, basic IRT observed score equating methods (mean-mean, mean-sigma), and varying whether the two Computation and Estimation items were included. Under the 2-PL model, differences in mean item difficulty between the two test forms were relatively small in all cases. However, the PEM form appeared more difficult than the standard form when the computation items were included; and the standard form appeared relatively more difficult when those items were excluded. Under the 1-PL model, the PEM appeared more difficult than the standard test form in both cases, but the difference was much greater when the computations items were included. Summary results of the linking analyses are presented in Table 6.

Table 6. Mean Item Difficulty of the Grade Six Standard Test Form and Mean Adjusted Item Difficulty of the PEM Test Form under a Variety of Linking Conditions Mean (sd)

Linking Condition	Standard Form	PEM
2-PL, mean-sigma, with computation items	-0.962 (.761)	-0.939 (.670)
2-PL, mean-mean, with computation items	-0.962 (.761)	-0.921 (.767)
1-PL, mean-sigma, with computation items	-0.972 (.743)	-0.795 (.739)
2-PL, mean-sigma, w/out computation items	-0.962 (.761)	-1.072 (.815)
2-PL, mean-mean, w/out computation items	-0.962 (.761)	-1.067 (.750)
1-PL, mean-sigma, w/out computation items	-0.972 (.743)	-0.957 (.739)

In this example using the grade six mathematics tests, the difficulty of the original items and modified items was quite similar. In general, however, when using the approach described in this report to develop a Plain English Mathematics test, it would not be unexpected to find a positive relationship between the linguistic complexity of the item (necessary or unnecessary) and the difficulty of the item. That is, it would not be unusual for the less difficult items to also have less linguistic complexity, and consequently require fewer modifications. In such a case, there would be a difference in overall performance on the linking and non-linking items for both groups. There may also be a difference in the relative performance of the groups on the linking and non-linking items. That is, the performance gap found on the less difficult items (i.e., the linking items) may be smaller than that found on the more difficult items (i.e., the non-linking items). As shown in the example in this study, this differential gap can impact the results of the linking study. Although not examined in this study, it may be the case that a larger mean difference in difficulty between the linking and non-linking items, but a constant gap in performance between the groups, might have less of an impact on the equating results than the relative difference found on the grade six mathematics tests in this example.

If there had been some variation in the relative difficulty of the standard and PEM test forms across grade levels (i.e., the standard test form was found to be more difficult in at least one grade) it is likely that few questions, if any, would have been raised by the results of the linking analyses. Having more closely examined the process of statistically linking the standard and PEM test forms, however, one important question does emerge:

- To what extent is the apparent increased precision (and equity) gained through the linking analyses real, and to what extent is it merely a function of the choices made during the linking process and the error associated with those choices?

In all cases examined, including the results presented by the state, the linking analyses showed relatively minor differences between raw score to scaled score tables generated for the standard and PEM test forms. In using a linking approach to evaluate the comparability of test results across forms, the state should determine beforehand whether the primary purpose of the linking study is to confirm, or validate, the comparability of the results from the two test forms or to statistically *create comparability* by developing unique raw score to scaled score conversion tables for each test form. If the former is the primary purpose of the linking study, then the end result might be to use a single raw score to scaled score conversion table for both test forms regardless of small differences found through the linking study. As Kolen and Brennan (2004)

state, “assuming that the test specifications, design, data collection, and quality control procedures are adequate, it is still possible that using the identify function [i.e., a single conversion table] will lead to less equating error than using one of the other equating methods” (p. 296). Approaches are available to examine the significance of differences among various equating approaches and to determine whether it is more appropriate to equate or not equate (Kolen & Brennan, 2004). Ultimately, a state’s answer to the question of whether results are comparable enough should include a consideration of these factors as well as the context in which the results will be used.

Comparability Coda

Before moving on from this discussion of linking the standard and PEM test forms, we would like to provide additional information regarding performance on the two unchanged computation items—Item 4 and Item 9—that behaved differently from all of the other items (see Figure 3 and Figure 4). As discussed previously, these were the only two items on which students taking the PEM test form outperformed, albeit slightly, students taking the standard test form. A review of the items revealed no linguistic- or mathematics-based explanation for the performance. As it turns out, the explanation lies in a threat to comparability that was not the focus of this study: the administration conditions of the item.

The computation items in question were administered during the non-calculator session of the mathematics test. However, some students with disabilities were able to use a calculator as an approved accommodation on the non-calculator portion of the test. As shown in Table 3, the vast majority (i.e., 77 percent) of the students completing the PEM are students with disabilities. The calculator accommodation was used by approximately 40 percent of the students completing the PEM. In comparison, slightly more than 1 percent of the students completing the standard test form used the calculator accommodation—too few students to impact overall item results. On both items, more than 80 percent of the students with the calculator accommodation responded correctly. The results in Table 7 show the impact of their performance on the overall item results. Without the use of the calculator accommodation, students completing the standard test form would have outperformed students completing the PEM on these two items as they did on the other unchanged linking items on the test forms.

Table 7. Impact of the Use of the Calculator Accommodation on Performance on Computation Items 4 and 9 (Percent Correct)

	Standard Test Form	All PEM Students	PEM Students with Calculator	PEM Students without Calculator
Item 4	76%	77%	88%	67%
Item 9	63%	67%	84%	57%

Item Comparability

In the previous section of this report, the focus was on the comparability of overall test scores between standard and linguistically modified test forms. In this section, the focus shifts to evaluating whether individual items are producing comparable results across groups. As stated by Kopriva et al. (2008) “one of the best ways of evaluating if a test is yielding valid inferences for different groups of students is to investigate how items are functioning for these groups” (p.

3). If the goal of states is to build linguistically modified assessments that produce comparable results, it is logical to conclude that an increased understanding of how to construct comparable items will lead them closer to achieving that goal.

Background

Kopriva et al. examine *distractor analysis* as an optimal approach for examining the comparability of item performance in cases such as the linguistically modified Plain English Mathematics test form examined in this study. The primary distinction between distractor analysis and an item-based approach such as Differential Item Functioning (DIF) is the ability to examine among students who answer the item incorrectly whether students in a particular group are drawn to a particular distractor more than students in another group(s). If so, there may be cause for concern that the item is measuring something different for each of the groups. Like DIF analysis, however, distractor analysis only indicates whether there is a significant difference; the post hoc task of interpreting what caused the difference and determining whether it is a reason to revise or remove the item remains.

Consistent with many DIF approaches, distractor analysis is based on using a chi-square test to examine the patterns of responses for two groups of students. In the case of distractor analysis, the response categories are the incorrect options (i.e., distractors) for each multiple-choice item. The correct response option is not included in the analysis. Kopriva et al. proposed that a separate response category be used for each distractor to provide the most useful information about the item. This was a variation of the method used by Abedi (2008) in which two response categories were used: the distractor receiving the greatest overall response was one category and the remaining distractors were collapsed into a second category. One disadvantage of the expanded approach described in the Kopriva et al. study is that when dealing with low incidence groups and field test data from a sample of students it is likely that there will be many distractor/group cells with very small response frequencies. Results of the chi-square tests are interpreted based on effect size and classified as small (0.00–0.12), medium (0.12–0.30), and large (0.30+) based on the effect size classification metric developed for DIF analysis by ETS.

Overall, Kopriva et al. found that distractor analysis identified a large percentage of the items in their study as having moderate or large effect size when examined with either students with learning disabilities or students who were English language learners. Across the various samples examined, between 44 percent and 77 percent of the items were identified. Those results reflect a much larger percentage of items identified than in the Abedi study (approximately 10–20 percent) and a disturbingly large percentage of items to be identified on a test.

Factors Impacting Distractor Analysis Results

Given the alarmingly large percentage of items identified through distractor analysis, we decided in this study to attempt to gain an understanding of the analysis-related factors that might impact the results. By *analysis-related factors* we are referring to characteristics of the design of the analysis rather than to factors related to the items themselves (e.g., difficulty, cognitive complexity). Two factors that seemed to be potentially relevant given the nature of the analysis were sample size and the relative performance of the two groups. When dealing with large sample sizes such as those found in statewide assessment, there is little question that there will

be large (and significant) values of chi-square. Accepting large values of chi-square as a given, the other variable used in determining effect size (i.e., sample size) seemed particularly important. Group performance seems particularly important within distractor analysis because within many test development frameworks there is an explicit expectation that students performing at different levels will show different patterns of response across distractors. It is not uncommon for at least one distractor to reflect a common misconception held by students with either limited or partial understanding of the construct being measured. It is also not unusual for distractors to possess varying levels of plausibility based on a student's level of mathematics knowledge and skills. For example, consider Figure 5 containing the Computation and Estimate item previously discussed with regard to equating. The content of the item is division with decimals. Without solving the problem, at some point along the performance continuum students will understand that dividing 6.596 by a positive number less than 1.0 will yield a result greater than 6.596, eliminating C and D as plausible distractors.

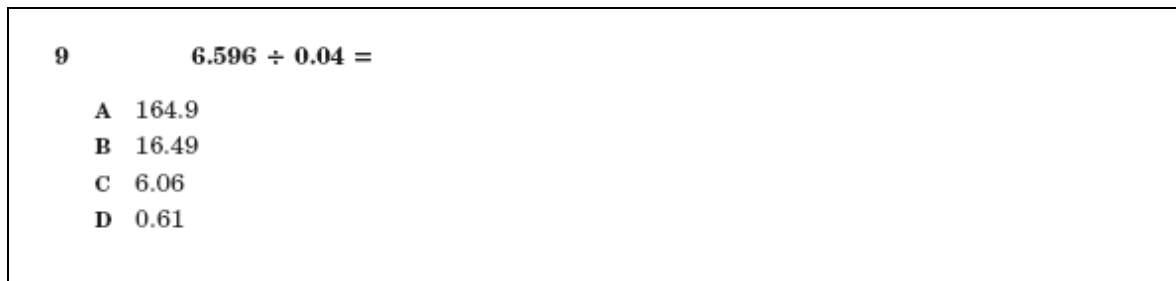


Figure 5. Sample Computation Item: Division of Decimals

Therefore, in this study we attempted to examine the impact on sample size and relative level of performance on the results of the distractor analysis. Four conditions were examined:

- different performance, different sample size
 - PEM sample (n=3,500) v. large standard sample (n = 30,000)
- different performance, similar sample size
 - PEM sample (n=3,500) v. small standard sample (n=3,500)
- similar performance, similar sample size
 - PEM sample (n=3,500) v. small matched standard sample (n=3,500)
- similar performance, different sample size
 - PEM sample (n=3,500) v. large matched standard sample (n=9,000)

Across the four conditions, the 50 multiple-choice items on the PEM and standard test forms, the method of analysis, and the criterion for identifying items also remained constant. Of course, also held constant across the four comparisons was the PEM sample and its performance on the items. The factor that varied was the comparison group for the PEM sample's performance.

Procedures

For these analyses, we used the expanded approach proposed by Kopriva et al. Given the large sample sizes in each group and the use of operational test items, small response frequencies for a particular cell was not a great concern. Each of the 50 items on the standard and PEM test forms

had four response options—one correct option and three incorrect options. Therefore, the chi-square analyses examined a series of 2-by-3 tables with students completing the standard and PEM tests as the two groups, and the three incorrect responses for each item as the response categories.

Samples of students, when needed, were drawn without replacement from the pool of students completing the standard test form. The samples were matched on the basis of raw score distribution. Given the relatively low performance of students completing the PEM, this did limit the maximum size of the sample matched on performance but different in sample size. That is, sampling without replacement, it was not possible to draw a large matched standard sample of more than 9,000 students and also replicate the raw score distribution of the students completing the PEM. Basic summary statistics on the performance of the five groups are provided in Table 8.

Table 8. Summary Performance Statistics For Distractor Analysis Samples

	% Passing	Scaled Score Mean (SD)	Number Correct Mean (SD)
PEM	20%	347.35 (68.17)	23.92 (8.78)
Standard Sample, Large	56%	414.43 (83.02)	34.20 (9.21)
Standard Sample, Small	56%	413.61 (80.58)	34.16 (8.95)
Matched Sample, Small	16%	330.89 (67.87)	23.97 (8.74)
Matched Sample, Large	18%	341.57 (64.58)	25.40 (8.33)

Selected background characteristics of the matched performance group in comparison to the PEM group are provided in Table 9.

Table 9. Selected Background Characteristics of the Plain English Mathematics Group and the Matched Samples Used in the Distractor Analysis

	Plain English	Matched, Small	Matched, Large
Number of students	3,420	3,410	8,965
Gender			
Female	37%	50%	50%
Male	63%	50%	50%
Disability	77%	10%	9%
LEP	35%	4%	5%
Race			
American Indian or Alaska Native	<1%	<1%	<1%
Asian	9%	2%	2%
Black (Not of Hispanic Origin)	22%	38%	38%
Hispanic	24%	7%	8%
White (Not of Hispanic Origin)	43%	51%	49%
Native Hawaiian/Other Pacific Islander	<1%	<1%	<1%
Unspecified	2%	1%	2%

Note: Percentages are rounded to the nearest whole number.

Results

The results of the distractor analysis, summarized in Table 10, show that there is a relationship between the composition of the comparison group and the number of items identified as problematic through the distractor analysis. Most clearly, many fewer items are identified when the PEM and comparison groups are matched on performance, regardless of the size of the comparison sample. When the comparison sample was matched for size alone, nearly 50 percent of the items were identified (a result consistent with the findings of previous studies). However, when the groups were matched on performance, the percentage of items identified shrinks to approximately 10 percent.

Table 10.: Comparison of Distractor Analysis Results Effect Size Classifications by Comparison Sample Number of Items Identified

Comparison Group	Effect Size		
	Small	Medium	Large
Standard, Large	36	14	0
Standard, Small	27	23	0
Matched, Small	45	5	0
Matched, Large	46	4	0

The results in Table 10 also suggest that the size of the sample has an impact on the number of items identified in the unmatched comparison groups. Nearly two-thirds more items were identified in the group matched on size (i.e., small) than in the large sample. A closer examination of those results indicates that the difference between the 14 items and 23 items identified is a function of the impact of sample size in the computation of effect size, *d*, in the equation

$$d = \sqrt{\frac{\chi^2}{n}}$$

As shown in Figure 6, there is a strong relationship between the effect size of the items in the small and large unmatched comparison groups. Of the 14 items identified in the large comparison group, 13 are also the items with the largest effect sizes in the small group. Also, the additional 9 items identified in the small group have effect sizes ranging from .08 to .11 in the large group (close to the small/medium borderline), and they have chi-square values that range from 42 to 175 and are significant ($p < .0001$). The difference in effect size appears to be due to the item level sample sizes for the large group being 4–5 times larger than those for the small group. The single item identified in the large group but not identified in the small group has a *p*-value greater than .90 and, consequently, a comparatively small sample size in the distractor analysis.

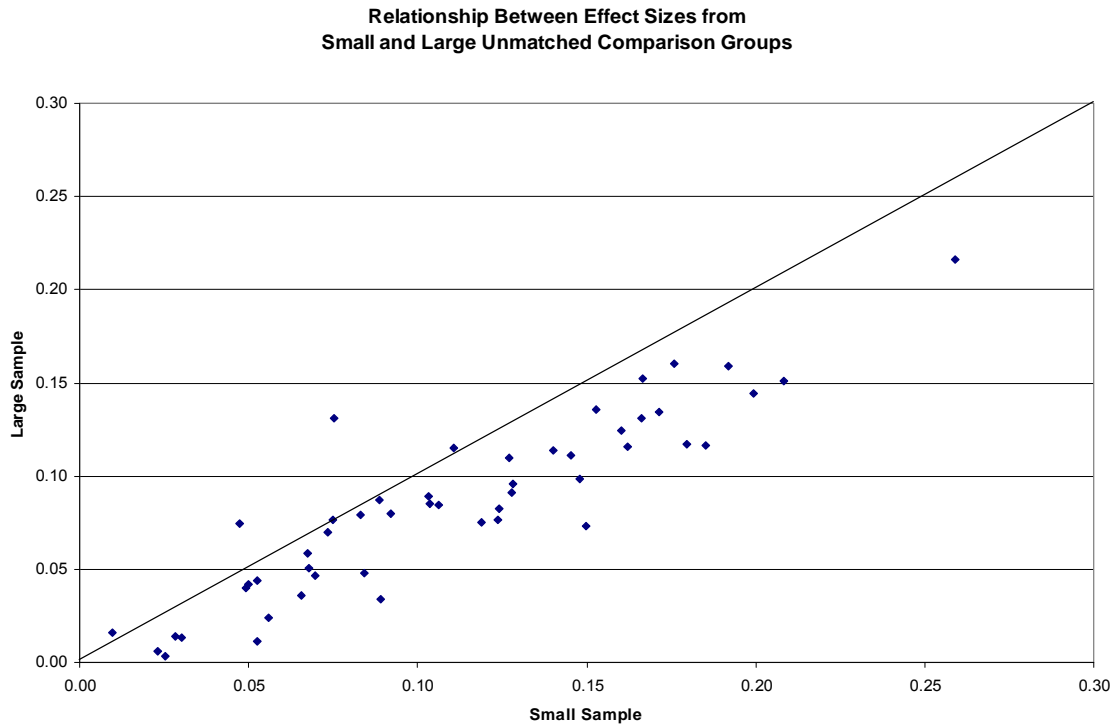




Figure 6: Relationship Between Effect Sizes from Small and Large Unmatched Comparison Groups

Items Identified through Distractor Analysis

Table 11 on the following pages presents the four pairs of items identified through the distractor analysis for the large matched comparison group. Information provided with the item includes the item key, p-value, and the percentages of students who chose each distractor. Note that the distractor percentages total 100 percent, and are based on the pool of students who responded incorrectly. The next step in the distractor analysis, consistent with DIF analysis, would be a review of the items by content, item development, and/or linguistic specialists to identify potential patterns among the items and explanations for the differences between groups.

For the purposes of this report, the items also provide examples of the types of linguistic modifications that were made to the original items. As in the case of Item 15, some modifications are quite minor, simply adding an abbreviation (GCF) that might be familiar to students.

Table 11. Items Identified through the Distractor Analysis

Standard Test Form	Plain English Mathematics Form																		
<p>6 Look at the table.</p> <p style="text-align: center;">Cost of Signs at Two Stores</p> <table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th>Store</th> <th>Neon Sign</th> <th>Wood Sign</th> </tr> </thead> <tbody> <tr> <td>A</td> <td>\$589</td> <td>\$227</td> </tr> <tr> <td>B</td> <td>\$534</td> <td>\$285</td> </tr> </tbody> </table> <p>What would be the <i>least</i> amount of money Jeremy's dad could spend if he bought one of each type of sign?</p> <p>F \$512 G \$761 H \$816 J \$819</p> <p>Key: G 44%, Distractors: F 40%, H 47%, J 13%</p>	Store	Neon Sign	Wood Sign	A	\$589	\$227	B	\$534	\$285	<p>6 Look at the chart.</p> <p style="text-align: center;">Cost of Signs at Two Stores</p> <table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th>Store</th> <th>Red Sign</th> <th>Blue Sign</th> </tr> </thead> <tbody> <tr> <td>A</td> <td>\$589</td> <td>\$227</td> </tr> <tr> <td>B</td> <td>\$534</td> <td>\$285</td> </tr> </tbody> </table> <p>Jena buys 1 red sign and 1 blue sign. What is the least amount of money she needs?</p> <p>F \$512 G \$761 H \$816 J \$819</p> <p>Key: G 27%, Distractors: F 23%, H 61%, J 16%</p>	Store	Red Sign	Blue Sign	A	\$589	\$227	B	\$534	\$285
Store	Neon Sign	Wood Sign																	
A	\$589	\$227																	
B	\$534	\$285																	
Store	Red Sign	Blue Sign																	
A	\$589	\$227																	
B	\$534	\$285																	
<p>15 What is the greatest common factor of 30, 42, and 48?</p> <p>A 2 B 3 C 6 D 8</p> <p>Key: C 77% Distractors: A 59%, B 20%, D 21%</p>	<p>15 What is the greatest common factor (GCF) of 30, 42, and 48?</p> <p>A 2 B 3 C 6 D 8</p> <p>Key: C 56% Distractors: A 46%, B 23%, D 31%</p>																		
<p>16 The picture shows the number of stars Angie received from her piano teacher for practicing.</p> <div style="text-align: center;">  </div> <p>What is the ratio of the number of striped stars to black stars?</p> <p>F 4 to 3 G 3 to 4 H 4 to 10 J 6 to 10</p>	<p>16 Look at the picture.</p> <div style="text-align: center;">  </div> <p>What is the ratio of white stars to black stars?</p> <p>F 4 to 3 G 3 to 4 H 4 to 10 J 6 to 10</p>																		

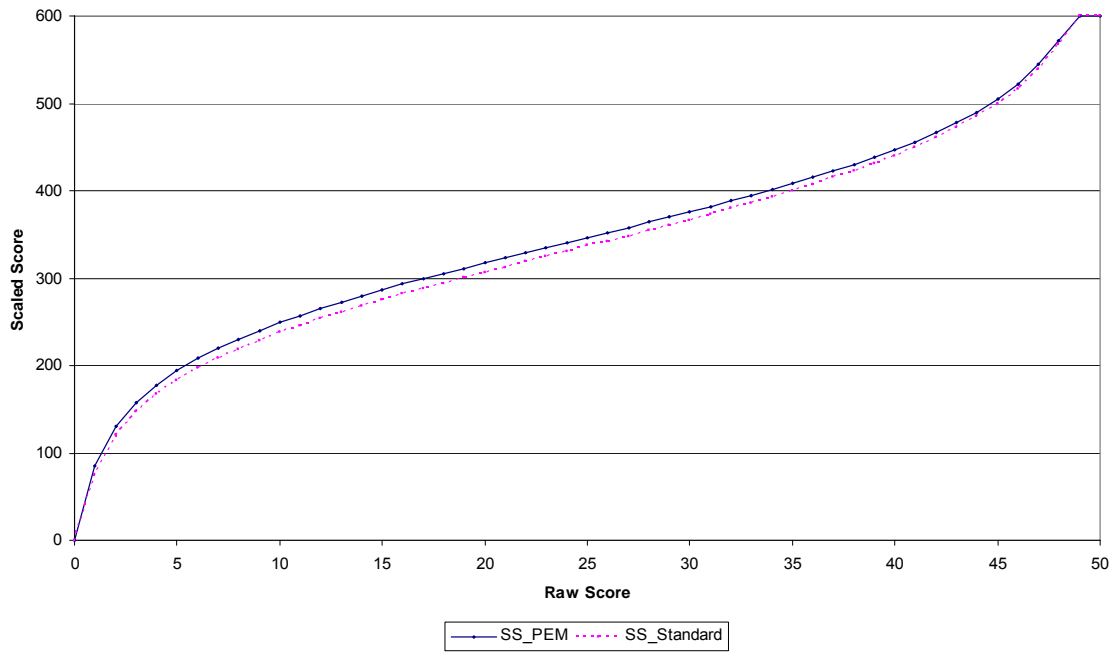
Standard Test Form	Plain English Mathematics Form																
Key: F 83% Distractors: G 68%, H 24%, J 8%	Key: F 78% Distractors: G 55%, H 27%, J 18%																
<p data-bbox="191 275 737 363">38 Mr. Warren requires his students to read 2 books: 1 book from list #1 and 1 book from list #2.</p> <table border="1" data-bbox="250 401 756 604"> <thead> <tr> <th data-bbox="250 401 483 443">List #1</th> <th data-bbox="483 401 756 443">List #2</th> </tr> </thead> <tbody> <tr> <td data-bbox="250 443 483 495"><i>A Trip to Asia</i></td> <td data-bbox="483 443 756 495"><i>Mystery at Chelsea</i></td> </tr> <tr> <td data-bbox="250 495 483 548"><i>Darlene's Hope</i></td> <td data-bbox="483 495 756 548"><i>Notes From Kent</i></td> </tr> <tr> <td data-bbox="250 548 483 604"><i>Sunset Hope</i></td> <td data-bbox="483 548 756 604"><i>A Clan of Many</i></td> </tr> </tbody> </table> <p data-bbox="250 638 737 695">What is the total number of different combinations for the 2 books?</p> <p data-bbox="250 726 305 856"> F 9 G 6 H 2 J 1 </p>	List #1	List #2	<i>A Trip to Asia</i>	<i>Mystery at Chelsea</i>	<i>Darlene's Hope</i>	<i>Notes From Kent</i>	<i>Sunset Hope</i>	<i>A Clan of Many</i>	<p data-bbox="878 285 1365 317">38 Jay has these 3 shirts and pants.</p> <table border="1" data-bbox="932 348 1352 552"> <thead> <tr> <th data-bbox="932 348 1136 390">Shirt Colors</th> <th data-bbox="1136 348 1352 390">Pants Colors</th> </tr> </thead> <tbody> <tr> <td data-bbox="932 390 1136 443"><i>White</i></td> <td data-bbox="1136 390 1352 443"><i>Brown</i></td> </tr> <tr> <td data-bbox="932 443 1136 495"><i>Green</i></td> <td data-bbox="1136 443 1352 495"><i>Black</i></td> </tr> <tr> <td data-bbox="932 495 1136 552"><i>Yellow</i></td> <td data-bbox="1136 495 1352 552"><i>Blue</i></td> </tr> </tbody> </table> <p data-bbox="932 583 1422 674">What is the total number of different combinations of 1 shirt and 1 pair of pants?</p> <p data-bbox="932 699 987 829"> F 9 G 6 H 2 J 1 </p>	Shirt Colors	Pants Colors	<i>White</i>	<i>Brown</i>	<i>Green</i>	<i>Black</i>	<i>Yellow</i>	<i>Blue</i>
List #1	List #2																
<i>A Trip to Asia</i>	<i>Mystery at Chelsea</i>																
<i>Darlene's Hope</i>	<i>Notes From Kent</i>																
<i>Sunset Hope</i>	<i>A Clan of Many</i>																
Shirt Colors	Pants Colors																
<i>White</i>	<i>Brown</i>																
<i>Green</i>	<i>Black</i>																
<i>Yellow</i>	<i>Blue</i>																
Key: F 69% Distractors: G 76%, H 18%, J 6%	Key: F 45% Distractors: G 56%, H 34%, J 10%																

References

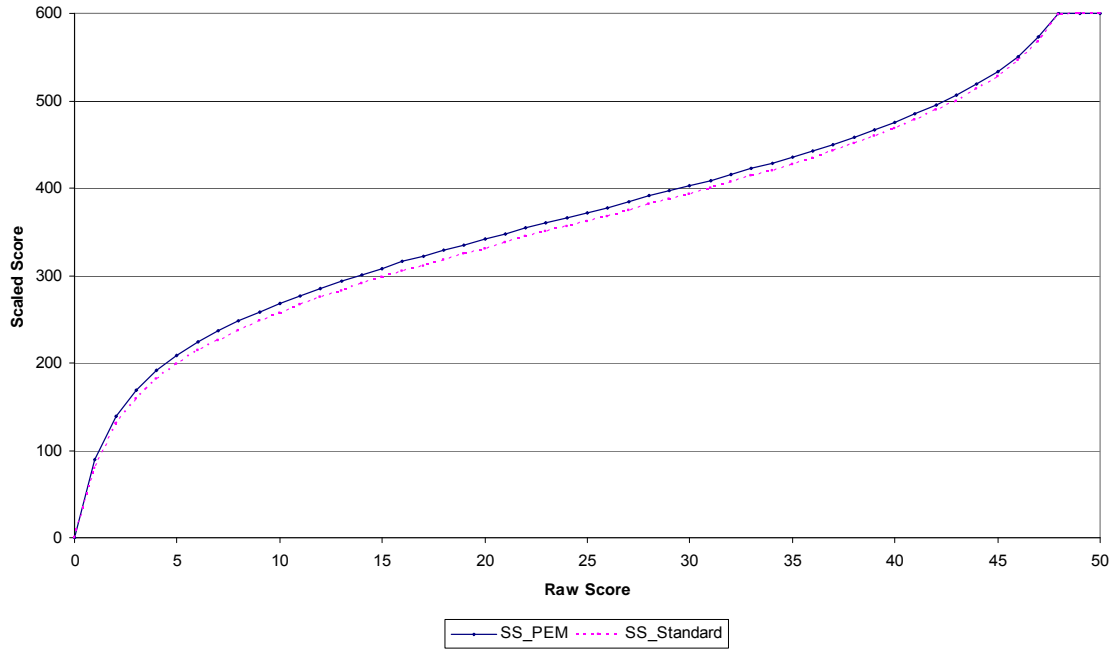
- Abedi, J, Leon, S., & Kao, J. (2008). *Examining differential distractor functioning in reading assessments for students with disabilities*. (CRESST Tech. Rep. No 744). Los Angeles: University of California. National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- duToit, M. (ed.) (2003). IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT. Lincolnwood, IL: Scientific Software International, Inc.
- Hambleton, R.K. & Swaminathan, H (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.
- Kolen, M.J. & Brennan, R.L. (2004). *Test equating, scaling, and linking: Method and practices (Second Edition)*. New York: Springer.
- Kopriva, R. J., Cameron, C, Carr, T, & Taylor, M. (2008). The limits of DIF: Why this item evaluation tool is flawed for English language learners, hearing impaired, and students with learning disabilities. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.
- MULTILOG for Windows (2003). Version 7.0.2327.3. Scientific Software International, Inc.
- Shaftel, J., Belton-Kocher, E., Glasnapp, D. R., & Poggio, J. P. (2003). *The differential impact of accommodations in statewide assessment: Research summary*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved January 1, 2009 from the World Wide Web: <http://education.umn.edu/NCEO/TopicAreas/Accommodations/Kansas.htm>.
- Sinharay, S & Holland, P.W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement*, 44(3), 249–275.

Appendix A

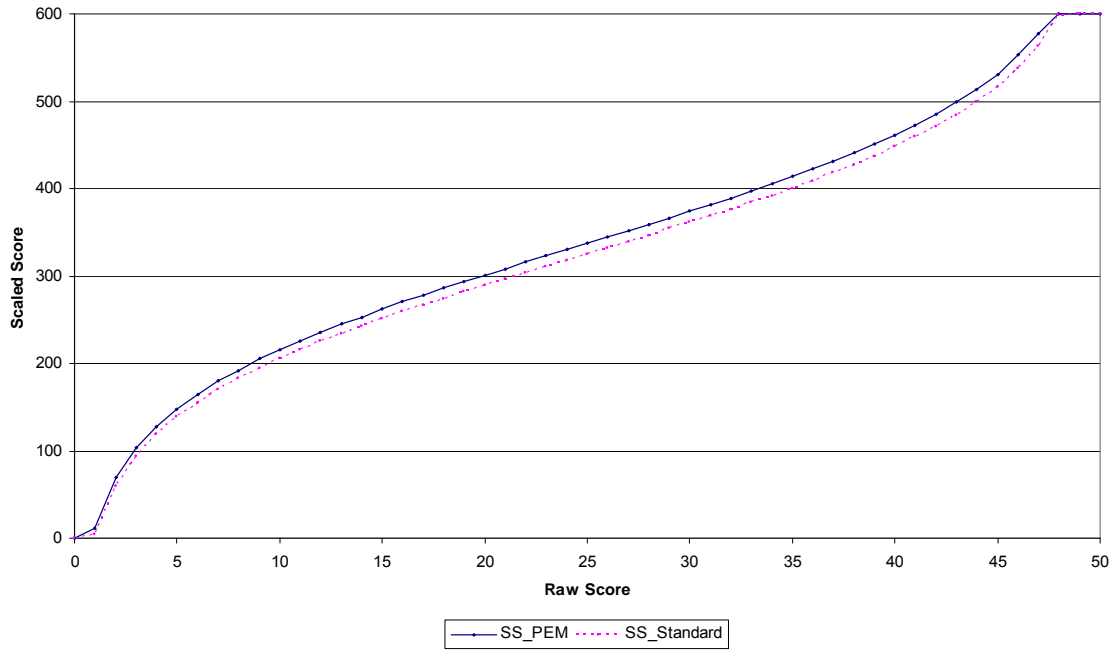
**Raw Score to Scaled Score Conversions: Grade 3
Standard and Plain English Forms**



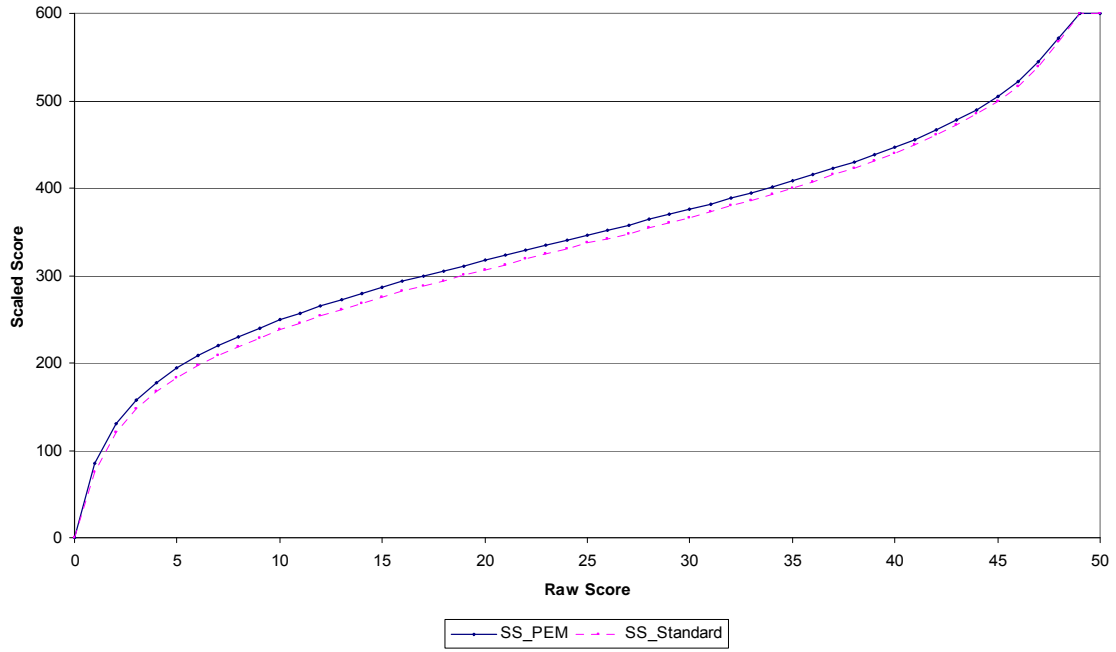
**Raw Score to Scaled Score Conversions: Grade 4
Standard and Plain English Forms**



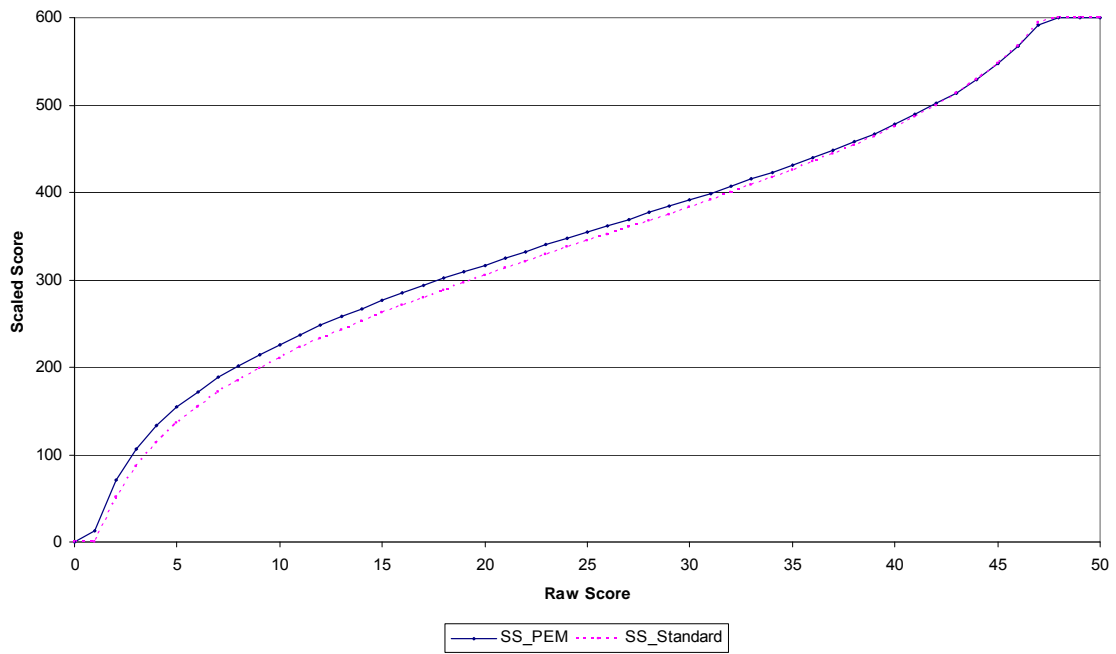
**Raw Score to Scaled Score Conversions: Grade 5
Standard and Plain English Forms**



**Raw Score to Scaled Score Conversions: Grade 7
Standard and Plain English Forms**



**Raw Score to Scaled Score Conversions: Grade 8
Standard and Plain English Forms**



Chapter 5: Evaluating the Comparability of Scores from an Alternative Format

Karen E. Barton
CTB/McGraw Hill

Phoebe C. Winter
Pacific Metrics

The focus of this planned set of studies was the evaluation of methods for determining the comparability of scores from a checklist-based assessment to those from the general assessment. Because of implementation fidelity and other practical issues, the methods were not used to evaluate comparability. However, because the design of the studies provides some direction for future research in this area, and a description of the implementation issues encountered here may be instructive for those carrying out such research, we have included descriptions of the design and issues in this chapter.

The state uses a checklist of grade level objectives in reading and mathematics, called here the Assessment Check List (ACL), supported by student work, to evaluate grade level achievement for some English language learners (ELLs) and some students with disabilities. These students are being taught grade level content and their performance is measured against the general grade level achievement standards, but they are not able to take the general test even with appropriate accommodations. Examples of such students include recently blinded students, some students with autism, and some students with physical disabilities that cannot be accommodated on the general test. English language learners identified as below Intermediate High in reading may participate in the ACL for up to two years. In the context of NCLB, the ACL is an alternate assessment of grade level achievement standards.¹²

The ACL

The ACL is a checklist of student performance on the content standards measured by the state's general test. The objectives within the content standards are sorted into four or five strands, depending upon subject area and grade level (e.g., algebra, reading comprehension). Each ACL-eligible student's teacher collects information related to the student's achievement of the objectives over the course of the school year. During the last 15–30 days of the school year (the last 15 days for half-year high-school courses), the teacher rates the student's performance on each objective based on the collections of student work, using the four-point rubric shown in Table 1.

¹² There are three types of alternate assessment defined by NCLB peer review guidance (U.S. Department of Education, July 2007): alternate assessment based on alternate achievement standards for students with significant cognitive disabilities, alternate assessment based on modified achievement standards for certain students whose disabilities have precluded them from achieving grade level proficiency and who meet criteria set by the state, and alternate assessment based on grade level achievement standards.

Table 1. Objective Level Scoring Rubric¹³

Level	Objective Performance Definition
Level 4	After appropriate modifications and accommodations, students performing at this level independently complete the requirements of the objective. Students performing at this level consistently demonstrate evidence of the depth of knowledge and skills necessary to comprehend the objective at a more complex level. Students are consistently and flexibly selecting appropriate strategies to understand and internalize the content.
Level 3	After appropriate modifications and accommodations, students performing at this level require minimal assistance to complete the requirements of the objective. Students performing at this level demonstrate sufficient knowledge and skills necessary to comprehend the objective. Students generally select, with occasional assistance, appropriate strategies to understand and internalize the content.
Level 2	After appropriate modifications and accommodations, students performing at this level generally require frequent assistance to complete the requirements of the objective. Students performing at this level demonstrate inconsistent evidence of the knowledge and skills necessary to comprehend the objective. Students occasionally select, with frequent assistance, appropriate strategies to understand and internalize the content.
Level 1	After appropriate modifications and accommodations, students performing at this level generally require continual assistance to complete the requirements of the objective. Students performing at this level demonstrate inadequate evidence of the knowledge and skills necessary to comprehend the objective. Students lack the ability to select appropriate strategies to understand and internalize the content.

After rating each objective, the student’s teacher considers the objective level profile within each content strand and holistically rates the student’s achievement on each strand, using a different four-point rubric, shown in Table 2.

Table 2. Strand Level Scoring Rubric

Score	Strand Performance Definition
4	Using appropriate instructional accommodations and/or modifications, the student, using various strategies, independently demonstrates an advanced understanding of subject matter content.
3	Using appropriate instructional accommodations and/or modifications, the student, using various strategies with minimal assistance, demonstrates an adequate understanding of subject matter content.
2	Using appropriate instructional accommodations and/or modifications, the student, using various strategies with frequent assistance, demonstrates a limited understanding of subject matter content.
1	Using appropriate instructional accommodations and/or modifications, the student, using strategies with continual assistance, demonstrates a lack of understanding of subject matter content.

Once the first set of ratings is completed by the teacher, a second rater reviews the student’s work and rates the student’s performance on each strand using the rubric shown in Table 2. Note

¹³ Labels for all sets of score levels have been changed to generic labels for the purpose of this report.

that the second rater does not rate the student’s achievement on each objective. If the two raters differ in their rating by more than one point, i.e., the strand scores are non-adjacent, a third rater reviews the collection of student work and rates the strand. Final strand level scores are submitted to the state department of education. The state applies weights to each strand score to mirror the weights given to each strand on the general test. Total scores range from 4 to 16.

Standard Setting

The state set cut scores on the ACL using a reasoned judgment approach (Cizek, 2001). Panelists with expertise in the content area and in teaching ACL-eligible students first discussed the ACL and the standard setting procedure. For each ACL grade level and content area, panelists reviewed the achievement level descriptors (mathematics grade five is shown in Table 3 for illustrative purposes) and six realistic exemplar scored checklists, without any accompanying student work. Panelists then recommended cut scores on the 4–16 point scale for each of four achievement levels, which are identical to those of the state’s general test.

Table 3. Grade Five Mathematics Achievement Level Descriptors

<p>Level 4</p>	<p>Students performing at this level consistently perform in a superior manner clearly beyond that required to be proficient at grade level work.</p> <p>Students performing at Level 4 commonly show a high level of understanding, compute accurately, and respond consistently with appropriate answers or procedures. They demonstrate flexibility by using a variety of problem-solving strategies.</p> <p>Students consistently demonstrate number sense for rational numbers 0.001 through 999,999. They consistently demonstrate ability in the addition, subtraction, comparison, and ordering of fractions, mixed numbers, and decimals. They correctly estimate the measure of an object in one system given the measure of that object in another system. Students commonly identify, estimate, and measure the angles of plane figures and commonly identify angle relationships. They consistently identify, define, and describe the properties of plane figures, including parallel lines, perpendicular lines, and lengths of sides and diagonals. Students are commonly able to identify, generalize, and extend numeric and geometric patterns. To solve problems, fifth graders at Level 4 consistently organize, analyze, and display data using a variety of graphs. They consistently use range, median, and mode to describe multiple sets of data. Students commonly use algebraic expressions to solve one-step equations and inequalities. They commonly identify, describe, and analyze situations with constant or varying rates of change.</p>
<p>Level 3</p>	<p>Students performing at this level consistently demonstrate mastery of grade level subject matter and skills and are well prepared for the next grade level.</p> <p>Students performing at Level 3 generally show understanding, compute accurately, and respond with appropriate answers or procedures. They use a variety of problem-solving strategies.</p> <p>Students generally demonstrate number sense for rational numbers 0.001 through 999,999. They generally demonstrate ability in the addition, subtraction, comparison, and ordering of fractions and decimals. They usually make correct estimates of the measure of an object in one system given the measure of that object in another system. Students generally identify, estimate, and measure the angles of plane figures and generally identify angle relationships. They generally identify, define, and describe the properties of plane figures, including parallel lines, perpendicular lines, and lengths of sides and diagonals. Students are usually able to identify, generalize, and extend numeric and geometric patterns. To solve problems, fifth graders at Level 3 generally are able to organize, analyze, and display data using a variety of</p>

	graphs. They generally use range, median, and mode to describe multiple sets of data. Students generally use algebraic expressions to solve one-step equations and inequalities. They generally identify, describe, and analyze situations with constant or varying rates of change.
Level 2	<p>Students performing at this level demonstrate inconsistent mastery of knowledge and skills in this subject area and are minimally prepared to be successful at the next grade level.</p> <p>Students performing at Level 2 typically show some evidence of understanding and computational accuracy and sometimes respond with appropriate answers or procedures. They demonstrate limited use of problem-solving strategies.</p> <p>Students demonstrate inconsistent number sense for rational numbers 0.001 through 999,999. They demonstrate limited ability in the addition, subtraction, comparison, and ordering of fractions and decimals. They inconsistently estimate the measure of an object in one system given the measure of that object in another system. They sometimes correctly identify, estimate, and measure the angles of plane figures and sometimes correctly identify angle relationships. Students inconsistently identify, define, and describe the properties of plane figures, including parallel lines, perpendicular lines, and lengths of sides and diagonals. Students are sometimes able to identify, generalize, and extend numeric and geometric patterns. In problem solving, fifth graders at Level 2 inconsistently organize, analyze, and display data using a variety of graphs. They have inconsistent success using range, median, and mode to describe multiple sets of data. Students sometimes are able to use algebraic expressions to solve one-step equations and inequalities. They inconsistently identify, describe, and analyze situations with constant or varying rates of change.</p>
Level 1	<p>Students performing at this level do not have sufficient mastery of knowledge and skills in this subject area to be successful at the next grade level.</p> <p>Students performing at Level 1 usually show minimal understanding and computational accuracy and often respond with inappropriate answers or procedures. They rarely use problem-solving strategies. Students rarely demonstrate number sense for rational numbers 0.001 through 999,999. They rarely demonstrate ability in the addition, subtraction, comparison, and ordering of fractions and decimals. They seldom can estimate the measure of an object in one system given the measure of that object in another system. They rarely identify, estimate, and measure the angles of plane figures and rarely identify angle relationships. Students rarely identify, define, and describe the properties of plane figures, including parallel lines, perpendicular lines, and lengths of sides and diagonals. Students are seldom able to identify, generalize, and extend numeric and geometric patterns. In solving problems, fifth graders at Level 1 rarely organize, analyze, and display data using a variety of graphs. They rarely are able to use range, median, and mode to describe multiple sets of data. Students rarely use algebraic expressions to solve one-step equations and inequalities. They rarely identify, describe, and analyze situations with constant or varying rates of change.</p>

Methods for Evaluating Comparability

Following Lottridge, Nicewander, Schulz, and Mitzel (2009, this volume), several conditions were formulated as criteria for investigating whether achievement level scores from the alternative format were comparable to those obtained from the general test¹⁴:

¹⁴ As noted in the introduction, because of the set of studies' focus on meeting NCLB peer review requirements, scores on the general test were the standard against which comparability was evaluated, rather than a more general

1. The ACL should measure the same constructs described in the achievement levels as the general test does.
2. Achievement level scores from the ACL should be based on the same degree of technical quality as those from the general test.
3. The ACL should require the same level of knowledge and skills to obtain each achievement level as the general test does.
4. The ACL and the general test should correlate the same with external criteria related to the tested construct described by the achievement levels.

Study Designs and Planned Analyses

The state department staff implemented a study and the researchers worked with department staff to design analyses and additional studies to evaluate the degree to which the ACL met these criteria. One goal of the studies was to compare the types of information and the inferences about comparability that could be made from the results of each study. Because of implementation fidelity issues and lack of teacher response to surveys, not all the studies were carried out. Issues preventing the researchers from completing these studies are described. The studies are discussed in this report because their designs serve as illustrations of methods that could be used to evaluate comparability, and the discussion of implementation issues may assist states in designing comparability studies.

1. Construct equivalence: The two assessments are designed to measure performance on the same set of objectives and standards, and strand scores on the ACL are weighted to match the weighting on the general test. To confirm that the student work on which ACL scores are based does cover the same objectives and standards, a study was designed to review sample collections of student work. In addition, an investigation of the relationship of strand scores on the general test and the ACL was planned from a study in which a sample of ineligible, general education students was assessed using the ACL.
2. Technical quality: Results of the study of general education students assessed with the ACL were intended to be used to compare selected technical characteristics of the ACL to those from the general test for these students.
3. Knowledge and skills required to meet achievement standards: Researchers planned to compare the achievement level scores from the ACL and general test for the group of students participating in the study. A study was designed to review the content of collections of student work from this group and compare them to test requirements at the cut scores.
4. Relationships with external criteria: A study was designed to compare student achievement level scores from the ACL and general test with teacher ratings of student performance on the skills covered by the assessments.

conception of comparability that might involve evaluating the degree of correspondence using a two-way framework (e.g., based on questions such as: In what ways are the two sets of scores comparable? How might the inferences from the scores be differentially useful in different situations?).

Comparing Scores from the ACL and the General Test

In the spring of 2007, the teachers in grades 3–8 who had a student eligible for the ACL completed an additional ACL in either reading or mathematics for one student in their classes who was not eligible for the ACL and would take the general assessment at the end of the year. This resulted in a matched set of data for each ACL-ineligible student that included ACL scores and scores on the general test. General test data included item, strand, total test, and achievement level scores. ACL data did not include task level (akin to item level) data but did include objective, strand, total test, and achievement level scores. Teachers in the study also submitted the collections of student work on which the ACL scores were based. A total of 605 and 865 students in reading and mathematics, respectively, participated in the special study. Qualitative information was also available for review and included test design documents (standards, blueprints, administration manuals, etc.), technical reports, and standard setting reports.

The researchers planned to use the data available from this study to investigate the criteria for score comparability noted above as follows:

1. Construct equivalence: comparison of scores within strand for students who took the checklist and the general test
2. Technical quality: analysis of technical properties of the checklist (e.g., rater agreement, reliability, classification consistency) compared to the technical properties of the general test
3. Knowledge and skills required to meet achievement standards: comparison of total scores, achievement level scores, and strand scores from the general test and the ACL

Judgment-based Review of Correspondence of Required Knowledge and Skills

Researchers developed a procedure for comparing the knowledge and skills required to meet each cut score on the ACL and the general test. The plan involved having teachers who had taught both ACL-eligible students and ACL-ineligible students in the relevant content area and grade level review materials and determine what scores on the ACL corresponded to each cut score on the general test. The panel would first be oriented to the process and provided a brief overview of the general test and ACL, including standards assessed, administration conditions and formats, and scoring procedures, followed by a review of sample ACL collections of work and a sample of the general test. Panelists were to review the achievement level descriptors and work in small groups to annotate them with additional information about required knowledge and skills based on an item map from the general test. Panelists would then review maps showing ACL objective level difficulty, accompanied by samples of student work. After discussion, the panelists would independently select the scores on the ACL that corresponded to the cut scores on the general test, based on equivalent knowledge and skills required. Several rounds of review and discussion were planned. Because this was a new method and the researchers wanted to evaluate the procedures, a number of evaluation activities were planned throughout the process.

The researchers planned to use the data available from this study to investigate the criteria for score comparability noted above as follows:

1. Construct equivalence: comparison of knowledge and skills required by the ACL and the general test, in terms of achievement level descriptor requirements
3. Knowledge and skills required to meet achievement standards: comparison of the types and degrees of knowledge and skills required to meet each achievement standard on the ACL and the general test

Relationship to Teacher Ratings

Researchers developed a survey of student performance on the knowledge and skills covered by the ACL and the general tests, using a methodology successfully implemented in earlier studies (Chen and Winter, June 2004). Four surveys were developed, for grade three reading and grades three, four, and six mathematics. Each survey item was based on the state's objectives in the appropriate grade level and content area, with the survey items structured so each asked about student performance at approximately the same grain size. The survey was web based, and teachers were asked to select a student at random, using a random selection procedure provided on the website, and rate the student on his or her classroom performance on each item in the survey, using a 3-point scale: below grade level, at grade level, or above grade level. Teachers were also given the option to select "not applicable for this student." The use of a 3-point scale rather than a 4-point scale, as in the state's other achievement level rubrics, was purposeful. This state has a mature accountability system, and to some extent, the achievement level descriptors have become reified, and in some respects almost normative. The teachers, when asked to rate students on the 4-point scale for contrasting-groups-type judgment procedures almost perfectly mimic the statewide distribution from the previous year. Toward the end of the survey window, teachers were asked to complete the survey for a student who had taken the ACL that spring (or another test that was the focus of a different study in the state) to increase the number of responses about those students. A total of 260 teachers responded to the surveys.

The researchers planned to use the data available from this study to investigate the fourth criterion for score comparability noted above as follows:

4. Relationships with external criteria: comparisons of the relationships of student scores from the ACL and general test with teacher ratings of student performance on the skills covered by the assessments

Summary of Planned Studies

Table 4 shows the planned studies and the criteria they were designed to address.

Table 4. Studies and Criteria

Criterion	Comparison of Scores	Judgment-based Review	Teacher Surveys
1. Construct equivalence	X	X	
2. Technical quality	X		
3. Knowledge and skills required to meet achievement standards	X	X	
4. Relationships with external criteria			X

Discussion: Implementation Issues

Researchers and teachers examined the content of the collections of student work submitted by teachers who had administered the ACL to ineligible students for the special study. The types of work collected were not representative of those expected from students who are eligible to take the ACL. Instead, the student work for a large number of the students in the study consisted of multiple-choice tests and work sheets. The nature of student work collected for ACL-ineligible students affected the planned studies that were based on comparisons of scores and affected the judgment-based review—the researchers could not be sure that teachers scored the work as intended or that ACL scores for the students represented what students would receive if the work had been more consonant with the intent of the ACL.

Careful training and ongoing monitoring of implementation may have improved the quality of ACL study results. Initial training should include a discussion of the purpose of the study and clear expectations for the type of student work to collect as well as thorough orientation to scoring procedures. Teachers should have access to assistance throughout the study period, through websites, message boards, and periodic contact with study researchers. At several points within the study period, researchers should check the contents of collections of student work and give teachers feedback and advice on their collection strategies.

The teacher survey was administered toward the end of the school year so that researchers could collect end-of-year data about student performance. Teachers were provided with incentives to participate (a \$10 gift card, plus a chance to win one of two \$100 gift cards). Communication about the survey from the state department of education was sent to district superintendents and building principals. The number of teachers who responded to the survey of student skills was much lower than expected—at least 50 responses for each test variation per grade level/content area combination. Only 8 teachers of the 260 who did respond across grade levels and content areas completed the survey for students who had taken the ACL. There were not enough responses for these students to compare the relationships between test scores and teacher ratings.

Several factors may have affected the response rate. Information about the survey may not have reached many teachers; direct solicitation of teachers, incorporating a clear explanation of the purpose of the study and the need for participation, with follow-up, may have increased the number of responses. When teachers reached the survey website, they had to register and wait

for an email with a code that would allow them to complete the survey. This two-stage process may have deterred some teachers; only 42 percent of teachers who registered completed the survey. The timing of the survey may have been too close to the end of the school year, when teachers are busy with year-end activities. Some teachers did not feel comfortable submitting confidential information about their students, even with the assurance that identifying information would be stripped from the data once survey responses were matched to test score data. In addition, the incentives may not have been sufficient, especially without providing strong reasons for participation related to the improvement of assessments for ACL-eligible students.

Conclusions

In conducting future research on similar instruments, there are lessons learned from these studies. Most importantly, it is critical that the data collected in such an assessment be based on high fidelity and at the smallest grain size of data possible. This will allow the state to collect specific evidence about the understanding by teachers of the process and consistency in the type of student work collected as well as the utilization of the scoring rubrics. For example, that the special study used teachers who were trained in collecting ACL was not a guarantee that those same teachers would provide the type of work necessary for comparison for their ACL-ineligible students. Assessment methods that use similar approaches, such as portfolios, collections of student work, or observational checklists should carefully attend to logistical strategies. This will lend credence to the process and open the way for rigor in the research.

References

Chen, C., & Winter, P.C. (June, 2004). Planning and conducting cognitive laboratories for developing large-scale assessments. Presentation at the annual National Conference on Large-Scale Assessment, Boston.

Cizek, G.J. (Ed.) (2001). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum.

Lottridge, S.M., Nicewander, W.A., Schulz, E.M., & Mitzel, H.C. (2009, in preparation). Comparability of paper-based and computer-based tests: A review of the methodology. In Lottridge, S.M., Nicewander W.A., Schulz E.M., Mitzel H.C. (Ed.), *Evaluating the comparability of scores from achievement test variations*. Washington, DC: Council of Chief State School Officers.

US Department of Education, July 2007. *Modified academic achievement standards: Non-regulatory guidance*. Washington, DC: Author.

Chapter 6: Modified Tests for Modified Achievement Standards: Examining the Comparability of Scores to the General Test

Charles A. DePascale

National Center for the Improvement of Educational Assessment

In April 2007, federal regulations were adopted allowing states to develop modified academic achievement standards for some students with disabilities and use those standards for making adequate yearly progress (AYP) under No Child Left Behind (72 FR 67 17748-17781). Those regulations and corresponding U.S. Department of Education guidance clearly indicate that measuring modified academic achievement standards requires states to develop a modified test, different from the state's general test and different from the state's alternate assessment(s) measuring alternate achievement standards¹⁵ (USED, 2007). The modified test will cover the same content as the general test, but with less difficult questions overall. States may include in AYP calculations proficient and advanced scores from the modified tests based on modified achievement standards for up to 2.0 percent of the students assessed in a particular grade level, resulting in the label "2% test" commonly being applied to these modified tests.

Comparability, in the strict sense of interchangeability of results, is not a concern with assessments designed to measure modified achievement standards. There is no expectation that the modified test classifies students into achievement levels based on the same degree of knowledge and skills as the general test. There is a clear understanding that the modified test, by design, is intended to measure students against standards that are less difficult than a grade level academic achievement standard.

In practice, however, there is a need to understand the relationship between the general test and the modified test. The general test and the modified test are designed to measure student achievement of a state's academic content standards for the grade in which the student is enrolled. Results from the modified test are intended to provide teachers and parents with information that will help guide instruction that, ultimately, will help students achieve the state's grade level academic achievement standard. Understanding the relationship between the two tests and their results can only enhance the usefulness of the results of the modified test.

There are also basic validity questions regarding the modified test that are best answered by understanding the relationship between the general test and modified test. A modified test is designed to provide more precise measurement at a lower point along the grade level proficiency continuum. Examining the relationship between the two tests will provide information on the relative difficulty of the two tests and whether the modified test provides more reliable

¹⁵ The U.S. Department of Education calls these modified tests "alternate assessments of modified achievement standards." For brevity, and to avoid confusion with alternate assessments of alternate achievement standards, we use modified test or assessment.

measurement than the general test in the area of the proficiency continuum that it is designed to measure.

In this study, we use classical approaches and item response theory (IRT) methods to examine the relationship between a state's general tests and modified tests in reading and mathematics. Our analyses focused on the results of groups of students who completed both the general test and modified test. One set of analyses examined the extent to which students performed consistently across the two instruments. A second set of analyses used IRT to place the results of the corresponding general and modified tests on the same scale to examine the relative difficulty of the two tests and their relative measurement precision at various points along the proficiency continuum. The primary purpose of the study was to investigate whether the methods and analyses used here provide useful information to support the design, interpretation, and use of modified tests. A secondary purpose of the study was to provide information to the state to assist them in interpreting whether their test development approach resulted in modified tests that met their needs as well as meeting federal guidelines for modified tests.

Method

Modified Tests

The modified tests examined in this study were reading and mathematics tests administered at grades 3–7. The assessments were developed to serve as the state's alternate assessments based on modified academic achievement standards tests under NCLB. Although design and development of the tests began prior to the issuance of guidance on the modified tests by the United States Department of Education, the fundamental design of the tests was consistent with those guidelines. The assessments are intended to be administered to students with an IEP who also meet the following criteria (among others):

- The student is not identified as having a significant cognitive disability.
- The student is not likely to achieve the grade level academic achievement standard for proficiency within the current school year.

Like the general grade level tests, the modified assessments are administered in paper-and-pencil, multiple-choice format. However, there are key distinctions in the design of the tests and items:

- Items on the modified tests contain three response options rather than four (one correct choice, two incorrect choices).
- In general, more simplified English is used on the modified tests.
- Fewer items are included on the modified tests (40 items v. 50+ items).
- Fewer items are displayed on each page on the modified tests.

Consistent with federal guidance, the modified tests are designed to measure the same grade level content standards as the general test, but to be less *difficult* than the general tests—that is, measuring grade level content standards at a lower level on the proficiency continuum than the general tests. A comparison of the basic test blueprints for the modified and general tests

indicates that the balance of emphasis across content strands is consistent for the corresponding general and modified tests. Additionally, a cursory review of released items on the general and modified tests revealed that across content strands items appearing on the modified test were consistent with the less difficult items in the same strands appearing on the general test. Note that a full content review of the tests was beyond the scope of this study, but would be part of the design and development process for producing modified tests.

Data

The data used in this study were from a state's general and modified tests in reading and mathematics at grades 3–7. The data were collected as part of a special administration conducted by the state in the fall of 2006 outside of the state's regular spring test administration window¹⁶. Although schools and students participating in the fall administration were aware that there were no stakes associated with performance on the tests, the fall administration was conducted with the gravitas associated with a state administered assessment (Department of Education, 2006).

The fall administration was designed specifically to collect data that could be used to examine the relationship between the general and modified tests. Specifically, the modified tests were administered at each grade level to a random sample of students who had participated in the state's general tests the previous spring. The intent was to collect student performance data on the two tests that could be compared to evaluate the relationship between the corresponding general and modified tests.

Description of Special Administration Samples

At each grade level, samples of students who participated in the general test the previous spring were selected. Two samples of students were selected at each grade level. One sample, serving as the *experimental group*, completed the modified test. The other sample, serving as a *control group*, completed the general test a second time. The primary function of the control group was to provide comparative information about the amount of change in student performance that could be expected due to factors other than the test instrument (e.g., time lag from spring to fall, differences in motivation, and differences in administration conditions). In this study, the control group was also used in the IRT analyses to examine the relationship between the general and modified tests.

Approximately 1,500–2,000 students were sampled at each grade level. The number of schools included in each sample ranged from 25–30 at the lower grades to 7–10 at the higher grades, reflecting the size differences between elementary and middle schools. One grade was sampled

¹⁶ The state collected data at grades 3–8 for the special study, but grade 8 data were not used in these analyses. Large performance differences between the state's operational data and the fall special study samples were found at grade 8. Although the mean scaled score performance was similar across the samples, there were large differences in the achievement level results and in the demographics of the samples. These differences at grade 8 may be a reflection of the limitations of sampling at the school level with a small number of larger schools. The data collected at grades 3–7 were sufficient to meet the needs of this study.

per school. All students in the selected grade were required to participate in the special administration in accordance with the state’s participation guidelines for state assessments.

The samples selected were similar in performance to the statewide sample completing the general test. In general, the fall samples were slightly higher performing than the state sample, particularly in terms of achievement level results. Table 1 shows the mean scaled score and percentage of students performing at or above the proficient level on the general test in the state sample, experimental group, and control group.

Table 1. Performance on the Spring Administration of the General Test for the State and Selected Study Samples

Grade	State Spring Administration			Experimental			Control		
	Reading Mean Scaled Score (sd) n	Math Mean Scaled Score (sd) n	% at or above Proficient in Both Areas	Reading Mean Scaled Score (sd) n	Math Mean Scaled Score (sd) n	% at or above Proficient in Both Areas	Reading Mean Scale Score (sd) n	Math Mean Scaled Score (sd) n	% at or above Proficient in Both Areas
3	248.6 (8.8) 103,627	343.2 (9.7) 104,205	65.8	249.1 (8.7) 1,687	344.0 (9.8) 1,687	67.5	249.4 (8.3) 1,931	343.3 (9.2) 1,931	70.2
4	253.1 (8.6) 101,654	348.9 (9.5) 102,306	63.8	252.7 (8.5) 1,835	348.9 (9.4) 1,835	63.1	253.1 (8.6) 1,715	349.1 (9.5) 1,715	63.8
5	257.1 (7.8) 102,429	353.7 (9.2) 103,067	63.2	257.4 (8.0) 1,989	354.5 (9.7) 1,989	64.6	257.1 (7.6) 1,934	353.9 (9.2) 1,934	64.3
6	259.2 (8.1) 105,660	354.9 (9.7) 106,036	60.6	260.4 (7.3) 1,527	356.1 (8.6) 1,527	68.0	260.4 (8.4) 1,722	356.5 (10.1) 1,722	64.5
7	261.9 (8.6) 105,502	357.8 (9.6) 105,764	61.5	262.0 (8.1) 1,725	357.6 (9.0) 1,725	60.9	261.5 (8.2) 1,614	357.5 (9.0) 1,614	62.3

Overall, the fall samples were also similar to the state sample in terms of demographic characteristics, although more variation across samples was found on these factors than on student performance. Table 2 shows the distribution of students by sex, economic status, disability status, and English language proficiency status in the state sample, experimental group, and control groups.

Table 2. Comparison of Selected Demographic Characteristics for the State and Selected Study Samples

	Total N	Female	Percentage of Students		
			Economically Disadvantaged	Limited English Proficient	Students with Disabilities
Grade 3					
State*	103,627	49.0	48.3	7.2	16.5
Experimental	1,687	50.5	44.5	7.2	13.4
Control	1,931	46.3	39.4	8.8	20.4
Grade 4					
State	101,654	48.9	47.5	6.3	16.7
Experimental	1,835	50.1	51.0	9.8	10.7
Control	1,715	47.6	45.4	6.0	16.9
Grade 5					
State	102,429	48.5	46.9	5.7	16.6
Experimental	1,989	48.9	41.7	6.0	13.0
Control	1,934	47.4	39.4	4.7	12.3
Grade 6					
State	105,660	48.6	47.2	4.5	15.6
Experimental	1,527	50.0	46.2	7.7	15.9
Control	1,722	49.7	37.1	4.7	20.1
Grade 7					
State	105,502	48.8	45.6	4.1	15.0
Experimental	1,725	48.5	46.5	4.2	15.2
Control	1,614	48.9	46.7	4.4	17.1

*Figures for state sample at each grade level are based on students completing the reading test.

Data Analyses

Consistency of Performance

To evaluate the consistency of performance across tests, two statistics were examined: percentage of students classified at the same achievement level on the spring and fall tests and the Spearman rank order correlation between student performance on the spring and fall tests.

The Spearman correlation provides information on the extent to which students are ordered the same on both test administrations. Given that the general and modified tests are designed to measure the same grade level content standards, it was the state's expectation that there should be consistency in the ordering of students on the two instruments. Aside from measurement error, in general, possible factors impacting the correlation in this case could be the time lag between spring and fall, differences in motivation between test administrations, and the potential of a ceiling effect on the modified test administered to general education students.

The examination of the consistency of the classification of students across achievement levels serves a different purpose than the Spearman correlation. Although there is an expectation that students will be ordered the same on both tests, there is not the same expectation that students will be classified into the same achievement level on the general and modified tests.

Achievement levels for the general and modified tests are set independently, and, in most cases,

are designed to have different meanings. To the extent that the achievement levels are modified on the modified test, one would expect differences in the students' achievement level classifications across the tests. In general, one would expect students to receive a higher achievement level classification on the modified test than on the general test.

Relationship between the General and Modified Tests

To evaluate the relationship between the general and modified tests, IRT analyses were performed to link the two tests and place them on a common scale. When the tests are placed on a common scale, an examination of the test information functions (TIF) of the two tests will provide evidence of the extent to which the design expectations have been met. In general, the TIF should reveal differences between the general and modified tests in the amount of information provided at various points along the proficiency continuum. In particular, the modified test should provide more information (i.e., more precise measurement) at lower levels of the proficiency continuum than the general test. An examination of the TIF will also reveal whether the modified test is providing sufficient reliability in the area of the achievement level cut scores; and whether the modified test is providing more reliable measurement than the general test in those areas.

In this study, a random groups design was used to link the general and modified reading and mathematics tests at grades three and six. As described above, random samples of students were assigned either the general test (control group) or the modified test (experimental group) in the fall administration. IRT methods were used to scale student responses and generate results for the two groups of students. Specifically, items were calibrated using a 2-parameter logistic model with the assumption that the underlying ability distribution was normally distributed with a mean of 0 and standard deviation of 1. Multilog for Windows software (v. 7.0.3) was used to perform the item calibrations. In the random groups design, test forms are randomly assigned to students and each student completes a single test form. Differences in performance between the two groups are “taken as a direct indication of differences in difficulty between the forms” (Kolen & Brennan, 2004).

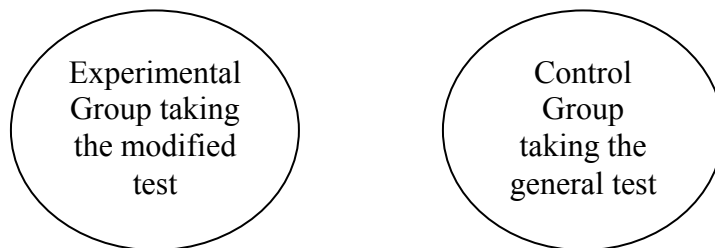


Figure 1. Random Groups Design to Link the General and Modified Tests

Of course, the equivalence of the random groups is critical to the design and interpretation of results. As shown in Table 1, student performance is similar between the control and experimental groups at each grade level, but there are some differences, particularly in terms of achievement level distributions. An alternative to using the control and experimental groups in the analysis was to use only the experimental group, linking the general test and modified test

through their performance in the spring on the general test and the fall on the modified test. This single group approach would offer the advantage of a single group of students. However, we decided that this advantage would be more than offset by the time lag between summer and fall and the potential for motivation differences between the two administrations.

To link the general and modified tests for operational purposes (e.g., to attempt to use scores from the two forms interchangeably), at a minimum much more care would be necessary in assuring the equivalence of the random groups. It is also possible that a different approach than the random groups design would be more appropriate. For the purposes of this study, however, to attempt to roughly describe the relationship in the difficulty of the corresponding general and modified tests, the use of the random groups design and the equivalence of the random groups sampled are appropriate and sufficient.

Results

Spearman Rank Order Correlation

Spearman rank order correlations between student scaled scores on the general test (completed in the spring) and modified test (completed in the fall) at each grade level were computed for the experimental group. For comparative purposes, correlations were also computed for the control group in which students completed the general test in both administrations. Results of the correlation analyses are shown in Table 3.

Table 3. Student Level Spearman Correlations between Spring Administration and Fall Special Study

Grade Level	Experimental Group Standard (Spring) & Modified (Fall) Tests (n)		Control Group Standard Test Both Administrations (n)	
	Reading	Mathematics	Reading	Mathematics
3	.72 (1,687)	.73 (1,687)	.82 (1,931)	.84 (1,931)
4	.72 (1,835)	.77 (1,835)	.85 (1,715)	.86 (1,715)
5	.74 (1,989)	.75 (1,989)	.81 (1,934)	.84 (1,934)
6	.68 (1,527)	.75 (1,527)	.83 (1,722)	.87 (1,722)
7	.76 (1,725)	.79 (1,725)	.80 (1,614)	.86 (1,614)

The correlation results indicate that there is a relationship between the ordering of students on the general and modified tests. As expected, correlations are higher for students taking the general test in both administrations (i.e., the control group), but are strong and positive between the general and modified tests as well. With few exceptions, results are consistent across grade levels. In both the control and experimental groups, the slightly higher correlations are found between the mathematics tests than the reading tests.

The correlations between the performances on the standard and modified tests are likely to be negatively impacted by a ceiling effect on the modified tests for the general population sample used in this study. Overall, score distributions were much more skewed on the modified tests than on the general tests. Figures 2 and 3 show scaled score distributions for the corresponding grade six reading and mathematics tests, respectively.

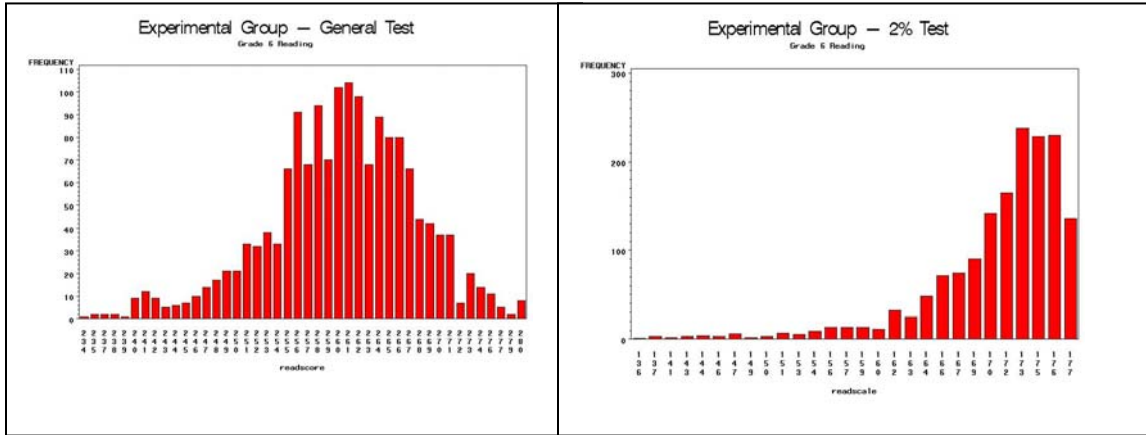


Figure 2. Scaled Score Distributions Grade Six Reading

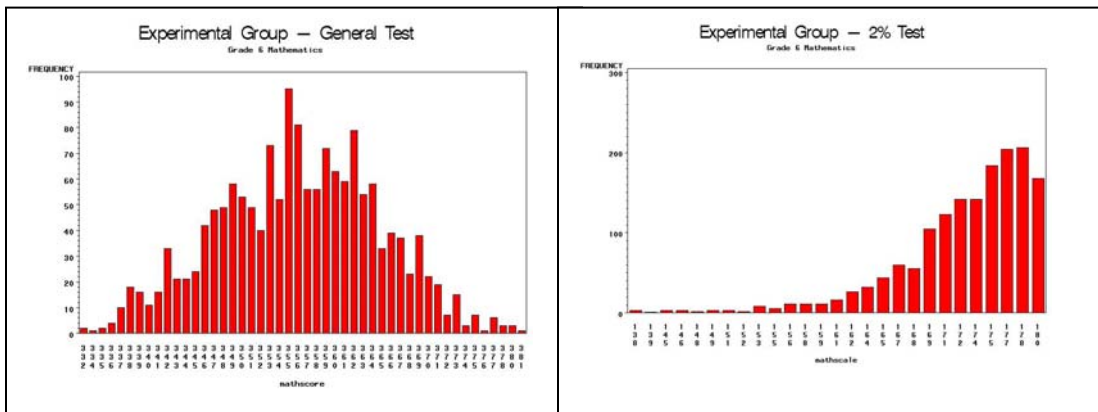


Figure 3. Scaled Score Distributions Grade Six Mathematics

Achievement Level Classifications

Student performance on each test is classified into one of four achievement levels. Table 4 shows the percentage of students classified into the same achievement level on each of the two tests that they completed. The results are fairly consistent across grade levels. The percentage of students assigned to the same achievement level is higher on the reading tests than on the mathematics tests in both the control and experimental groups, although the difference is much greater in the experimental group. With the exception of grade seven reading, the percentage of students

assigned to the same achievement level is higher within the control group (students taking the general test in both administrations) than in the experimental group (students taking the general test and the modified test).

Table 4. Percentage of Students Receiving the Same Achievement Level Classification, Spring and Fall Test Administrations

Experimental Group Standard (Spring) and Modified (Fall) Tests				Control Group General Test Both Administrations			
Grade Level	# Students	Reading	Mathematics	Grade Level	# Students	Reading	Mathematics
3	1,687	66.2	53.6	3	1,931	71.6	68.7
4	1,835	65.6	51.4	4	1,715	67.1	65.4
5	1,989	67.8	42.4	5	1,934	69.1	64.8
6	1,527	65.4	41.8	6	1,722	69.1	67.6
7	1,725	69.0	30.3	7	1,614	65.7	65.7

A primary function of the control group in this special administration was to provide a reference point, or baseline, for evaluating the results of the experimental group. It is clear from the results presented in Table 4 that the reading results are much more similar across the experimental and control groups than the mathematics results. With respect to comparability of results, in general, it is also worthwhile to consider these results in the context of the percentage of students one would expect to be assigned the same achievement level classification in a test-retest situation. Achievement level classification match on state tests is commonly estimated through a *consistency* statistic (Livingston and Lewis, 1995). Consistency is used to estimate the classification match between two independent administrations of parallel test forms. A small sample of technical reports from two state testing programs was searched for consistency figures for reading and mathematics tests at grades 3–8. Across 42 reading tests, consistency figures, reported as percentage of students with the same achievement level classification, ranged from 68 percent to 79 percent with a median of 74 percent. Across 42 mathematics tests, consistency figures, reported as percentage of students with the same achievement level classification, ranged from 64 percent to 75 percent with a median of 72 percent.

As discussed previously, the interpretation of these results is dependent upon the expectations established in the development of the achievement level descriptions and the process for setting achievement level cut scores for the tests. If the state’s goal were to establish achievement levels with similar meanings across the two tests then the results found on the reading tests might be expected and accepted as evidence of comparability. The mathematics results, in contrast, would not support a claim of comparability of achievement level classifications across the two tests. However, if the state’s goal were to establish lower achievement level standards on the modified test than on the general test then the results found on the mathematics test would be more likely.

Although the summary mathematics results in Table 4 do not show that students were classified into higher achievement levels on the modified test than on the general test, this can be easily confirmed through an examination of a two-way table of achievement level classifications on the two tests as shown in Table 5. The shaded cells below the diagonal in Table 5 show that 56.4 percent of the students received a higher achievement level classification on the modified test than on the general test, and the cells above the diagonal show that only 0.4 percent of students received a lower achievement level classification on the general test.

Table 5. Achievement Level Classifications on the General and Modified Tests Grade Six Mathematics

		General Test (Percentage of students)				
		Level 1	Level 2	Level 3	Level 4	Total
Modified Test	Level 1	0.1	0.1	0.1	0	0.3
	Level 2	1.1	1.2	0.1	0.1	2.5
	Level 3	4.2	20.8	22.0	1.4	48.4
	Level 4	0.1	2.4	27.8	18.5	48.8
Total		5.4	24.6	50.0	20.0	100.0

In the case of this state, the high level of agreement between achievement level standard classification on the general and modified reading tests is not unexpected given overall reading test results within the state. Across grades 3–8 in 2006, 83–91 percent of all students achieved the grade level academic achievement standard in reading, and there has been a significant increase in the percentage of students achieving the standard since 2001. Under those conditions, one might not expect major differences between the grade level and modified achievement level standards. In contrast, in mathematics, the percentage of students achieving the grade level academic achievement standard was closer to 60 percent across grade levels as a new, higher achievement standard was introduced in 2006.

Relationship between the General and Modified Tests

IRT analyses were conducted to link the general and modified reading and mathematics tests at grades three and six. The purpose of the analyses was to place the results on the same scale so that direct comparisons could be made regarding the measurement precision of the two tests at various points along the proficiency continuum. Specifically, the goals were to determine that the modified tests were less difficult than the general tests, and to determine that the modified tests provide more reliable information than the general test in the area of interest for its target population of the modified test.

Figure 4 contains test information function (TIF) for the general and modified tests in grade three reading. We estimate that the grade level academic achievement standard for the grade three reading test is near 0.0 on this scale. The graph shows that the TIF for both tests provide maximum information at a point below that grade level standard, with the general test TIF centered near -1.0 and the modified test TIF centered closer to -2.0. Using an information value of 10 as a standard for sufficiently reliable information, it is clear from the TIF that there is considerable overlap in the areas in which the general and modified tests provide reliable measurement. However, there are important differences between the tests as well. The modified test continues to provide reliable information at the extreme low end of the scale, which may be particularly important for the target population. Conversely, the general test provides reliable information and more information than the modified test in the area of the grade level academic achievement standard.

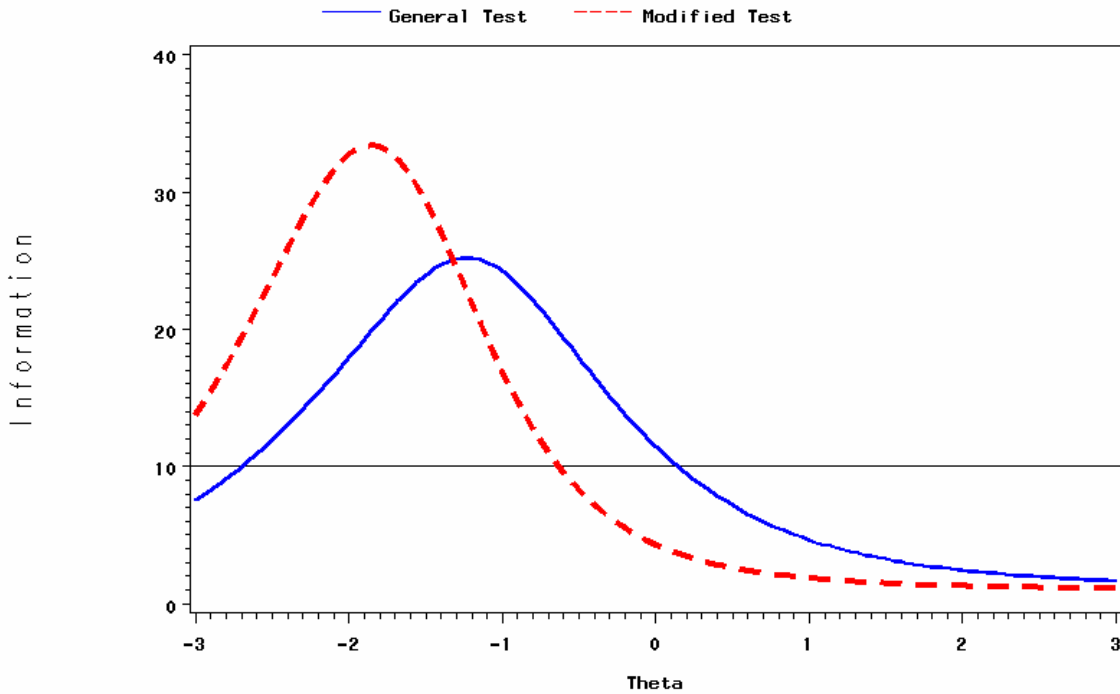


Figure 4. Test Information Function for General and Modified Tests: Grade Three Reading

Figure 5 contains TIF for the grade six mathematics general and modified tests. We estimate that the grade level academic achievement standard for the grade six mathematics test is approximately -0.5 on this scale. The general test, therefore, appears to be providing maximum information near the grade level achievement standard, and the modified test is providing maximum information well below this point on the proficiency continuum. In contrast to the corresponding graphs for the grade three reading tests shown in Figure 4, there is much less overlap in the areas where the general and modified mathematics test are providing maximum and sufficiently reliable information. The general test provides sufficient information in an approximate range between -2.0 and 1.0 on the scale and the modified test provide sufficient information from the low end of the scale to approximately -1.0.

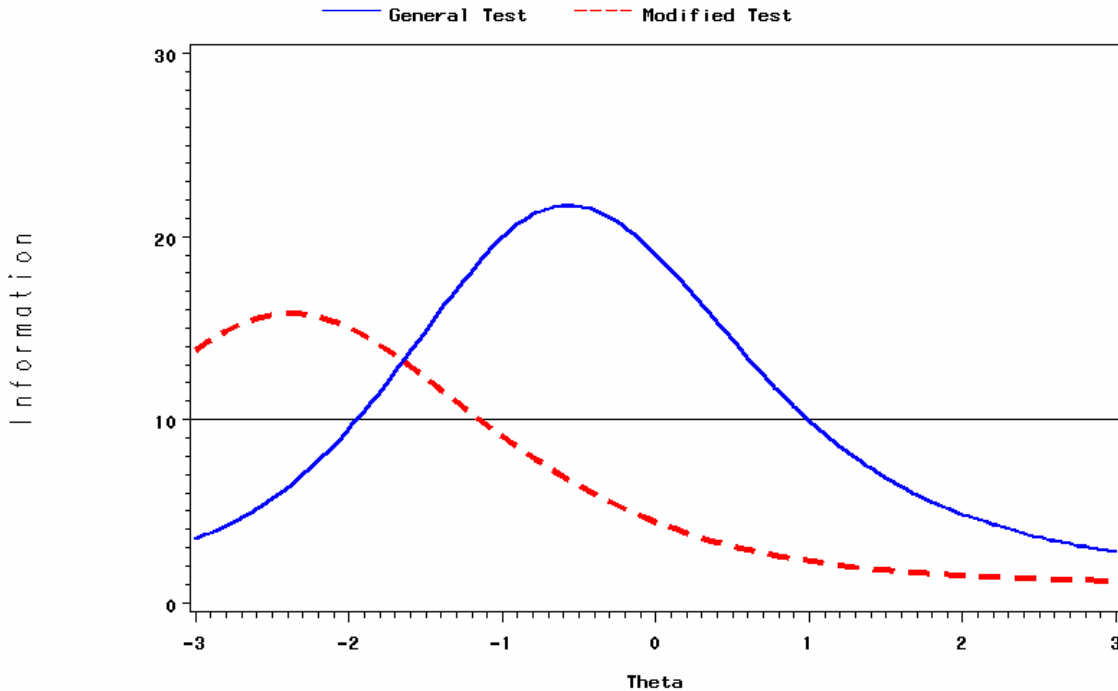


Figure 5. Test Information Function for General and Modified Tests: Grade Six Mathematics

The analyses for the grade three mathematics tests and the grade six reading tests are consistent with the grade three reading and grade six mathematics results shown in Figures 4 and 5, respectively. In all cases, the modified test appears to be less difficult than the general test, providing more reliable information at a lower point along the proficiency continuum. At both grade levels, there is much more separation between the TIF for the mathematics tests than for the reading tests.

Discussion

The primary purpose of the study was to investigate whether the methods and analyses used here provide useful information to support the design, interpretation, and use of modified tests. A secondary purpose of the study was to provide information to the state to assist them in interpreting whether their test development approach resulted in modified tests that met their needs as well as met federal guidelines for modified tests. With respect to the second purpose, the results of this study do provide evidence that the modified tests developed by the state in reading and mathematics are less difficult than the corresponding general tests, and they provide sufficiently reliable information at points well below the grade level academic achievement standard.

The results of these quantitative analyses suggest that the general and modified tests in reading are much more similar to each other than the corresponding mathematics tests. A qualitative analysis of the content of the two tests would be necessary to determine the degree of similarity. However, given the general manner in which reading and mathematics content standards tend to change across grade levels, the result of finding significantly more overlap between the general and modified tests in reading was not unexpected. Attempting to maintain fidelity to the grade

level content standards, including the reading level of passages included on the assessment, it appears that there would be less room to extend the proficiency continuum on a reading test than on a mathematics test. This may be particularly true in the case of exclusively multiple-choice tests in which there may be a tendency to measure lower level skills. (Although it certainly is possible to measure higher-level thinking skills with multiple-choice items, and that question has not been examined this study.)

The level of overlap between the general and modified reading tests does raise questions about the purpose of the modified assessments and the United State Department of Education requirement that the modified test must be a different instrument than the general test, not simply a different achievement standard. The regulations and guidance regarding the modified tests reflect the dual concerns of access and level of proficiency that are meant to be addressed by the modified test. With regard to level of proficiency, quantitative analyses such as those presented in this report may clearly demonstrate that the general test provides sufficiently reliable information at the level of proficiency appropriate for the target population(s). Whether the general test provides students in that population with sufficient access to demonstrate their level of proficiency is a separate issue. The access question cannot be addressed through a study in which the two tests are administered to the general population.

With regard to whether the methods described in this report provide useful information to those designing and using modified tests, the answer to that question will be left to those designing and using modified percent tests. It is clear that the analyses conducted in this study do provide information that addresses key questions regarding the difficulty of the modified tests. However, the effort on the part of the state to design and implement the special study used to gather the data used in the study was quite substantial. This issue, of course, is part of the larger issue facing states of determining the return on investment of developing a series of modified tests that meet appropriate technical quality standards and are compliant with federal requirements.

In cases where a state does decide to develop modified tests, however, it seems certain that there will be a demand to establish a link between the results of the modified test and the general test in terms of the content mastered and of the distance from the grade level academic achievement standard. Such efforts to establish links between tests are already commonplace in the case of general tests across grade levels and are not uncommon in the case of alternate assessments and general tests within a grade level. The need for and appropriateness of such a link is almost certainly stronger in the case of the modified percent tests where

1. the students in the target population are expected to achieve the same grade level academic achievement standard at some point in time
2. the two tests are aligned to the same grade level content standards

If the modified tests are to meet the goal of providing useful information to support instruction, and not merely serve an accountability purpose, it is likely that states will have to engage in some type of effort, such as those described in this study, to establish the relationship between their general and modified tests.

References

- “34 CFR Parts 200 and 300 Title 1 – Improving the Academic Achievement of the Disadvantaged; Individuals With Disabilities Education Act (IDEA) – Assistance to State for the Education of Children With Disabilities,” 72 Federal Register 67 (April 9, 2007) pp. 17748–17781.
- Department of Education (2006). Test Administrator’s Manual: Special Study for Reading and Mathematics Grades 4–9 (Form A).
- Department of Education (2006). Test Administrator’s Manual: Special Study for Reading and Mathematics Grades 4–9 (Form B).
- Kolen, M.J. & Brennan, R.L. (2004). Test equating, scaling, and linking: Methods and practices (second edition). New York: Springer.
- Livingston, S.A. & Lewis, C (1995). Estimating the consistency and accuracy of classifications bases on test scores. *Journal of Educational Measurement*, 32, 179–197.
- Thissen, D. Chen, W., & Bock, D. (2003) MULTILOG for Windows, version 7.0.2327.3. Lincolnwood, IL: Scientific Software International.
- United States Department of Education (2007). Modified Academic Achievement Standards: Non-regulatory Guidance. Downloaded on March 1, 2009 from <http://www.ed.gov/policy/special/guid/nclb/twopercent.doc>

Section 3: Literature Reviews Related to Comparability of Various Types of Test Variations

Chapter 7: Comparability of Paper-based and Computer-based Tests: A Review of the Methodology

Susan M. Lottridge
W. Alan Nicewander
E. Matthew Schulz
Howard C. Mitzel
Pacific Metrics

Introduction

The number of computer-delivered tests is increasing in schools as a result of accountability legislation and the perceived advantages of computer-delivered tests. In 2003, an *Education Week* survey reported that 13 states were piloting or using computer-based tests (Olson, 2003). In 2006, 21 states and the District of Columbia offered computerized tests in some form (Educational Research Center, 2006). Perceived advantages of computer-based testing include more flexible scheduling, the ability to tailor tests more specifically to student needs, and more rapid scoring and reporting.

Comparability issues arise when a new mode of test delivery replaces or is used alongside an established mode, such as paper-and-pencil testing. Because paper-and-pencil tests (PBTs for paper-based tests or testing) have precedence of use, they represent the gold standard to which computer-based tests or testing (CBTs) is compared. As PBTs are replaced in the K–12 testing arena by CBTs, the CBT scores need to be tied back in some way to the original PBT scale in order to maintain continuity. Also, the development of a CBT-only test for which there is no PBT precedent within a given state or school district may nevertheless require consideration of its relationship to a paper-based counterpart in another district or state.

This report is intended to be useful to educational policymakers and researchers concerned with the comparability of computer-delivered and paper-and-pencil tests. The purpose of this report is to review the *methods* used in the literature to investigate comparability rather than the results of such studies. Excellent reviews of comparability study results are already available in the literature (Paek, 2005; Bennett, 2003; Gaskill, 2006).

Comparability

Comparability can be examined on two levels. First, comparability can be examined in terms of score equivalence. In other words, one can investigate whether the two modes (PBT or CBT) produce similar score distributions, such as similar means and standard deviations. Second, comparability can be examined in terms of construct equivalence. Here, the term “construct” refers to an unobservable property of persons that is being measured using a test. An example of a construct is math proficiency; a person’s proficiency at math cannot be directly observed but a

person can be thought to have a level of this skill. Constructs can be narrowly defined (e.g., keyboarding proficiency) or broadly defined (e.g., reading comprehension). Because construct comparability involves determining whether the tests in two modes are measuring the same construct to the same degree, it is a complex and difficult task.

Why examine score equivalence? If the distribution of the scores is the same across the two modes, then two important inferences logically follow. First, the two modes can be said to be functionally comparable in terms of overall scoring of a sample. Second, the constructs measured can be *reasonably* assumed to be the same; there is no counter evidence that differing constructs are involved when the score distributions are comparable. In fact, meta-analyses on comparability generally indicate that mode differences, if they exist, are small for untimed tests (Bergstrom, 1992; Mead & Drasgow, 1993; Kim, 1999). However, comparing score distributions alone may be misleading. Tests from two modes might produce the same overall score distributions, but the scores for any one examinee may differ substantially between tests. In other words, the scores from two modes might produce a different rank ordering of examinees. Such an occurrence is an indication that the modes are measuring different constructs. Using methods to examine the validity of the scores, one can compare the extent to which scores are measuring the same construct.

Score equivalence and construct equivalence are especially important for accountability. It would be difficult or impossible to monitor trends in student learning for purposes of accountability if scores from two alternative modes of assessment had different meaning, or if students in a particular demographic category tended to earn different scores on one mode than on the other. The meaning of change in a summary statistic, such as the percent of students scoring at or above a given achievement level, could be confounded by whether or not the test was administered in a paper-based or computer-based mode.

Various guidelines have been published for examining comparability between CBT and PBT. Standards from the American Psychological Association (1986) and the International Test Commission (2005) emphasize the need for similar score distributions, reliabilities, ranking of examinees, and correlations with external criteria.

“Scores from conventional and computer administrations may be considered equivalent when (a) rank orders of scores of individuals tested in alternative modes closely approximate each other, and (b) the means, dispersions, and shapes of the score distributions are approximately the same, or have been made approximately the same by rescaling the scores from the computer mode” (APA, 1986, p. 18).

“Provide clear documented evidence of the equivalence between the CBT/Internet test and non-computer versions (if the CBT/Internet version is a parallel form). Specifically, to show that the two versions: have comparable reliabilities; correlate with each other at the expected level from the reliabilities; correlate comparably with other tests and external criteria; and, produce comparable means and standard deviations or have been appropriately calibrated to render comparable scores” (ITC, 2005, p. 21).

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) provide less specific guidance on statistics to be used, and instead instruct the test developer to conduct studies relevant to the use and interpretation of the test scores.

Standard 4.10: “A clear rationale and supporting evidence should be provided for any claim that scores earned on 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 the procedures for establishing score comparability have been sufficiently satisfied. The specific rationale and the evidence required will depend in part on the intended uses for which score equivalence is claimed” (p. 57).

Together, these guidelines call for comparisons of score distributions across testing modalities, comparisons of relationships of scores across modes, and also comparisons of relationships with other criterion measures. While the guidelines outline methods for investigating comparability, they do not specify criteria by which to determine whether comparability has been achieved. Rather, the investigator must use judgment when interpreting results. The remainder of this report will describe methodologies for making such comparisons, and will outline the work other researchers have conducted to determine comparability.

Use of the Principles of Construct Validation in the Design of Comparability Studies

The area of psychology from which one may draw methods for addressing comparability is *construct validation* (Cronbach & Meehl, 1955). Construct validation is the process by which evidence is gathered to assess the degree to which a test is measuring the construct it is intended to measure and the degree to which test scores are supporting intended inferences. The validation procedure consists of the accumulation of positive evidence *and* the inability to uncover negative evidence. At some point, enough evidence is acquired that enables a judgment regarding the validity of the test scores.

In terms of the philosophy of science, construct validation falls under the general topic of *theory testing*. Construct validation uses theory and/or logic to develop hypotheses that should be true if a test is valid for the construct that it claims to measure. These hypotheses form predictions that are tested using experimental data. In the case of comparability studies, the construct validation paradigm is simplified a bit since the nature of the construct being measured by two tests (or two testing modes) does not have to be identified. Rather, the question is whether the constructs assessed by the two tests are the same. In the determination of the degree of consonance between the constructs measured by CBT and PBT, the following is a partial list of the logical deductions that can be derived and then tested with experimental data:

- If the construct measured by the two modes is the same, then their content and content specifications should be the same (evidence from content validation).
- If identical constructs underlie these two testing modes, then they should have the same factor structure (psychometric-statistical evidence).

- If CBT and PBT measure the same construct—and are to be used interchangeably—they should have the same measurement precision.
- If the constructs being measured are the same, then these two testing modes should yield scores that differ only because of difficulty. This difference can be then removed using equating (psychometric-statistical evidence).
- If the underlying constructs are identical for two testing modes, then their intercorrelation, corrected for unreliability, should be unity—within sampling error. This evidence would establish that the tests are *congeneric* (i.e., have perfectly correlated true scores); it does not confirm that they are parallel forms (i.e., yield identical true scores and the same error variance) (psychometric-statistical evidence).
- If the CBT and PBT measure the same construct, then they should have the same predictive validity coefficient (evidence from predictive validation), or, similarly, tests measuring the same construct should have *equal correlations* with external measures (concurrent validity).

Of course, the support (or lack thereof) for these hypotheses involves human judgment, probabilistic reasoning, and the strengths and limitations of study design. Thus the interpretation of evidence is crucial to a decision regarding comparability. The key idea here is that comparability can and should be addressed experimentally using a hypothesis-testing approach in a construct validation framework.

Comparability Research Methodology Review

The purpose of this report is to provide an overview of the methods used in studies investigating comparability between PBT and CBT. The list of studies examined appears in Table 1. The focus of the comparability studies centered on score and/or construct differences between PBTs and CBTs, and all studies conformed to one or more of the six construct validation hypotheses listed above. No studies examined all six hypotheses, but many examined at least two. The most comprehensive studies considered the key aspects of validity and reliability, taking into consideration both score distributions and the relationships of CBT and PBT scores to the construct being measured.

Table 1. List of Studies Included in Methodological Review

Study
Choi and Tinkler (2002)
Eignor (1993)
Fitzpatrick and Triscari (2005)
Higgins, Russell, and Hoffman (2005)
Hollenbeck, Tindal, Stieber, and Harniss (1999)
Johnson and Green (2006)
Olson, Maynes, Slawson, and Ho (1989)
Poggio, Glasnapp, Yang, and Poggio (2005)
Pommerich (2004)
Pomplun and Custer (2005)
Pomplun, Frey, Becker, and Hughes (2000)
Russell and Haney (1997)
Russell and Plati (2001)
Russell and Tao (2004)
Russell (1999)
Sandine, Horkay, Bennett, Allen, Braswell, Kaplan, and Oranje (2005)
Schwarz, Rich, and Podrabsky (2003)
Way and Fitzpatrick (2006); Way, Davis, and Fitzpatrick (2006); Keng, McClarty, and Davis (2006)
Zhang and Lau (2006)

This section outlines the various approaches used in the literature that can serve as examples of potential comparability designs. Each design description will be accompanied by potential threats to the validity of the results. Only studies using K–12 students, having sufficient detail to understand the methodology employed and incorporating reasonably strong design and implementation are presented.

General Methodological Issues

Study design issues generally span four major dimensions: sampling, instrumentation, administration, and scoring. The comparability literature (including but not limited to studies listed in Table 1) reviewed often did not provide sufficient detail or exhibited problems related to one or more of these dimensions. Most design issues arose due to real-world constraints in implementation; however, some were also due to inadequate descriptions of the implementation. The dimensions listed below are organized by the internal and external validity framework provided in the work of Campbell and Stanley (1963). External validity refers to the extent to which the study results can be generalized to the population. Internal validity refers to the extent to which the study design sufficiently controls variables to test the hypotheses.

General factors influencing external validity (generalizability of results). Each of the four dimensions of sampling, instrumentation, administration, and scoring are presented as they relate to the generalizability of the study results.

- **Sampling.** The selection and solicitation of the sample should be described. This description should describe how the original sample (e.g., schools and/or students) was selected, how many of the sample participated in the study, and the extent to which the study participants were similar to the original sample. In particular, with voluntary samples it is important that attention be paid to the type of schools that volunteer since

participating schools may differ from non-participating schools. For instance, a volunteer school may have more access to computers or may consist of more computer-savvy students.

- Instrumentation. The instruments used in the study should be comparable to those that would be used operationally. In the case of K–12 testing, this issue is less of a concern since it is likely that operational forms would be used in any comparability study. Studies that use researcher-assembled instruments need to provide details on test specifications, test construction, reliability, and, if possible, validity.
- Administration. The conditions under which the tests are administered should be similar to those normally practiced. These conditions include having the same motivation level (e.g., low-stakes versus high-stakes testing), similar computers and computer access, similar bandwidth issues, similar student practice testing and training, and similar proctor training.
- Scoring. The method of scoring should reflect the methods used operationally. For instance, if examinees in the population are assigned into performance levels based on test scores, then the sample examinees should be similarly assigned.

General factors influencing internal validity (interpretation of results). Issues relating to the four dimensions and whether their influence on test scores could be attributed to mode and not other influences are listed below.

- Sampling. The key areas of sampling with respect to internal validity are sample size and assignment to groups.
 - Sample size. It is important to collect a large enough sample in order to reflect the breadth of the population, compute stable statistics, and use the sample for future analyses (such as equating studies or IRT item calibrations). For instance, calibration of IRT item parameters is often an element of comparability studies. For the three-parameter IRT model, a minimum of 1,200 examinees is recommended for accurate calibration. For the Rasch model, a minimum of 500 observations is recommended. Equipercentile equating can require up to 5,000 examinees.
 - Group assignment. A key element of an experimental study is random assignment to help ensure group equivalence. Random assignment can occur at the group level (school, classroom) or student level. The equivalence of the groups should be demonstrated using background variables, such as demographic or testing data. Group assignment into counterbalanced conditions should also be random and not left up to the study participants. If random assignment is not possible, then efforts should be made to ensure the groups are similar (e.g., matching methods, use of covariates).
- Instrumentation. Instruments need a sufficient number of items to represent the construct of interest and to be reliable measures. This issue is particularly salient when using constructed-response items, because generally a small number of items are used and examinees often respond differently to similar-looking items. For instance, a student may perform and score differently on two items designed to be equivalent. As an example, two story problems might use the same underlying math but apply it to different scenarios. A student might know the underlying math, but be able to apply the

computation to only one scenario. This examinee-by-task interaction is important to consider if different items are administered across mode since the interaction, rather than mode, may be the source of score differences.

- Administration. Tests of the two populations should occur at the same time whenever possible to rule out extraneous variables (e.g., maturation). Studies that use a volunteer sample and compare the sample results to already existing test data risk extraneous factors influencing test scores.
- Scoring. The test scores need to be placed on the same scale prior to any comparison. This linking process should be reported. In addition, if IRT item parameters are used in CBT or computer adaptive tests (CAT), then the origin of those parameters should be described. For instance, the item parameters may be from scores gathered from a PBT or items may be recalibrated on the CBT examinees. In addition, inter-rater reliability of constructed-response measures should be examined and presented. It is also important to identify which type of score is being used in any comparison (e.g., raw score, standard score, IRT-estimated theta, IRT-estimated true score).

Types of Designs

Two basic designs are dominant in the comparability literature: the within-subjects design and the between-subjects design. In the within-subjects design, examinees take both a CBT and a PBT. In the between-subjects design, examinees are divided into two groups, with one group taking the CBT and the other taking the PBT. The studies reviewed were fairly evenly divided between the two designs. Variations exist within each design, and these are discussed in the relevant section.

Within-subjects designs

In the within-subjects design, a single group of students is administered a PBT and a CBT. Counterbalancing is used to moderate any effects that might arise from test order (such as fatigue, practice, or motivation). With respect to presentation order, counterbalancing is a process by which examinees are divided into two groups: examinees in one group take the CBT first and the PBT second, and examinees in the other group take the tests in the opposite order. In addition, counterbalancing can be used to moderate other testing effects. For example, in the case of comparability, a within-subjects design often requires that examinees take one set of items on the PBT (e.g., Form A) and another set of items on the CBT (e.g., Form B). Thus a researcher can counterbalance for both order and test form. Order can also be used as a blocking variable to reduce experimental error and to check whether one order leads to better performance than another. The within-subjects design is somewhat limited in its application because of the administrative problems caused by testing the same students twice. However, this design is the richest source of information concerning comparability. The only threat to the validity of this design is that taking two tests may not generalize to a population where only one test is administered.

There are three general variants of the within-subjects design. The first variant administers the same items (or draws items from the same testing pool) for the CBT and PBT. The second uses different items (i.e., test forms) for the CBT and PBT. The third variant does not administer a test

twice to examinees. Rather, constructed responses are scored by raters in different formats (i.e., typed and handwritten). The studies representing each variation appear in Tables 2–4.

The first variant of within-subjects studies used the same set of items for the PBT and CBT. Table 2 summarizes details of these three studies. Key elements of the design of the studies are outlined in the bulleted text below.

- Pomplun and Custer (2005) administered the same items via a computer-based and paper-based test, counterbalancing the order. They administered the Initial Skills Analysis Test, which had three subtests (basic skills, comprehension, and language) to children in kindergarten to grade three. Schools were randomly assigned to the counterbalanced condition. The purpose of the study was to conduct a confirmatory factor analysis on the raw subtest scores and mode to determine levels of factorial equivalence (factor structure equivalence, same factor loadings, same factor loadings and errors). Separate factor analyses were conducted at each of the four grades. A potential threat to the validity of this study is that examinees were administered the same items twice, increasing the potential for a practice or memory effect. However, the random assignment and counterbalancing presumably minimized this threat.
- Eignor (1993) studied the performance of a group of high school juniors, high school seniors, and college freshmen in motivated conditions on computerized adaptive and paper-based versions of the SAT. The purpose of this study was to examine the degree of differences in equating scores from the two methods. The study design called for randomly counterbalancing the modes, but the instructions for counterbalancing were not closely followed. The item pool for the adaptive test shared items with the paper-based test, and presumably the adaptive test used item parameters calibrated from a previous paper-based administration, although this was not specified. The major threats to this design are that examinees may encounter the same items twice and thus increase the potential for a practice or memory effect, the problems with counterbalancing, and the potential confounds of using PBT-derived item parameters for the computer adaptive test.
- Olson, Maynes, Slawson, and Ho (1989) examined mode differences for CAT, CBT, and PBT in math for third and sixth graders. The researchers counterbalanced the mode for the adaptive and computer-based test and counterbalanced the mode for the adaptive and paper-based test. Students were randomly assigned into groups. The item pool for the adaptive test shared items with the paper-based and computer-based tests, and presumably the adaptive test used item parameters calibrated from a previous paper-based administration, although this was not specified in the report. The major threats to the validity of this design are similar to those listed for Eignor (1993).

Table 2. Studies Using a Within-subjects Design (Counterbalancing for Order) in Which the Same Items Were Administered across Modes

Study	Study details
Pomplun and Custer (2005)	<i>Instruments:</i> Initial Skills Analysis Test (basic skills, comprehension, language), multiple choice w/reading passages <i>Sample:</i> kindergarten (n=537), 1 st grade (n=457), 2 nd grade (n=498), 3 rd grade (n=467) <i>Other measures:</i> parental income, indicated by whether examinee received free lunch <i>Purpose:</i> score and construct equivalence
Eignor (1993)	<i>Instruments:</i> SAT-Verbal and Quantitative, multiple choice <i>Sample:</i> HS/college students (n=506) <i>Other measures:</i> none <i>Details:</i> CBT was an adaptive test <i>Purpose:</i> score and construct equivalence
Olson, Maynes, Slawson, and Ho (1989)	<i>Instruments:</i> California Assessment Program math, multiple choice <i>Sample:</i> 3 rd grade (n=350), 6 th grade (n=225) <i>Other measures:</i> testing time <i>Details:</i> two CBTs: one adaptive, one linear <i>Purpose:</i> score and construct equivalence

In the second variation of within-subjects studies, examinees are administered different items (i.e., different forms) across modes and these forms are counterbalanced as well. In one particularly interesting design, all examinees are administered the same test in the same mode, and then are administered both a computer-based and paper-based test using different items. The initial test is used for common-item equating, and permits equating studies to determine mode effects. Table 3 summarizes details of these studies, and key elements of the design are described in the bulleted text below.

- In Johnson and Green (2006), participants were randomly assigned into four groups in which counterbalancing occurred for testing order and form. Forms were not equated, and comparisons of scores were made within forms. The forms consisted of eight mathematics items, and students' processes in answering items were also captured by analyzing student worksheets and interviews about their process. Participants were 10–11-year-old school children. Threats to the validity of this design were in the instrumentation (the instruments used were quite short and researcher-created) and in the small sample size (about 50 examinees per form).
- In Pomplun, Frey, Becker, and Hughes (2000), participants were randomly assigned into four groups where counterbalancing occurred for order and form. Six schools were randomly assigned form and mode; schools were asked to randomly assign students to test orders. The researchers administered the Nelson Denny Reading test at the high school and college levels. The focus of the study was the effect of mode on reading rate. However, different stopping criteria were used for measuring the reading rate. On CBT, the student clicked on the last word read and on the PBT, the student marked the last line read (and the middle word for that line was used). The threats to the validity of this design were the use of different stopping criteria, which could influence the calculation of test scores, and the small sample size.

- Poggio, Glasnapp, Yang, and Poggio (2005) conducted a comparability study on four already-equated forms from a state seventh-grade mathematics assessment in Kansas. Forms were randomly assigned but the counterbalancing for mode order was not random; rather, volunteer schools chose which mode to administer first. The majority of schools chose to administer the CBT first. Because forms in either mode were randomly assigned, a subset of students took the same form in both modes. The major threat to the validity of this design is the non-randomized counterbalancing procedure because this could potentially results in non-equivalent groups. In addition, the sample size was small given the analyses conducted on these data.
- In Choi and Tinkler’s (2002) study, participants in two randomly assigned groups took three tests. Participants first took a set of common items via computer and then took either a computer-based test and then a paper-based test or a paper-based test and then a computer-based test. Random assignment into testing order was conducted at the classroom level. Two test forms were used. Both test mode and form were counterbalanced, although order differences were not presented. The instruments were state math and reading assessments and were administered to 3rd and 10th graders. The threat to the validity of this design is the unknown influence of the first CBT.

Table 3. Studies Using a Within-subjects Design (with Counterbalancing for Order and Form) in Which Different Items Were Administered Across Modes

Study	Study details
Johnson and Green (2006)	<i>Instrument:</i> mathematics test items aligned to the British National Curriculum, multiple choice <i>Sample:</i> British 10–11 year olds (n=104) <i>Other measures:</i> observations of examinee test taking, analysis of examinee test-taking strategies, interviews about mode preference <i>Purpose:</i> score and construct equivalence
Pomplun, Frey, Becker, and Hughes (2000)	<i>Instrument:</i> Nelson Denny Reading Test (vocabulary, reading comprehension, and total score), multiple choice <i>Sample:</i> high school, 2 and 4 yr college (n=185) <i>Other measures:</i> none <i>Purpose:</i> score and construct equivalence
Poggio, Glasnapp, Yang, and Poggio (2005)	<i>Instrument:</i> Kansas Computerized Assessment in mathematics multiple choice <i>Sample:</i> 7 th grade (n=646) <i>Other measures:</i> gender, socio-economic status (lunch support), and academic placement (general education, gifted, special education) <i>Purpose:</i> score and construct equivalence
Choi and Tinkler (2002)	<i>Instruments:</i> math and reading tests were portions of operational tests, multiple choice with stimulus material <i>Sample:</i> 3 rd grade (n~800), 10 th grade (n~800) <i>Other measures:</i> additional CBT administered for equating, analysis of reading item characteristics <i>Purpose:</i> score and construct equivalence

In addition to the design above, a single-group within-subjects design has been used in studies whose focus was to determine whether *typed* constructed-response essays would be graded differently by scorers than *handwritten* essays. This approach could also be used to compare digitized handwritten essays and typed essays. In these studies, handwritten essays were transcribed into computerized text. Essays in both modes were then scored by human raters using

a scoring rubric, and mean comparisons of the scores were calculated. These studies have been used in operational testing of elementary school students (Russell & Tao, 2004), middle school students (Hollenbeck, Tindal, Stieber, & Harniss, 1999; Russell & Tao, 2004), and high school students (Russell and Tao, 2004) in English Language Arts. In one study (Russell & Tao, 2004), raters also marked writing errors (e.g., spelling, punctuation, capitalization, awkward transitions, confusing passage) and compared the proportion of errors across mode. These studies involved no additional participation on the part of the student, since the student essays were re-scored. In Russell and Tao (2004), essays were transcribed twice into text as a single-spaced essay and as a double-spaced essay. The only threat to the internal validity of these types of studies is that no comparative group existed to separate rater reliability and bias from mode effects. Interestingly, none of these studies calculated rater agreement across mode. These studies also used small samples. Table 4 summarizes details of these studies.

Table 4. Special Studies Using a Within-subjects Design (Same Responses, Different Raters)

Study	Study details
Hollenbeck, Tindal, Stieber, and Harniss (1999)	<i>Instrument:</i> single Oregon ELA writing essay item <i>Sample:</i> middle school students (n=80) <i>Other measures:</i> none <i>Purpose:</i> score equivalence
Russell and Tao (2004)	<i>Instrument:</i> Massachusetts Comprehensive Assessment System (MCAS) Language Arts Test, essay <i>Sample:</i> 4 th grade (n=52), 8 th grade (n=60), 10 th grade (n=60) <i>Other measures:</i> analysis of essay features <i>Purpose:</i> score equivalence

Between-subjects Designs

In the between-subjects design, participants are divided into two or more groups and each group is administered the same items via a computer-based test or a paper-based test. There are no serious threats to the internal validity of these designs if the groups are assumed to be randomly equivalent.

Three variations of the between-subjects design are found in the studies. The first variation uses random or pseudo-random methods to assign groups. The second variation uses matching methods to form groups. Finally, the third variation uses additional test data to perform post hoc covariance analysis or for equating. Tables 5–7 provide information on these studies.

In the first variation of between-subjects studies, examinees are divided into groups using random or pseudo-random methods (e.g., stratified random sampling). There are no general threats to the validity of this type of design. Table 5 summarizes details of these studies, and the studies are described in the bulleted list below.

- A series of studies on the Texas statewide graduation exams (TAKS) in reading, mathematics, social studies and science (Way, Davis, & Fitzpatrick, 2006; Way & Fitzpatrick, 2006; Keng, McClarty, & Davis, 2006) used the between-subjects approach. Eleventh graders who had failed the graduation exam were offered an additional testing opportunity and were randomly assigned to take a PBT or CBT. The major threats to the

validity of this design are the restricted range of scores due to using a sample that had scored relatively low on the original administration and problems of generalizing of the retest results to another population.

- Fitzpatrick and Triscari (2005) randomly assigned subjects to two groups (CBT or PBT) in order to examine the comparability of the operational Virginia end-of-course Algebra I, earth science, and English language arts tests. Scores from a previously-taken PBT were used to examine the similarity of the two groups, and the groups were found to be too different to compare scores. As a result the study used a non-equivalent groups common-item linking design to examine equating parameters.
- Stratified random samples are also used to divide examinees into groups. Russell and Plati (2001) used English grades from a previous year to assign participants to a CBT or a PBT group in eighth- and tenth-grade samples. In this study, responses to a single extended response item were examined to determine whether writing an essay on computer differed from writing an essay on paper. All handwritten essays were transcribed and subsequently scored by two raters blind to the testing mode. The raters' scores were then summed. The authors indicated that the motivation levels may have differed between the two modes due to the tests being administered in a low motivation condition and the use of different proctors. Also, a single high-performing district was used. In another study with a similar design, Russell (1999) used the grade seven Stanford Achievement Test 9 (SAT-9) normal curve equivalents (NCE) to assign examinees for an eighth-grade sample in math, language arts, and science. This study focused on the influence of mode on constructed-response items. The major threats to the validity of these designs were the limited sample, the instrumentation consisting of few items, and the potential differences in motivation.
- Various computer-based test characteristics were studied (e.g., enabling examinees to scroll line-by-line through long text passages or to scroll page-by-page). Researchers (Higgins, Russell, & Hoffman, 2005; Pommerich, 2004) divided examinees randomly into various computer-based test conditions and one paper condition. Pommerich (2004) conducted two studies in an effort to examine the impact of a CBT interface. One examined mode differences between PBT and CBT scores. The second study used results learned in the first study, and examined the influence of various item presentation modes in CBTs. In this study, two forms of automated scrolling were used in English. In one variation, the relevant portion of the passage automatically scrolled if it did not appear in the passage window. In the second variation, the relevant portion of the passage automatically scrolled to the top of the page for each item. The influence of line-by-line and page-by-page scrolling was examined in reading and science reasoning. The other examined the influence of two methods for automated scrolling through stimulus passages. Higgins, Russell, and Hoffman (2005) examined the impact of item presentation on paper, on computer with line-by-line scrolling, and on computer with page-by-page scrolling. The authors used a fourth-grade reading test with long passages in their study. In addition, the authors examined results in relation to external measures such as a computer fluidity test.

Table 5. Studies Using a Between-subjects Design with Random or Pseudo-random Assignment

Study	Study details
Way and Fitzpatrick (2006); Way, Davis, and Fitzpatrick (2006); Keng, McClarty, and Davis (2006)	<p><i>Instrument:</i> Texas Assessment of Knowledge and Skills (TAKS) math, ELA, social studies, science; multiple choice.</p> <p><i>Sample:</i> 11th-grade retest. Math (n=2156), Science (n=2201), Social Studies (n=743), ELA (n=1368)</p> <p><i>Other measures:</i> survey administered regarding computer skills, use, and preference for testing</p> <p><i>Purpose:</i> score and construct equivalence</p>
Fitzpatrick and Triscari (2005)	<p><i>Instrument:</i> Virginia state end-of-course tests in algebra, earth science and ELA, multiple choice</p> <p><i>Sample:</i> high school students (n=2205)</p> <p><i>Other measures:</i> additional PBT administered prior to study</p> <p><i>Purpose:</i> score equivalence</p>
Russell and Plati (2001)	<p><i>Instrument:</i> Single Massachusetts Comprehensive Assessment System (MCAS) writing prompt</p> <p><i>Sample:</i> 8th grade (n=144) and 10th grade (n=145)</p> <p><i>Other measures:</i> keyboarding test, survey of prior computer use, midterm grade used as covariate</p> <p><i>Purpose:</i> score and construct equivalence</p>
Russell (1999)	<p><i>Instrument:</i> items from NAEP and Massachusetts Comprehensive Assessment System (MCAS); math, science, and ELA, all items were constructed-response</p> <p><i>Sample:</i> 8th grade (n=229)</p> <p><i>Other measures:</i> keyboarding test, computer use survey, and SAT-9 NCE scores used to stratify sample</p> <p><i>Purpose:</i> score and construct equivalence</p>
Higgins, Russell, and Hoffman (2005)	<p><i>Instrument:</i> Reading Comprehension Test, multiple choice with passages. Released items from NAEP, Progress in International Reading Literacy Study (PIRLS), and NH state assessments</p> <p><i>Sample:</i> 4th grade (n=219)</p> <p><i>Other measures:</i> computer fluidity test, computer literacy test, computer use survey</p> <p><i>Purpose:</i> score and construct equivalence</p>
Pommerich (2004)	<p><i>Instrument:</i> English, reading, and science reasoning content areas; test original unknown; most items had stimulus materials; all items were multiple choice</p> <p><i>Sample:</i> 11th and 12th grade (Study 1: n= 5612, Study 2: n= 9473)</p> <p><i>Other measures:</i> none</p> <p><i>Purpose:</i> score and construct equivalence</p>

The second variation of the between-subjects design used groups that took the tests at different times. In these designs, the CBT group was a volunteer sample, and the PBT group was assembled from prior testing. In some studies this group is the entire census tested group or a subset of the entire group. There are a number of threats to the validity of this type of design. First, if the administration of the CBT is not comparable to that of the PBT, then the data will likely not be comparable. Second, maturation or additional learning may have occurred if the CBT testing occurs much later than the PBT testing. Third, the sample of students who agree to participate in the CBT testing may not have the same characteristics as the students participating in the PBT testing. Table 6 summarizes details of these studies.

- Schwarz, Rich, and Podrabsky (2003) examined the mode comparability of the analytical reasoning and quantitative reasoning subscales of InView Test. The CBT sample data were collected following the much larger PBT standardization sample. The CBT sample was selected based upon region city size and lunch status. The authors did not provide information on the method for selecting participants. A threat to this design is that the researchers used the entire norming sample rather than a subsample selected to be equivalent to the examinees participating in the study.
- In Way, Davis, and Fitzpatrick's (2006) study of the TAKS, the small number of study volunteers did not allow for enough participants to randomly assign subjects to CBT and PBT conditions. The investigators administered a CBT to all participants and used a set of matched samples from all students taking the operational PBT tests. Eighth graders were the subjects of study, and the seventh-grade TAKS reading and math test scores were used as the matching variables.
- In a National Assessment of Educational Progress (NAEP) study, Sandine, Horkay, Bennett, Allen, Braswell, Kaplan, and Oranje (2005) used this design to study mode effects of writing on a computer and on paper. The study data were collected in the months immediately following the operational NAEP administration. A random sample of schools participating in NAEP was chosen for the study, and students with existing NAEP reading or writing data were selected to be in the sample. One half of the sample had taken a NAEP writing item, and one half had taken the NAEP reading (and not the writing item) item. The latter sample was collected to examine whether taking a NAEP writing item previously influenced performance; it did not. Study participants answered two constructed-response prompts on the computer. The CBT scores were compared to their PBT scores and to a comparable sample of respondents answering the same essay during the operational NAEP assessment. Potential threats to this design are that the sample did not reflect the nationally representative NAEP sample, and that only two items were used in the instrument.

Table 6. Studies Using a Between-subjects Design with Volunteer CBT Sample and Pre-existing Sample

Study	Details
Schwarz, Rich, and Podrabsky (2003)	<i>Instrument:</i> InView analytical reasoning and quantitative reasoning subtests, selected response (i.e., multiple choice, matching) <i>Sample:</i> analytical reasoning: grades 4 and 5 (n=2295), 6 and 7 (n=1839), 8 and 9 (n=1455); quantitative reasoning: grades 4 and 5 (n=2103), 6 and 7 (n=1623), 8 and 9 (n=1260) <i>Other measures:</i> none <i>Purpose:</i> score and construct equivalence
Way and Fitzpatrick (2006); Way, Davis, and Fitzpatrick (2006); Keng, McClarty, and Davis (2006)	<i>Instrument:</i> TAKS reading, math, social studies, multiple choice and constructed response <i>Sample:</i> grade 8. math (n=1273), social studies (n=1449), reading (n=1840) <i>Other measures:</i> none <i>Purpose:</i> score equivalence
Sandine, Horkay, Bennett, Allen, Braswell, Kaplan, and Oranje (2005)	<i>Instrument:</i> two essays used in the main NAEP writing assessment <i>Sample:</i> 8 th grade (n=1255) <i>Other measures:</i> background questions that asked for demographic information, computer experience, and writing instruction; also, an online computer skills measure <i>Purpose:</i> score and construct equivalence

In the third and final variation of the between-subjects design, examinees take additional tests aside from the tests being examined for comparability. These additional tests are used to equate the item parameters and test scores, or the test scores are used as covariates. The additional tests are often administered prior to the studied tests. Collecting the additional test data enables the measurement of group differences, the use of regression methods to control for differences, and the information for equating the PBT and CBT. The threat to these designs is the potential influence of the initial PBT which is often administered first and is not counterbalanced during the experiment. Taking the initial PBT may produce practice or fatigue effects that are not controlled through counterbalancing. Table 7 summarizes the details of these studies and the bulleted list below provides additional information.

- Russell and Haney (1997) randomly drew examinees from all students enrolled in a special school. All examinees were administered a set of PBT open-ended items in writing, science, math, and reading. Examinees then took a set of NAEP items (science, math, and language arts) and one extended writing item either in a computer-based or paper-based format. The open-ended items taken on paper by all examinees were scored by a single rater. For the two tests examined for mode differences, the hand-written responses to open-ended items were typed verbatim into the computer to minimize rater bias. These items were scored by three raters, and the average score across the three raters was used as the dependent variable. The groups were found to differ on the open-ended items taken on paper, so these scores were used as a covariate in analyzing the NAEP items and writing item. Speededness was also examined across the two groups, and the two groups were found to differ in their test completion status, with the CBT examinees having higher test completion rates. The researchers noted that the proctors

may have allowed extra time to students in the CBT condition because it was a new experience. Threats to the validity of this design are the use of an instrument having potentially significant error (the open-ended items) when using covariance methods, and the possibility of differences between test conditions.

- In a NAEP study (Sandine et al., 2005), researchers administered a block of twenty items on paper to all examinees, and then administered either a computer-based or paper-based test. All examinees answered a set of background questions at the end of the test administration session. The subject under study was math, and participants were eighth-grade students. A multi-stage, probability-based sampling strategy was used. Open-ended responses were scored in the mode in which they were taken. The paper-administered common items were used both as a covariate for mean comparisons and to enable item level comparisons. Item response theory was used to estimate proficiency values, which were then transformed into scale values. Aside from the threats described above, there were no threats to the validity of this design.
- Zhang and Lau (2006) used SAT-9 scores as the common test, and then used a state test for their comparability study in reading (fifth and eighth graders) and math (eighth graders). The students in the study were those requiring a retest, having been assigned into “below standard” and “well below standard” from the first test. The common test was used to equate the item parameters and create RS-SS tables, and then compare the RS-SS tables for the two modes. Students were not randomly assigned into conditions as assignment into a condition was based upon consultation with the student. Thus the threats to the validity of this design are potential group differences due to self-selection into groups, as well as restriction of range and issues with generalizability due to the use of retest students.

Table 7. Studies Using a Between-subjects Design and Additional Tests

Study	Details
Russell and Haney (1997)	<p><i>Instruments:</i> subtest of NAEP multiple choice and constructed response items in language arts, science, and math; locally-created performance writing item; locally-created subtest using all open-ended items in writing, science, math, and reading (used as a covariate) <i>Sample:</i> 6th, 7th, and 8th grade (n=89) <i>Other measures:</i> locally-created subtest (administered on paper) used as a covariate <i>Purpose:</i> score equivalence</p>
Sandine, Horkay, Bennett, Allen, Braswell, Kaplan, and Oranje (2005)	<p><i>Instruments:</i> NAEP math items, multiple choice and constructed response <i>Sample:</i> 8th grade (n=1970) <i>Other measures:</i> additional PBT administered as a covariate, independent review of items for extent of modification for CBT, background survey asking about demographic variables and computer use, familiarity, and skills <i>Purpose:</i> score and construct equivalence</p>
Zhang and Lau (2006)	<p><i>Instruments:</i> Delaware state assessment, SAT-9 reading and math, mostly multiple choice with some short-answer and extended-response items <i>Sample:</i> 5th-grade (n=570) and 8th-grade (n=330) reading, 8th-grade math (n=801) <i>Other measures:</i> additional PBT administered and used for common-item equating, survey of test use, observations of testing conditions <i>Purpose:</i> score equivalence</p>

Matching Designs

Matching designs are a viable alternative when random assignment and/or repeated measures studies are not possible. In matching designs for examining mode comparability, a small study sample is administered a test in one mode (often, a CBT) and a comparable sample based on key matching variables is drawn from a larger sample administered a test in another mode (often, a PBT). The small study sample is likely a special sample of volunteers who take a test on computer. The larger PBT sample is likely gathered from operational PBT testing. A primary goal in matching designs is that the matched sample is comparable on the critical variables of concern. Two promising approaches to creating a matched sample are described below: the matched samples comparability analysis (MSCA) and propensity score matching.

Matched Samples Comparability Analysis (MSCA)

Way, Davis, and Fitzpatrick (2006) introduced a matching design to conduct a comparability analysis in the context of linking. They called their design “Matched Samples Comparability Analysis (MSCA).” In their study, a small sample of CBT responses was collected and compared to a much larger sample of operational PBT responses. The MSCA uses a bootstrap design (i.e., random draws with replacement) with equating to determine the magnitude of equating error across modes and at all score points on a test. The MSCA design for this study used the previous year’s test scores as matching variables to conduct studies each of math, science, and social studies. Presumably, any combination of matching variables can be used.

The MSCA involves implementing the following procedure:

1. Collect a bootstrap sample from the CBT sample equal to the size of the CBT sample.
2. Collect a bootstrap sample of the same size from the PBT sample and matched on both matching variables.
3. Conduct raw score to raw score equating using IRT true score equating.
4. Transform the raw scores to scale scores using the operational RS-SS tables.

The authors used 500 bootstrap replications in their study, although presumably any reasonable number of replications can be used. Once all samples have been collected, the CBT RS-SS tables can be created by averaging the scale scores across all samples at each score point. In addition, the standard deviation at each score point can also be calculated and reflects the error in linking at each raw score point.

In the same publication, the authors reported the results of a simulation study that examined two situations when examinees differ on ability based on prior test performance: the performance of MSCA when mode differences do not exist, and the performance of MSCA when mode differences do exist. Using PBT data in math and reading, the authors generated six datasets, each with different frequency distributions. They examined the performance of MSCA in the context of mode effect differences of 0, .25, .5, and 1.0. The MSCA appropriately did not detect significance using a 95 percent confidence interval when no mode effects existed, and did detect significance when mode effect differences were .5 or 1. The MSCA had difficulty detecting significance when the mode effect was .25.

The MSCA approach has so far been limited to use in the TAKS testing program. Potential threats to the validity of this design are improper use of a matching variable (resulting in non-comparable groups on key variables) and differences in testing conditions such as low motivation or Hawthorne effects.

Propensity Score Matching Designs

Propensity score matching (D’Agostino, 1998; Rosenbaum, 1995; Rosenbaum & Rubin, 1983; Rubin, 2006) is a relatively new and efficient method for producing a matched group design. The method is a refinement of a more general matching or covariate design. Matched groups are

created to reduce bias that may result in a two-group design where randomization is difficult or impossible to implement fully, or when a within-subjects design is ruled out by administrative difficulties. Three examples where propensity score matching has been used are listed below.

- a study of the effect of offering accommodations, such as additional testing time, on student performance (Rudner & Peyton, 2006)
- the effect of taking a test before versus after graduation (Rudner & Peyton, 2006)
- the effect of a change in contractors and item pools on results of a computer-based test (There is currently no publicly available documentation for this example but, like the previous two examples, the study occurred recently with the Graduate Management Admissions Test [Schulz, M., personal communication, March 1, 2007].)

In each of these examples, the experimental group was a relatively small, non-randomly selected group, and the control group was selected from a much larger population that experienced the control treatment. Generally, in a matching design, one selects, for each experimental subject, a control subject with the same standing on one or more matching variables. If only one variable or covariate is used for matching, there is essentially no difference between propensity score matching and traditional matching or covariate analysis. When two or more variables are used for matching, propensity score matching is more powerful and effective. The propensity matching technique uses multivariate logistic regression to form a weighted composite of the covariates—the propensity score—for matching. Propensity score matching is more effective because it uses a weighted composite of covariates, and because it can accommodate missing data and therefore include more covariates and even subjects.

In a comparability study, propensity score matching would work as follows:

1. A group of students is tested using a CBT, and background variables are measured or taken from existing student records.
2. An existing group of students (possibly quite large in number) which has already taken the PBT version of the test is assembled, complete with PBT score and background variables.
3. Logistic regression is used to predict group membership, and each member of both groups is assigned a value of the likelihood of belonging to each group (i.e., a propensity score).
4. For each examinee who took the CBT, an examinee who took the PBT with the nearest propensity score is selected.

Analysis of the CBT and PBT scores follow the matching. As with the MSCA approach, potential threats to the validity of this design are improper implementation of the match (resulting in non-comparable groups) and differences in testing conditions.

Types of Analyses

In order to be considered comparable, the CBT and PBT should measure the same construct and should measure these constructs with the same degree of precision. The types of analyses used in comparability studies can be thought of in terms of the types of hypotheses tested. For example,

it was mentioned earlier that if the CBT and PBT measure the same construct, then the following should be true:

- The test content and content specifications must be the same.
- The scores should have the same factor structure.
- The scores should have the same measurement precision.
- The score distributions should differ only in difficulty, and hence, be equitable.
- The scores should be highly related to one another.
- The scores should have the same relationship to other related measures.

The comparability studies included in this review were classified above in terms of the experimental design employed. They are now classified in terms of the six content validity hypotheses (listed directly above) that were examined using the experimental data gathered by the chosen research design.

Test Content

Content comparability between the paper-based and computer-based modes can be conceptualized as requiring that tests have the same test specifications, have similar items measuring each construct, and require the same skills to answer those items. This issue can be studied regardless of the study design chosen as it relates more to instrumentation than design.

Test Specifications

In the comparability studies considered for this review, the computer-based and paper-based tests were built using the same test specifications. An exception to this rule is computer adaptive testing, where the administered tests vary for examinees, although presumably conditions are placed on the item pool and test administration to ensure parity. The items administered via the computer-based test were often transferred to the computer directly from paper-based administrations, so essentially the “same” items were administered in each mode.

Item Similarity

The degree to which the items are similar across modes is related to the extent to which items need modification. Most studies simply described the process for transferring the paper-based items to the computer format. One method for ensuring similarity was that the same font and style are used (Higgins, Russell, & Hoffman, 2005), or that participants received a paper copy of the test booklet (Russell, 1999). One study had students use the same computers so there was no variation of resolution or screen size across CBTs (Higgins, Russell, & Hoffman, 2005). Sandine et al. (2005) examined the results for participants using study-provided laptops versus school computers. In addition, Sandine et al. (2005) also used an external review to characterize the amount of adaptation needed to use the item on the computer and used this characterization in its analysis of differences in item difficulty across mode. Another issue influencing item similarity is the use of online toolkits (such as a compass, calculator, or ruler). Additionally, the use of color as a navigational or attentional aid may have a differential effect for some examinees (such as color-blind students).

Skills Required

Taking a test on computer may require skills unrelated to the construct. The studies attempted to reduce the influence of potential construct-irrelevant skills by attending to potential differences in the test-taking process. The types of interface issues for the CBTs are listed below.

- use of a tutorial (Higgins, Russell, & Hoffman, 2005; Sandine et al., 2005; Poggio, Glasnapp, Yang, & Poggio, 2005)
- enabling item review throughout the test (Schwarz, Rich, & Podrabsky, 2004; Higgins, Russell, & Hoffman, 2005; Poggio, Glasnapp, Yang, & Poggio, 2005) or sections of the test (Pommerich, 2004) to better mimic a paper-based test
- allowing only one item per screen to be presented (Schwarz, Rich, & Podrabsky, 2004; Poggio, Glasnapp, Yang, & Poggio, 2005), or allowing multiple items to be presented for a testlet (Pommerich, 2004)
- enabling or disabling spell-checking, grammar, and copy/paste functionalities (Russell & Plati, 2001; Sandine et al, 2005; Zhang & Lau, 2006)
- using scrolling or whole-page advancement for text passages that do not fit on a single screen (Pommerich, 2004; Higgins, Russell, & Hoffman, 2005)
- using highlighting or automatic scrolling through stimulus text as an attentional aid (Higgins, Russell, & Hoffman, 2005; Pommerich, 2004)

Related to the issue of skills required, one study (Zhang & Lau, 2006) also reported on the importance of ensuring that staff involved with the CBT administration was properly trained.

Factor Structure

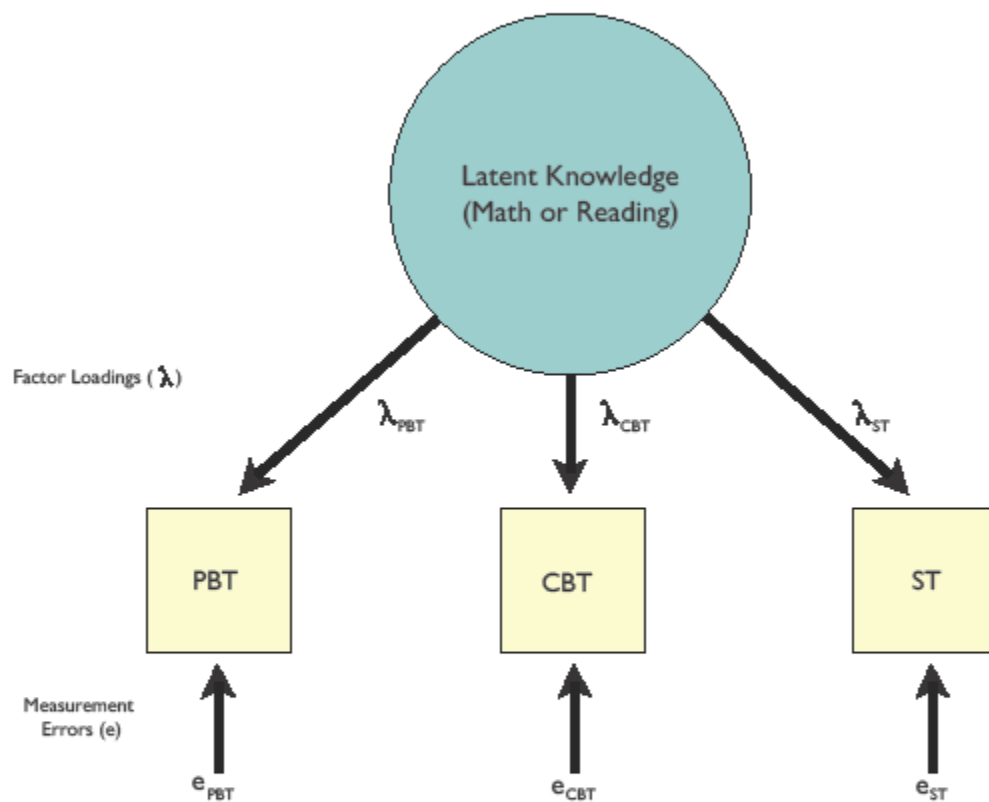
If two instruments measure the same construct, then they should have the same factor structure. In comparability studies, this hypothesis can be examined by using confirmatory factor analysis techniques or a full structural equation model to examine the fit of various hypothesized models. This type of study can only be examined using a within-subjects design because data from paper-based and computer-based tests using the same items need to be collected from each participant. Figure 1 provides a visual depiction of a confirmatory factor analysis approach that might be used. Only one study (Pomplun & Custer, 2005) from the K–12 comparability literature examined the invariance of factor structure across mode. This technique has been used in comparability studies on adults (Boo, 1997; Neumann & Baydoun, 1998).

The Pomplun and Custer (2005) study used a confirmatory factor analysis to determine whether a CBT with three subtests differed in its factor structure from the PBT. Three models were tested. All models used a three-factor model with subtests from each model loading on the appropriate factor. The three factors were allowed to covary, and no error covariances were allowed among the observed variables. The first model placed no constraints on the factor loadings or error variances, and therefore just looked at pattern consistency (i.e., congenicity). The second model constrained the factor loadings to be the same within a factor, and thus required the modes to have the same relationship to the factor (i.e., tau equivalence). The third model constrained the factor loadings and error variances to be the same within a factor, and thus

required the modes to have the same relationship to the factor and same error (i.e., parallelism). Model fit was tested using statistical tests and fit indices.

Although not a K–12 study, Moreno and Segall’s (1997) structural equation modeling approach on Armed Services Vocational Aptitude Battery (ASVAB) data bears noting. In this study, participants with a prior PBT test score were randomly assigned to two groups. In one group, participants took two additional forms of the PBT. In the other group, participants took two computer-adaptive tests (CAT-ASVABs) built to the same specifications as the PBT. Instead of testing model fit for parallelism, tau-equivalence, or congeneric equivalence, these researchers made assumptions regarding relationships among PBT scores and CBT scores (assuming latent correlations among scores were 1 within mode) and estimated the test reliabilities and the relationship between the CBT and PBT. This analysis was conducted for 10 subtests of the ASVAB. The researchers then compared the estimated reliability among tests and presented the disattenuated correlations between PBTs and CBTs (all of which were close to 1). Significance tests for model fit supported the hypotheses that the CBT-PBT correlation was the same as the PBT-PBT and CBT-CBT correlations.

**Confirmatory Factor Analysis Model for
Examining the Comparability of
Paper-Based and Computer-Based Tests (PBT_s and CBT_s)
[Using an Additional Summative Test (ST)]**



Hypotheses	Statistical Tests
1) PBT, CBT and ST have perfectly correlated True Score	Fit of overall model
2) PBT and CBT are parallel	$\lambda_{PBT} = \lambda_{CBT}$ and $\sigma^2(e_{PBT}) = \sigma^2(e_{CBT})$
3) PBT and CBT have same validity for predicting ST	$\lambda_{PBT} = \lambda_{CBT}$

Figure 1. Confirmatory Factor Analysis Model Used for Assessing Comparability of a PBT and CBT

Item level differences can also be considered an indication of factor structure. Item differences were examined using a variety of methods. Once item level differences were determined, the items were reviewed by researchers to explain why such differences occurred.

- Schwarz, Rich, and Podrabsky (2003) used differential item functioning using PBT as the reference group and CBT as the focal group. Poggio, Glassnap, Yang, and Poggio (2005) conducted item level and category level DIF studies with small sample sizes.
- Pommerich (2004) and Higgins, Russell, and Hoffman (2005) used confidence intervals to determine whether item p -values varied significantly across mode. Pommerich (2004) presented confidence intervals for each item studied and also presented the proportion of items statistically favoring each mode.
- Sandine et al. (2005) compared item p -values and IRT b parameters, as well as presented scatterplots of a and b parameters across mode. The comparison of parameters was also categorized by the amount of modification needed to computerize the items. This study had a set of common items administered in the same mode, and these item parameters were constrained to be equal in the calibration process, whereas the parameters of items administered in the two modes were allowed to vary.
- Choi and Tinkler (2005) used a similar methodology as Sandine et al. (2005). Choi and Tinkler (2005) also categorized items by the level of comprehension required (comprehension at the word, phrase, sentence, or discourse level) to answer the item, and then examined bias and error between the b values by mode for each category.
- Johnson and Green (2006) compared p -values of items as well as conducted a qualitative examination of error types (e.g., transcription error, place value error, partial answer, computation error, misunderstanding) made by students in each mode as well as strategies across modes.
- Keng, McClarty, and Davis (2006) examined differences in mode at the item level by comparing p -values, differences in choices across response options, and computing IRT-based differential item functioning tests. These researchers also reviewed a sample of test booklets to determine whether scratchwork on the booklet was associated with predicted item performance.

Measurement Precision

If two tests are comparable, then they should have the same measurement precision both overall and across proficiency levels. This issue can be studied using any design presented, although the indicator of measurement precision may vary across designs. Measurement precision can be examined at the overall test level and can be examined by looking at the consistency of individual constructed-response item ratings.

Two methods for examining measurement precision at the test level are Cronbach's alpha and IRT information curves. Cronbach's alpha and the standard errors of measurement should be the same across modes, as should the test information functions. Classical test theory reliability was estimated in a few studies (Olson, Maynes, Slawson, & Ho, 1989; Zhang & Lau, 2006). Surprisingly, only one study provided IRT test information curves to provide a comparison of standard error conditional on theta for modes (Poggio, Glassnap, Yang, & Poggio, 2005).

Rater agreement statistics (such as Cohen's kappa, exact agreement rates, and the proportion of essays needing adjudication by an external rater) can be used to compare consistency for constructed-response items. While researchers often provided indices of measurement precision across modes, few provided indices within modes and then compared precision across modes. Sandine et al. (2005) provided exact agreement rates for constructed-response items for both modes in the NAEP assessments. Way and Fitzpatrick (2006) compared rater agreement (kappa, exact agreement, and proportion needing adjudication by a third rater) for a single essay by mode for the TAKS. This study also examined the impact of mode using automated essay scoring. Five samples (71 in human scorer calibration, 300 PBT, 300 CBT, 300 PBT and 300 CBT, and 150 PBT and 150 CBT) were used to calibrate the automated essay scoring engine, and then the five trained engines were used to score both handwritten and computer-entered essays. Mode effects were detected when the engine trained in one mode scored same-mode essays more consistently than other-mode essays.

Score Distributions

If two tests are comparable across modes, then their score distributions (e.g., means, standard deviations, frequency distributions) should be the same. A comparison of score distributions can be made for any of the designs described. Equating studies and the computation of RS-SS tables also provide information about score distribution.

Score Comparisons

Comparisons of mean test scores were made in all the studies considered. Studies compared raw score means (Hollenbeck, Tindal, Stieber, & Harniss, 1999; Russell & Tao, 2004; Schwarz, Rich, & Podrabsky, 2004; Way, Davis, & Fitzpatrick, 2006; Russell, 1999; Sandine et al., 2005; Russell & Haney, 1997; Zhang & Lau, 2006; Eignor, 1993; Johnson & Green, 2006; Olson, Maynes, Slawson & Ho, 1989), scale score means (Schwarz, Rich, & Podrabsky, 2004; Way, Davis, & Fitzpatrick, 2006; Zhang & Lau, 2006; Pomplun & Custer, 2005; Pomplun, Frey, Becker, & Hughes, 2000; Poggio, Glasnapp, Yang, & Poggio, 2005), and IRT proficiency (theta) score means (Choi & Tinkler, 2002; Sandine et al., 2005). For studies using IRT, item parameters were constrained to be equal across mode so that any mode differences appeared in the proficiency scores. Other measures of distributional differences were computed in only a few studies. Pomplun and Custer (2005) examined equivalence of variance of mean scores. Eignor (1993) presented histograms of scores. When groups differed on the pretest, researchers often used multiple regression to obtain means adjusted for those differences (Russell & Haney, 1997; Russell, 1999).

Equating

Researchers conducted equating studies to examine mode effects using a variety of different methods.

- Choi and Tinkler (2002) calibrated items in an IRT program twice, once indicating group membership and once not indicating membership. They then used a statistical test to compare model fit. A test administered to the two groups in the same mode was used as a basis for comparison.

- Zhang and Lau (2006) examined the mode effect using a PBT test administered to all examinees prior to the test examined for comparability. This PBT was used for common-item equating and so differences in equated parameters could be examined across mode.
- Fitzpatrick and Triscari (2005) equated the item parameters from each mode to existing Rasch item parameters gathered from an operational paper-based test administration. A difference between equating parameters would then be attributed to mode differences (although no overall difference was detected).
- Way, Davis, and Fitzpatrick (2006) used a bootstrap technique to determine standard errors in equating paper-based and computer-based test scores. Regions of the IRT proficiency scale were noted where scale score differences were outside the 95 percent confidence interval. The researchers also conducted a simulation study to determine whether this bootstrap methodology would be able to distinguish between differences in mode and differences in group ability. This study examined the standard errors produced when no mode effects existed but group effects existed, and when mode effects existed but group effects did not.

Raw Score to Scale Score Tables

Closely related to equating was the comparison of raw score to scale score tables. Three studies computed RS-SS tables following an equating procedure and then compared the scale scores across the raw score range between the two modes (Way, Davis, & Fitzpatrick, 2006; Fitzpatrick & Triscari, 2005; Eignor, 1993). Because Way, Davis, and Fitzpatrick (2006) had computed standard errors in the equating process, they could determine where the scale scores differed significantly across the two modes. Eignor (1993) conducted both linear and curvilinear (equipercentile) equating to determine which method was appropriate and then computed RS-SS tables for each mode and testing order. In one test, the RS-SS tables were similar within mode and across order, so those data were combined. On another test, the RS-SS tables differed within mode and across order. As a result, these results were combined using a weighted sum. Fitzpatrick and Triscari (2005) generated separate RS-SS tables for the CBT and PBT, applied cut-scores derived for the PBT score to the CBT scores, and then examined the proportions of PBT and CBT examinees scoring at each proficiency level. Two studies (Zhang & Lau, 2006; Choi & Tinkler, 2002) presented differences in test characteristic curves across the proficiency continuum to reflect mode differences.

Relationship of Scores

If the two tests are comparable, scores gathered from both tests should be highly related. If norm-referenced comparisons are being made, then the correlation corrected, for unreliability, should be 1.¹⁷ If criterion-referenced judgments are made, then the agreement among levels should be very high (80 percent or higher agreement). Relationships among scores can only be computed in within-subjects designs (and possibly in matched-group designs). Correlations of raw, scaled,

¹⁷ A note of caution is in order here: if coefficient alpha is used in correcting for unreliability, this will often result in an over-correction because alpha is a lower-bound estimate to reliability.

and/or IRT proficiency scores were computed in a number of studies (Sandine et al., 2005; Eignor, 1993; Pomplun & Custer, 2005; Olson, Maynes, Slawson, & Ho, 1989; Pomplun, Frey, Becker, & Hughes, 2000; Poggio, Glasnapp, Yang, & Poggio, 2005). None of the within-subjects designs reported a statistic on rater agreement.

The analysis of assignment of examinees into performance categories is an important issue that was not often considered in the studies. Analyses of performance category placement are compelling because these placements are of primary concern to many K–12 stakeholders. Only one study (Zhang & Lau, 2006) compared the proportion of examinees assigned into performance levels by mode. This study also presented agreement tables of performance level assignment from each mode of the studied test and from a test administered a few months prior. Thus assignment into categories could be compared across mode. Court (2006)¹⁸ conducted a follow-up study to Poggio, Glasnapp, Yang, & Poggio (2005) using separately collected data from a within-subjects design. This study presented the overall proportion of examinees assigned to performance levels by mode as well as the proportion of examinees being scored in the same or a different category based upon the mode of test.

Relationship with Other Variables

If two tests are comparable, then they should relate to other factors to the same degree. These factors may be specifically related to the test under study (such as completion rate, time spent) or related to the construct under study.

Test Issues

Researchers have examined a variety of factors related to the test itself. The list of studied characteristics is listed below, along with the studies in which these factors were examined.

- completion rate (Higgins, Russell, & Hoffman, 2005; Pommerich, 2004; Sandine et al., 2005; Russell & Haney, 1997)
- time to complete the test (Russell & Plati, 2004)
- surface features of constructed-response text such as number of characters, number of words, variation in sentence length (Russell & Plati, 2004; Way & Fitzpatrick, 2006; Russell and Haney, 1997; Sandine et al., 2005)
- proportion of valid responses to constructed-response items (Sandine et al., 2005)

Other Factors

The relationship of the test to other constructs has also been studied. This work has generally been done using regression or ANOVA. In these models, the impact of mode was examined controlling for other factors (such as a pretest, typing speed) or interactions were examined for

¹⁸ This study was not included in detail since it did not provide enough methodological details of the design to warrant inclusion.

other factors (such as gender and race). One problem with this analysis approach is that regression methods assume the variables have no measurement error, and the influence of error on weights is unknown. Structural equation modeling is one methodology that could be used to account for error in the measure and compare relationships among factors.

- Higgins, Russell, and Hoffman (2005) used regression to predict raw test scores using a variety of related measures or examinee characteristics and mode of test. The measures used were computer fluidity, computer literacy, home computer use, and school computer use. The examinee characteristics used were gender and whether the student had an individualized education plan (IEP).
- Russell and Plati (2001) used regression to predict constructed-response trait scores and total scores using mode and other variables (typing speed, mid-term English grades).
- Russell (1999) examined typing speed as well as gender.
- Way and Fitzpatrick (2006) examined the relationship of examinee self-rated computer skills and computer use and student performance using ANCOVA.
- Sandine et al. (2005) conducted a series of ANOVAs on raw scores with a host of other factors (gender, race, parental education level, region of country, school type, type of computer used, typing speed, typing accuracy, editing skill).
- Pomplun and Custer (2005) presented the mean difference of examinees receiving free lunch and those not receiving free lunch by mode. No statistical analyses were conducted on the means.
- Poggio, Glasnapp, Yang, and Poggio (2005) presented the mean difference across a number of dimensions: gender, SES (indicated by examinee receiving no lunch support, free lunch, or reduced lunch), and academic placement category (general education, gifted, or special education).

Summary and Recommendations

The successful evaluation of the comparability of computer-based and paper-based test scores requires a strong inference study design with a focus on key comparability issues. Researchers have used a variety of designs and analyses to examine different aspects of comparability. Table 8 outlines the key issues to consider when designing a study that can generalize to the population of interest and can best identify the impact of testing mode on score and construct equivalence. It is expected that these issues would be addressed in any comparability study. Table 9 presents key comparability questions and hypotheses, organized under the areas of score and construct equivalence. Table 9 also suggests the types of designs and analyses that could be used to test hypotheses.

Because K–12 comparability studies occur in a social and political context, their design and implementation will be shaped by administrative conditions that may conflict with a strong inference design. The investigator must often deal with volunteer samples, cannot control all variables, cannot use random assignment, and/or cannot use a within-subjects design. While these limitations influence the ability to identify the impact of mode and the generalizability of the results, such studies are still important to pursue. The investigator should disclose any design limitations and provide an analysis of the results in light of those limitations.

The assessment of whether scores from a paper-based test are comparable to a computer-based test is, finally, a matter of judgment. The evidence used to support the hypotheses is either judgmental or statistical. In the case of expert review, the investigator needs to interpret the results in light of problems with the limitations of human judgment. In the case of statistical methods, the interpretation involves estimates of both statistical and practical significance. Presumably, the final assessment occurs when a satisfactory amount of evidence that supports the hypotheses has been collected and little or no evidence contradicts the hypotheses.

Table 8. Issues to Consider in Comparability Study Design

Validity Issue	Suggested Design Features
Sampling	<ul style="list-style-type: none"> • identify the population from which to sample (e.g., general population, ESL students, IEP students) • recruit a sample of sufficient size for statistics used • outline incentives for school or examinee participation • monitor attrition rates • present characteristics of participants <i>and</i> non-participants • use counterbalancing in within-subjects designs • assign examinees to groups to minimize nuisance effects (using random assignment, matching, post hoc statistical methods) • monitor the assignment of examinees to groups to ensure sampling plan is properly implemented
Instrumentation	<ul style="list-style-type: none"> • study a test that is comparable to one used in practice • study a test with enough items to adequately represent the construct • ensure that the CBT is built to the same test specifications as the PBT • monitor extent to which items are modified for the CBT • consider influence of item type (multiple choice, constructed response) • report on test presentation features (e.g., ability to review items, number of items presented on the screen) • report on item presentation features (i.e., ability to scroll) • identify editing features (e.g., grammar- or spell-checking) permitted
Administration	<ul style="list-style-type: none"> • ensure test administration conditions are similar to those in practice • provide practice tests or a tutorial for CBT examinees • provide a tutorial or other training for test administrators • ensure computers have the appropriate software and hardware • consider examinee access to computers • consider bandwidth issues • record technical problems of the CBT administration • ensure examinee motivation levels are reasonably similar to those in practice • administer tests to groups within a reasonably similar time period
Scoring	<ul style="list-style-type: none"> • use scoring methods that reflect those used in practice • identify the type of score used (e.g., raw, scale, IRT theta) • outline the equating process for scores and item parameters • present inter-rater reliability of constructed responses, if applicable • report measurement precision statistics • report score distribution information • report performance category assignment, if applicable • disaggregate data if there is reason to suspect group differences

Table 9. Study Design and Analyses by Type of Comparability Question

Score comparability		
Focus	Design	Suggested Analyses
<ul style="list-style-type: none"> The scores should have the same measurement precision. 	BS or WS	<ul style="list-style-type: none"> overall reliability values overall standard error of measurement conditional SEM/test information
<ul style="list-style-type: none"> The score distributions should differ only in difficulty, and hence, be equitable. 	BS or WS	<ul style="list-style-type: none"> frequency distribution of scores or histogram measures of central tendency and dispersion (mean, standard deviation) RS-SS tables distribution of examinees assigned to performance levels
Construct comparability		
Focus	Design	Suggested Analyses
<ul style="list-style-type: none"> The test content and content specifications should be comparable. 	None, BS or WS	<ul style="list-style-type: none"> blueprint comparisons expert review of item CBT modifications which may require additional skills comparisons of different test formats
<ul style="list-style-type: none"> The scores should have the same factor structure. 	BS or WS	<ul style="list-style-type: none"> tests of dimensionality at item level tests of dimensionality at item parcel (item groups) level item level comparisons of CBT and PBT parameters (IRT, CTT)
<ul style="list-style-type: none"> The scores should be highly related to one another. 	WS	<ul style="list-style-type: none"> correlation (corrected and uncorrected for unreliability) agreement of assignment into performance levels
<ul style="list-style-type: none"> The scores should have the same relationship to other related measures. 	BS or WS, C	<ul style="list-style-type: none"> correlation (corrected and uncorrected for reliability) theory should drive inclusion of measures (e.g., test anxiety influencing PBT scores) other measures can include tests of ability, attitude measures, examinee characteristics (such as gender and race)

Design: WS=Within-subjects, BS=Between-subjects, C=Criterion

References

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2004). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- American Psychological Association (1986). *Guidelines for computer-based tests and interpretations*. Washington, DC: Author.
- Bennett, R.E. (2003). Online Assessment and the Comparability of Score Meaning (ETS Research Report RM-03-05). Princeton, NJ: Educational Testing Service.
- Bergstrom, B.A. (1992). *Ability measure equivalence of computer adaptive and pencil and paper tests: A research synthesis*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.
- Boo, J. (1997). Computerized versus Paper-and-Pencil Assessment of Educational Development: Score Comparability and Examinee Preferences. Unpublished doctoral dissertation, University of Iowa.
- Campbell, D. T. & Stanley, J.C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (ed.), *Handbook of Research on Teaching*. Chicago: Rand McNally.
- Choi, S.W., & Tinkler, T. (2002, April). *Evaluating comparability of paper-and-pencil and computer-based assessment in a K–12 setting*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Court, S. (2006, April). *The interchangeability of dual-mode testing results (CBT vs. PPT)*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- D’Agostino, R.B. (1998). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, *17*, 2265–2281.
- Educational Research Center (2006, May 4). Technology Counts 2006: The information Edge. *Education Week*, *25*(35).
- Eignor, D.R. (1993). *Deriving comparable scores for computer adaptive and conventional tests: An example using the SAT*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta, GA.

- Fitzpatrick, S., & Triscari, R. (2005). *Comparability studies of the Virginia computer-delivered tests*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.
- Gaskill, J. (2006). *Comparisons Between Paper- and Computer-Based Tests: A Literature Review*. Kelowna, BC, Canada: Society for the Advancement of Excellence in Education.
- Higgins, J., Russell, M., & Hoffman, T. (2005). Examining the effect of computer-based passage presentation on reading test performance. *The Journal of Technology, Learning, and Assessment*, 3(4).
- Hollenbeck, K., Tindal, G., Steiber, S., & Harniss, M. (1999). *Handwritten vs. word processed statewide compositions: Do judges rate them differently?* Unpublished manuscript, University of Oregon, BRT. Retrieved November 30, 2006 from http://brt.uoregon.edu/files/Hdwrtn_vs_Typed.pdf
- International Test Commission. (2005). *International guidelines on computer-based and internet-delivered test*. Granada, Spain: International Test Commission.
- Johnson, M., and Green, S. (2006). On-line mathematics assessment: The impact of mode on performance and question answering strategies. *The Journal of Technology, Learning, and Assessment*, 4(5).
- Keng, L., McClarty, K.L., & Davis, L.L. (2006). *Item-level comparative analysis of online and paper administrations of the Texas assessment of knowledge and skills*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.
- Kim, J.-P. (1999). *Meta-analysis of equivalence of computerized and p and p tests on ability measures*. Paper presented at the Annual Meeting of the Mid-Western Educational Research Association, Chicago, IL.
- Mead, A.D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 114(3), 449–458.
- Moreno, K.E., & Segall, D.O. (1997). Reliability and construct validity of CAT-ASVAB. In W.A. Sands, B.K. Waters., and J.R. McBride (Eds.), *Computerized Adaptive Testing: From inquiry to operation* (pp. 169–174). Washington, DC: American Psychological Association.
- Nuemann, G., & Baydoun, R. (1998) Computerization of paper-and-pencil tests: When are they equivalent? *Applied Psychological Measurement*, 22(1), 71–83.
- Olson, J.B., Maynes, D.D., Slawson, D., & Ho, K. (1989). Comparison of paper-administered, computer-administered and computerized achievement tests. *Journal of Educational Computing Research*, 5, 311–326.

- Olson, L. (2003, May 8). Legal twists, digital turns: Computerized testing feels the impact of “No Child Left Behind.” *Education Week* 12(35), 11–14, 16.
- Paek, P. (2005). Recent Trends in Comparability Studies (PEM Research Report 05-05). Austin, TX: Pearson Educational Measurement.
- Poggio, J., Glasnapp, D.R., Yang, X., & Poggio, A.J. (2005). A comparative evaluation of score results from computerized and paper and pencil mathematics testing in a large scale state assessment program. *The Journal of Technology, Learning, and Assessment*, 3(6).
- Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *The Journal of Technology, Learning, and Assessment*, 2(6).
- Pomplun, M., Frey, S., Becker, D., & Hughes, K. (2000). *The validity of a computerized measure of reading rate*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.
- Pomplun, M., & Custer, M. (2005). The score comparability of computerized and paper-and-pencil formats for K-3 reading tests. *Journal of Educational Computing Research*, 32(2), 153–166.
- Rosenbaum, P.R. (1995). *Observational studies*. New York: Springer-Verlag.
- Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
- Rubin, D.B. (2006). *Matched sampling for causal effects*. New York: Cambridge University Press.
- Rudner, L. M., & Peyton, J. (2006, May). *Consider Propensity Scores to Compare Treatments*. Research Report RR-06-07. Graduate Management Admissions Council, McLean, VA.
- Russell, M. (1999). Testing on computers: a follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives (online)*. Retrieved November 30, 2006 from <http://epaa.asu.edu/epaa/v7n20/>
- Russell, M., & Haney, W. (1997). Testing Writing on Computers: Results of a Pilot Study to Compare Student Writing Test Performance via Computer or Via Paper-and-Pencil. *Educational Policy Analysis Archives*, 5(3).
- Russell, M., & Plati, T. (2001). Effects of computer versus paper administration of a state mandated writing assessment. *Teachers College Record*. Retrieved November 30, 2006 from <http://escholarship.bc.edu/cgi/viewcontent.cgi?article=1012andcontext=intasc>

- Russell, M. & Tao, W. (2004). Effects of handwriting and computer-print on composition scores: a follow-up to Powers, Fowles, Farnum, & Ramsey. *Practical Assessment, Research and Evaluation*, 9(1). Retrieved November 30, 2006 from <http://PAREonline.net/getvn.asp?v=9andn=1>.
- Sandine, B., Horkay, N., Bennett, R.E., Allen, N., Braswell, J., Kaplan, B., & Oranje, A. (2005). Online Assessment in Mathematics and Writing: Reports from the NAEP Technology-Based Assessment Project, Research, and Development Series (NCES 2005—457). U.S. Department of Education, National Center for Education Statistics. Washington, DC. U.S. Government Printing Office.
- Schwarz, R.D., Rich, C., & Podrabsky, R. (2003). *A DIF analysis of item-level mode effects for computerized and paper-and-pencil tests*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.
- Way, W.D., & Fitzpatrick, S. (2006). *Essay responses in online and paper administrations of the Texas assessment of knowledge and skills*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.
- Way, W.D., Davis, L.L., & Fitzpatrick, S. (2006). *Score comparability of online and paper administrations of the Texas assessment of knowledge and skills*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.
- Zhang, L, & Lau, C.A. (2006, April). *A comparison study of testing mode using multiple-choice and constructed-response items – Lessons learned from a pilot study*. Paper presented at the Annual Meeting of the American Educational Association, San Francisco, CA.

Chapter 8: Validity Issues and Empirical Research on Translating Educational Achievement Tests

Stephen G. Sireci
University of Massachusetts

Abstract

Contemporary educational assessments in the United States are typically administered in English even though many students are not fully proficient in English. Whenever an educational assessment is designed to measure skills other than English proficiency, students' proficiency with the English language can interfere with valid measurement of their achievement. This potential invalidity is particularly likely for English language learners (ELLs). To address this problem, several states use translated versions of the assessments for testing ELLs. In this report, we review the issues associated with test translations and review the empirical literature that has evaluated translated educational tests. Empirical work in this area indicates that statistical and qualitative checks throughout the test development and validation processes are needed to evaluate the comparability of scores from original and translated tests, and that these tests may not always be comparable to their original English-language counterparts. Caution should be exercised whenever comparisons are made across students who take different language versions of an assessment, and more research on translated tests is needed to guide proper interpretation of scores from these assessments.

Introduction

Mandated assessments are a common component in contemporary educational reform efforts. Standardized tests are used in these efforts because they ensure uniform content, test administration, and scoring procedures. However, in many cases not all students have equal access to an assessment due to limited proficiency in the language in which the test is administered. When linguistic diversity is present in an educational system, a test administered in the dominant language of the system may not provide a level playing field for all students. Specifically, administering a test in one language may disadvantage students who are not fully proficient in that language, or worse yet, it may make the test inaccessible to them altogether.

The two most recent versions of *Standards for Educational and Psychological Testing* have acknowledged the problem of language proficiency interfering with proper measurement of students' knowledge, skills, and abilities (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999; APA, AERA, & NCME, 1985). For example, the most recent version of the standards stated,

...any test that employs language is, in part, a measure of their language skills... This is of particular concern for test takers whose first language is not the language of the test... In such instances, test results may not reflect accurately the qualities and competencies intended to be measured. (AERA, et al., 1999, p. 91)

To address this concern, many testing programs have developed different language versions of a test (Stansfield & Bowles, 2006). To maximize comparability of these adapted tests¹⁹ to the original versions, these alternate test forms are most often translated versions of the standard English-language test. The logic behind test translation is that it allows for more accurate measurement of students' skills, since it removes construct-irrelevant variance due to second language proficiency. This logic is reflected in recent legislation regarding the assessment of English language learners (ELLs²⁰). For example, Stansfield (2003) points out that the No Child Left Behind Act (NCLB) requires states to assess ELLs using "assessments in the language and form most likely to yield accurate data on what such students know and can do in academic content areas, until such students have achieved English language proficiency" (NCLB, 2002, cited in Stansfield, 2003, p. 203).

A variant of the administration of a translated test form that has been used in some cases is development and administration of a *dual-language* test. This variant involves creating test forms that display two different language versions of each item side-by-side in the same test booklet.

In the United States, there are several states that use translated or dual-language versions of their state assessments to measure the proficiencies of students who are ELLs. These alternate test forms are typically offered in math, science, or social studies, where measurement of English proficiency is *not* a purpose of the test. For example, Delaware administers Spanish versions of their math, science, and social studies tests; Massachusetts administers dual-language versions of math and science tests; and Michigan provides oral translations of their math, science, and social studies tests in Arabic and Spanish. Other states that offer alternate-language versions of their state assessments include Florida, New Mexico, Ohio, New York, Rhode Island, and Utah. In many cases the alternate-language versions are translations into Spanish, which represents the largest non-English language group in U.S. schools. However, in addition to the aforementioned Arabic and Spanish assessments, statewide tests have been translated into over a dozen languages including Cambodian, Chinese, Haitian Creole, Portuguese, Russian, and Vietnamese (for examples, see the web site for Second Language Testing, Inc. at http://www.2lti.com/htm/testta_Clients.htm).

¹⁹ The terms *translation* and *adaptation* are used interchangeably in this paper. Test translation is the more familiar term, but adaptation is the more appropriate term since it refers to the broader process of conveying the same meaning across different language versions of a test.

²⁰ We use the term English language learner or ELL in this report to refer to students within the United States whose native language is not English, but have an educational goal to increase their English proficiency. Such students are sometimes referred to as having limited English proficiency (LEP), however, that term is problematic because complete proficiency in English can never be unequivocally demonstrated by anyone. In any case, the issues discussed in this paper apply in full force whenever someone is tested in a language in which they are not sufficiently proficient to understand and respond to the test items.

Translated versions of statewide assessments attempt to increase the validity of inferences derived from ELLs' test scores by removing English proficiency as a barrier to proper measurement of the skills targeted by an assessment. However, this practice raises several concerns. One concern is whether the translated version of a test is easier or more difficult than the original, English-language version. Another concern is whether the translation changes the construct measured by the assessment. A third concern is whether scores from the original and translated versions of these assessments are comparable. This concern is particularly important whenever scores from English and translated-language versions of a test are aggregated for program accountability purposes.

Are such concerns valid, or can it be assumed that translated tests are equivalent to their original, English-language counterparts? Research provides a clear answer to that question. It has been repeatedly demonstrated that simply translating an exam into another language, even when done with great care, is not enough for ensuring the validity of score interpretations or the comparability of scores across the original and adaptive versions (Angoff & Cook, 1988; Bechger, van den Wittenboer, Hox, & De Gloppe, 1999; Gierl & Khaliq, 2001; Hambleton, Sireci & Robin, 1999; Robin, Sireci, & Hambleton, 2003; Sireci, 1997; Sireci & Khaliq, 2002). Extensive quality control procedures that involve both statistical and judgmental analyses should be put in place to achieve defensible levels of validity and comparability (Hambleton, 1994, 2001, 2005; van de Vijver & Hambleton, 1996).

The purpose of the present study is to summarize the literature on test translation with a focus on research and practice in K–12 assessment in the United States. In the following sections, we

- describe contemporary assessment systems in which test translations are being used
- discuss validity issues in testing ELLs and in translating educational tests
- review the literature that evaluates the validity of translated/adapted educational tests

Given that this paper was commissioned by policymakers and practitioners in K–12 education, we focus on K–12 testing to the greatest extent possible.

Adapting Educational Tests for Use across Linguistically Diverse Populations

The past several decades have seen an explosion in the adaptation of tests for use across multiple languages. In addition to the aforementioned statewide educational assessments in the U.S., other educational areas in which test translation activities are abundant include licensure and certification testing (e.g., Fitzgerald, 2005; Robin et al, 2003; Sireci, Foster, Olsen, & Robin, 1998), employment testing (Sireci, Harter, Yang, & Bhola, 2003; Sireci, Yang, Harter, & Ehrlic, 2006), educational assessment in Canada (Davis, Buckendahl, & Plake, 2006; Ercikan & Koh, 2005; Gierl, 2000; Gierl & Khaliq, 2002), and international comparisons of educational achievement, such as the Trends in International Mathematics and Science Study (TIMSS) (measuring the math and science proficiencies of students in 45 countries), the Program for International Student Assessment (PISA [Organization for Economic Co-operation and Development, 2000]), which assesses the reading, math, and literacy skills of 15 year olds in 32 countries, and the Progress in International Reading Literacy Study (PIRLS, [Campbell, et al., 2001]), which assessed the reading skills of fourth-grade students in approximately 40 countries.

In addition, admissions tests have also been translated for use in the U.S. (Cook & Schmitt-Cascallar, 2005) and Israel (Beller, Gafni, & Hanani, 2005; Sireci & Allalouf, 2003).

Given the widespread activity of test adaptation across languages, it is clear that adapted versions of tests are seen as a useful means for overcoming linguistic barriers in assessing and comparing examinees who differ with respect to their dominant language. However, popularity in use certainly does not signify validity. That is, the degree to which test adaptations are appropriate for their intended purposes must be evaluated whenever tests are adapted for use across languages. In the next section, we discuss the validity issues involved in test adaptation.

Validity Issues in Test Adaptation

Thus far, we have discussed English proficiency as a potential barrier to valid measurement of ELLs' knowledge and skills and we noted that test adaptations are often used to remove this barrier. Before extending our discussion of issues relevant to the validity of adapted tests, we must first formally define validity. To do that, we turn to the *Standards for Educational and Psychological Testing* as the authoritative source.²¹ According to the *Standards*, validity refers to “the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (AERA et al., 1999, p. 9). From this definition, we see that it is not a test per se that is validated, but rather the decisions or interpretations made on the basis of test scores. Therefore, in considering validity issues involved in the assessment of ELLs, we must consider the purpose of the assessment. In statewide educational testing, the purposes are typically to evaluate students' achievement with respect to well-defined curriculum frameworks and performance standards, and provide information useful for evaluating teachers, schools, and school districts. Thus, evaluations of translated versions of statewide educational tests must keep these purposes in mind.

Standards for Educational and Psychological Testing describes the desirable qualities of tests that are needed to promote valid test score interpretations. Although most of the guidelines pertain to measurement within a single language, the standards also state, “when multiple language versions of a test are intended to be comparable, test developers should provide evidence of score comparability” (AERA et al., 1999, p. 99). When tests are used for accountability purposes and scores from English and translated versions are aggregated, comparability of scores across language versions is assumed. Such an assumption should be empirically verified, which may involve systematic sources of bias that would threaten comparability.

Van de Vijver and his colleagues stated that at least three types of bias could lead to a lack of comparability of test scores across languages: construct bias, method bias, and item bias (van de

²¹ The standards are authoritative in the sense that they represent consensus standards from the three major organizations involved in appropriate test use and that the courts have used them widely whenever tests are challenged (Sireci & Parker, 2006). However, it should be noted that they represent the cumulative thinking of several prominent validity theorists over the years (e.g., Cronbach, 1971; Cronbach & Meehl, 1955; Kane, 1992; Messick, 1989; Shepard, 1993).

Vijver & Poortinga, 1997, 2005; van de Vijver & Tanzer, 1998). Construct bias refers to the situation where the construct measured, as operationally defined by the assessment, is nonexistent in one or more cultures or is significantly different across cultures. This type of bias can be logically ruled out in state-mandated testing because the curriculum frameworks for a state define the content to be measured and the frameworks and test specifications are common for the original and translated versions of the test. Method bias refers to a systematic source of construct-irrelevant variance that manifests at the test score level. Examples of method bias include improper test administration conditions, inappropriate or unfamiliar item formats, or improper test translations that make all test items easier or difficult in one language, relative to the original language. Item bias refers to construct-irrelevant variance that affects performance at the item level.

Comprehensive and careful processes for adapting test material across languages go a long way toward avoiding method and item bias. However, empirical studies are needed to rule out such biases. For this reason, *Guidelines for Adapting Educational and Psychological Tests* (developed by the International Test Commission, see Hambleton, 1994, 2001, 2005) underscores the need for statistical procedures to evaluate test comparability across cultures and languages. The guidelines encourage test developers to use appropriate statistical techniques to evaluate item equivalence and to identify areas of a test that may be inadequate for one or more of the intended groups. For example, the guidelines recommend that test developers conduct differential item functioning (DIF) analyses to evaluate test items designed to be used in two or more cultural or language groups. DIF analyses evaluate whether examinees from different groups (e.g., ELL or non-ELL) who are of comparable ability have equal probabilities of success on an item. Although DIF analyses are useful for identifying problematic items, an evaluation of the dimensionality of adapted tests is prerequisite for ruling out systematic biases at the total test score level that are not detectable at the item level (Sireci, 1997, in press; van de Vijver & Tanzer, 1998).

A summary of the guidance provided by *Standards for Educational and Psychological Testing* and *Guidelines for Adapting Educational and Psychological Tests* with respect to test adaptation is provided in Table 1 (from Sireci, in press). These guidelines emphasize the types of empirical evidence states could gather to support the use of translated tests. Common themes in these guidelines and standards are that rigorous quality control steps should be included in the translation process, and that when translating items, both judgmental and statistical techniques should be used to ensure item comparability across languages. Specifically, evidence based on internal structure can be used to evaluate the consistency of factor structure across original and adapted versions of a test, and evidence based on test content can evaluate the linguistic and statistical equivalence of specific test items across their original and translated versions. In the next sections, we discuss procedures for proper translation of educational tests, and we review the empirical research that has been conducted on state assessments that have employed test translations.

*Table 1. Selected Excerpts from Professional Guidelines Related to Comparability of Translated Tests**

Standards for Educational and Psychological Testing (Source: AERA et al., 1999)	Guidelines for Test Adaptations (Source: Hambleton, 2005)
A clear rationale and supporting evidence should be provided for any claim that scores earned on different forms of a test may be used interchangeably (p. 57).	Instrument developers/publishers should compile judgmental evidence, both linguistic and psychological, to improve the accuracy of the adaptation process and compile evidence on the equivalence of all language versions (p. 22).
When substantial changes are made to a test, the test's documentation should be amended, supplemented, or revised to keep information for users current and to provide useful additional information or cautions (p. 70).	When a test is adapted for use in another population, documentation of the changes should be provided, along with evidence to support the equivalence of the adapted version of the test (p. 23).
When a test is translated from one language to another, the methods used in establishing the adequacy of the translation should be described, and empirical and logical evidence should be provided for score reliability and the validity of the translated test's score inferences for the uses intended in the linguistic groups to be tested (p. 99).	Test developers/publishers should ensure that the adaptation process takes full account of linguistic and cultural differences in the intended populations (p. 22.).
When multiple language versions of a test are intended to be comparable, test developers should report evidence of test comparability (p. 99).	Test developers/publishers should apply appropriate statistical techniques to (a) establish the equivalence of the language versions of the test, and (b) identify problematic components or aspects of the test that may be inadequate in one or more of the intended populations (p. 22).
When there is credible evidence of score comparability across regular and modified tests...no flag should be attached to a score. When such evidence is lacking, specific information about the nature of the modification should be provided...(p. 98).	Comparisons across populations can only be made at the level of invariance that has been established for the scale on which the scores are reported (p. 23).

*Adapted from Sireci (in press).

Quality Control Processes in Test Translation

The importance of the quality of the translation process for supporting comparative inferences across groups cannot be understated. For this reason, much attention has been paid to conducting proper test adaptations. Brislin, Hambleton, van der Vijver, and others have discussed the pros and cons of different models for adapting tests from one language to another (e.g., Brislin, 1970, Hambleton, 1994; Hambleton, Sireci, & Robin, 1999; Solano-Flores, Trumbull, & Nelson-Barber, 2002, Stansfield, 2003; van der Vijver & Tanzer, 1998). The models discussed in the literature include forward translation, back translation, forward translation with subsequent review and revision (iterative forward translation), and parallel test development.

The consensus of research in this area suggests that independent translators should be convened to adapt items across languages and to validate the translations. Back translation (Brislin, 1970) is also suggested as a further quality control check. In addition, many suggestions in this area

focus on the quality of the translators. For example, Hambleton and Kanjee (1995) stated that translators should be fully proficient in both languages of interest, familiar with the cultures associated with the different language groups, and have an understanding of the subject domain measured. Hambleton, Sireci, and Robin (1999) added the suggestion that translators also be proficient with respect to principles of good item writing. Solano-Flores et al. (2002) suggested that translation teams also be knowledgeable with respect to “cognitive and subject-area developmental levels” (p. 126).

After stressing the need for qualified translators, Hambleton and Patsula (1998) suggested that a rigorous instrument adaptation process would involve at least three steps:

1. translating the test from source to target language
2. translating the test back into the source language (back translation)
3. using independent teams of qualified translators to review the original, back-translated, and target language versions of the instrument to examine equivalence and resolve discrepancies

Literature in this area also stresses the need for *decentering*, which avoids literal word-for-word translations in favor of those that use different words but preserve the same meaning across languages (Brislin, 1970). As van de Vijver and Tanzer (1998) described, “An appropriate translation requires a balanced treatment of psychological, linguistic, and cultural considerations” (p. 266).

An alternate approach to translating a test from one language to another is simultaneous development of parallel versions of an exam in two or more languages. Solano-Flores et al. (2002) recommended that when simultaneous development is used, it is helpful to first develop item shells, which are highly constrained descriptions of the items to be written, and then develop items to fit those constraints in all the languages to be tested. They also pointed out that when tests are developed in one language and then translated into another language, only the original version of the test typically undergoes several rounds of review, piloting, and revision. An advantage of the simultaneous approach is that equal attention is paid to all language versions of the test throughout the development process. Although the simultaneous approach has not yet seen wide application, its merits make it an attractive option worthy of further study.

Quantitative Approaches to Evaluating Comparability

After care has been taken to adapt a test for use across different languages, statistical analyses are needed to evaluate the adaptation. *Standards for Educational and Psychological Testing* provides a framework for evaluating appropriate test use and interpretation that involves five “sources of evidence that might be used in evaluating a proposed interpretation of test scores for particular purposes” (AERA et al., 1999, p. 11). The sources are evidence based on

- test content
- response processes
- internal structure
- relations to other variables
- consequences of testing

With respect to the evaluation of translated versions of educational tests, the most common sources of validity evidence that have been investigated are test content and internal structure. Analysis of test content has involved statistical comparisons of original and translated versions of test items, followed by qualitative reviews of items flagged for potential translation problems. Analysis of internal structure has involved comparison of the factor structures of original and adapted tests. Sireci (in press) and van der Vijver (van de Vijver & Poortinga, 2005; van de Vijver & Tanzer, 1998) have pointed out that the factor structures of original and adapted tests should be evaluated first, before moving to analyses at the item level, because similarity of factor structure helps rule out potential method bias that would not be detected at the item level.

Evaluating Translated Tests Based on Internal Structure

Validity evidence based on internal test structure includes information regarding item homogeneity, such as estimates of internal consistency reliability, as well as dimensionality analyses that reflect the major dimensions needed to represent students' responses to the items and tasks making up an assessment. These techniques can be used to evaluate whether the intended dimensionality for an assessment is supported by empirical analysis of the data. With respect to test translations, studies in this area often involve an evaluation of whether the dimensionality of an assessment is consistent across its original and translated versions. Several studies have been done to compare the internal structure of standard and translated assessments. These studies, which are reviewed later, have primarily used exploratory and confirmatory factor analysis, or multidimensional scaling.

Evaluating Translated Tests Based on Differential Item Functioning

Validity evidence for translated assessments based on test content has involved statistical screening of items that may be potentially biased. This evidence has been in the form of *differential item functioning* (DIF) analysis. DIF analysis involves *matching* test takers from different groups on the characteristic measured and then looking for performance differences on an item.²² The logic behind this matching is that test takers from different groups who have similar proficiency (i.e., who are matched) should respond similarly to a given test item. If they do not, the item is said to “function differently” across groups. However, such differential functioning does not necessarily signify item bias. It merely flags the item for a statistical difference; it does not explain the cause of the difference. Item *bias* is present when an item has been statistically flagged for DIF *and the reason for the DIF is traced to a factor irrelevant to the construct the test is intended to measure*. Therefore, for item bias to exist, a characteristic of the item that is unfair (construct-irrelevant) to one or more groups must be identified. DIF is a necessary, but insufficient, condition for item bias.

²² Comprehensive descriptions of DIF methodologies can be found in Clauser and Mazor (1998) or Holland and Wainer (1993). It should be noted that DIF can be thought of as multidimensionality at the item level and so such analyses are also relevant to the consistency of the internal structure of an assessment across subpopulations.

DIF analyses have been used to evaluate test translations by identifying items that function differentially (i.e., are easier or more difficult) across their original and translated versions. Translated items that are flagged for DIF are sent to content review committees for adjudication. If it were judged that the DIF was due to a translation problem, the item is likely to be removed. The degree to which DIF is present on a translated assessment provides information regarding the comparability of the original and translated versions.

Linking Score Scales from Translated Assessments

In addition to evaluating test structure and item equivalence, methods have also been suggested for trying to achieve or approximate *scalar equivalence*, which is when scores from different tests are on a common scale (van de Vijver, & Tanzer, 1998). Mislevy (1992), Linn (1993), and others (e.g., Cascallar & Dorans, 2005; Dorans, 2004; Kolen & Brennan, 2004) have discussed the different degrees to which different tests may approximate scalar equivalence. This area is presently referred to as “linking,” which is a general term to describe the various ways in which scores on different tests, or on different forms of the same test, may be related to one another. As Linn (1993) described, “linking is a generic term that includes a variety of approaches to make results of one assessment comparable to those of another” (p. 84).

Mislevy (1992) and Linn (1993) described five levels of linking that vary according to the method used to make scores from different tests comparable and the assumptions regarding similarity of the assessments to be linked. This taxonomy of linking methods is relevant to the comparability of scores from translated tests and so a brief summary is provided here.

Equating is the strongest form of linking in that when tests are equated, examinees are indifferent to which form they would take, because they would get essentially the same score on either test, within the expectations of measurement error. Equating requires that the same construct is measured and the different tests are developed from the same test specifications.²³ Statistical methods for equating tests have strict data collection designs requiring common groups, randomly equivalent groups, or common items. Translated versions of educational tests typically involve the same construct and content specifications, but the items cannot be considered “common” after they have been translated, nor are the groups of examinees who take different language versions of an assessment typically considered randomly equivalent. For this reason, it is not considered possible to strictly equate translated tests (Sireci, 1997, 2005).

²³ Lord (1980) stated that for tests to be equitable, they must measure the same construct with equal reliability, symmetry, equity, and population invariance. The equal construct, equal reliability, and population invariance conditions are satisfied when tests to be equated measure the same construct with equal reliability and are related in the same way across different subpopulations. The equity condition holds if, after equating, scores on the tests can be used interchangeably such that it should not matter which test an examinee chooses to take. The condition of symmetry requires that equating be the same regardless of which test is taken. Therefore, symmetry holds if the equating function for transforming scores on Test A to scores on Test B is an inverse of the function for transforming scores from B to A.

The next level of linking is *calibration*. Calibration also assumes the tests to be linked are measuring the same construct and are targeted to the same content specifications. However, they are not assumed to be measuring the construct with the same precision. Calibration involves linking scores on tests that are intended to measure the same construct, but with different levels of reliability or difficulty. Linking scores on a short version of a test to scores on a longer version is an example of calibration (Linn, 1993). This form of linking also does not describe the typical situation in test translation, because common items or persons (who can take both assessments) are needed to conduct the calibration.

Statistical moderation involves the use of an external measure to link scores from different assessments. An example given by Linn (1993) is the use of scores from a standardized test to adjust the scores from performance assessments that might be administered at the local level. In this example, the mean and standard deviation for each school on the local assessments would be moderated to have the same mean and standard deviation as the local sample of students had on the standardized assessment. With respect to test translation, Wainer (1993) argues that scores from different language assessments could be statistically moderated through an external validity criterion. However, such criteria are elusive and this model has not yet seen application on translated educational tests.

In *prediction*, the scores on one test are used to predict the scores on another test. The predictive relationship is typically sample-dependent and is not reciprocal (i.e., a different prediction equation is used to predict X from Y than is used to predict Y from X). With translated tests, it is typically not possible to have students take both forms of a test and so this form of linking is rare on translated tests (but see Sireci, 2005 for a theoretical example using bilingual test takers, and see Boldt, 1969; CTB, 1988; and Cascallar & Dorans, 2005 [described in a subsequent section], for empirical examples).

The last form of linking, called *social moderation*, is essentially judgmental rather than statistical. Linking scores from different assessments (or from different groups who score different students taking the same assessment) is accomplished through the use of a common rubric and process for deriving the scores. That is, the scores from different assessments are not linked in any formal, statistical sense. Instead, they are scored using a common process of how scores should be assigned to students who operate at different levels of proficiency. Davis, Buckendahl, and Plake (2006, described in a subsequent section) provide an example of how this was applied in a dual-language testing context. In this example, common standard setting procedures were used to set a presumably common passing standard across different language versions of reading and writing assessments.

The taxonomy of linking approaches described by Mislevy (1992) and Linn (1993) is helpful for understanding the degree to which scores from different test forms may be considered comparable. With respect to translated tests, since translated items cannot be considered identical (and so used to anchor the score scales), strict equating and calibration are not possible. Statistical moderation is theoretically possible when a valid, language independent, criterion is available, but of course such criteria are rare, if they exist at all. The weakest form of linking, social moderation, is certainly applicable, but the evidence of score comparability it provides may not be enough to defend score comparability. For these reasons, research has primarily

focused on evaluating and comparing the psychometric properties of different language versions of an assessment to support various degrees of score comparability. We turn now to reviews of the empirical work in this area.

Reviews of Empirical Evaluations of Translated Educational Assessments

In previous sections, we provided very brief descriptions of how analyses of factor structure and DIF can inform an evaluation of the validity of inferences derived from scores from translated assessments, and how linking procedures might be used to facilitate score comparability. In this section, we summarize the empirical literature in this area that has involved evaluation of the comparability of original and translated versions of educational assessments. Empirical research on translated versions of state-mandated tests is rare, and there are very few examples in the published literature. Therefore, to illustrate some of the methods that can be applied to this problem, studies that involve international assessments and admissions tests are also reviewed. A list of the studies reviewed is presented in Table 2, along with very brief descriptions of the testing context, validity evidence analyzed, statistical methods applied, and summary of the results.

As indicated in Table 2, most studies of translated assessments have investigated consistency of factor structure (structural equivalence) or DIF, and many studies have investigated both. Interestingly, the two studies conducted on state-mandated achievement tests in the U.S. both focused on dual-language versions of these tests, and both are unpublished (Anderson et al., 2000; Sireci & Khaliq, 2002). We begin with the evaluation of dual-language tests because these studies represent the only empirical analyses of state-mandated educational assessments that we could find, and they illustrate methods that can be used to evaluate tests adapted for use across languages.

Empirical Evaluations of Dual-language Tests

Three studies have looked at the value of dual-language tests and the comparability of these tests to the original English language version. All three studies involved test booklets that included items in both the English and Spanish languages. The first two studies focused on state-mandated tests (Anderson et al., 2000; Sireci & Khaliq, 2002), and the third involved a pilot study of a dual-language version of a math test from the National Assessment of Educational Progress (NAEP; Duncan et al., 2005).

Table 2. Empirical Analyses of Translated Educational Assessments

Citation	Context	Validity Evidence	Statistical Methods	Findings
Sireci & Khaliq (2002)	dual-language state math test	internal structure, DIF	SIBTEST, EFA, CFA, MDS	Some differences in test structure. Attributed to both DIF and non-overlap of score distributions
Anderson et al. (2000)	state reading test with passages in English, but directions and test items in both English and Spanish	experimental design	ANOVA	No statistically significant difference between ELLs who took dual-language version of assessment and those who took English language version.
Duncan et al. (2005)	pilot dual-language (English/Spanish) version of NAEP math test	analysis of group differences	ANOVA	Little empirical support for dual-language booklet, but ELLs with low English proficiency appreciated the dual-language version and responded to the Spanish versions of the items.
Allalouf, Hambleton, & Sireci (1999)	Hebrew and Russian versions of the verbal reasoning subset from a college admissions test used in Israel	test content, DIF	Mantel-Haenszel	DIF was found primarily on analogy & sentence completion items and was explained by changes in difficulty due to translation (e.g., word difficulty), item format, or cultural relevance.
Gierl & Khaliq (2001)	English and French versions of math and social studies tests in Alberta, Canada	test content, DIF	SIBTEST	Bilingual translators and content specialists identified causes of DIF that were confirmed by content and statistical analyses on a similar test.
Sireci & Gonzalez (2003)	TIMSS Science test for 13 year olds: 9 countries	internal structure	EFA, MDS	Slight differences in test structure for some countries/languages. Differences were related to item difficulty.
Ercikan & Koh (2005)	English and French versions of TIMSS math and science tests	internal structure, DIF	CFA, IRT DIF analyses	Structure of the assessments was not consistent across languages in some countries and substantial DIF was found.
Cascallar & Dorans (2005)	English-Spanish bilinguals who took SAT, PAA, and ESLAT	prediction, concordance	Multiple regression	Bilinguals can be used to compute predicted scores on SAT from PAA and ESLAT. Method has several limitations.
Davis, Buckendahl, & Plake (2006)	reading and writing assessment for high school students in Canada	test content	NA (standard setting)	Setting standards on each version of the exam simultaneously using bilingual translators and facilitators helped ensure consistent standard setting processes across exams.
Baxter et al. (2007)	NAEP Trial PR assessment in math	item fit	IRT residual	30% of items misfit and needed to be excluded, but still claimed PR results could be reported for NAEP.

Anderson, Liu, Swierzbina, Thurlow, and Bielinski (2000) evaluated the accommodation of providing dual-language test booklets on a reading test. The dual-language booklets presented all reading passages in English, but all other test information, including directions, items, and response options, were written in both English and Spanish and presented side-by-side. The directions, items, and response options were also presented aurally in the native language on a cassette tape. The dual-language assessment was created from unused items and reading passages associated with the Minnesota Basic Standards Reading Test. The participants were 206 eighth-grade students from two consecutive eighth-grade classes from five schools in Minnesota. They were separated into three test groups: an accommodated ELL group (n=53), a non-accommodated ELL group (n=52), and a control group of general education students (n=101). Most of the ELLs were randomly assigned to the standard or dual-language condition. A focus of the study was whether ELLs who took the dual-language version of the assessment would perform better than ELLs who took the English-only version.

The only statistically significant finding was that the non-ELL group outperformed the standard and dual-language ELL groups. Although there was no statistically significant difference for ELL students across the standard and dual-language conditions, the trend was in the expected direction with the dual-language group having a higher mean score (about 3/10 of a standard deviation higher) and it was estimated that about twice as many students in this condition would pass the test relative to the standard condition (9 percent versus 4 percent). In addition, Anderson et al. found that students tended to primarily use one version of the written test questions (either English or Spanish) and then refer to the other version when they felt it would be helpful. They concluded that most students used the Spanish translations as an occasional reference or when they encountered difficult terminology. Students made little use of the oral presentation of the test questions in Spanish. However, when asked whether they would want the written Spanish translations if they took the test again, about two-thirds of the students responded affirmatively.

Sireci and Khaliq (2002) also evaluated an English-Spanish dual-language version of a statewide assessment, but unlike Anderson et al. (2000), the focus of their study was on the comparability of the dual-language assessment to the English version. Their study involved analysis of data from an operational administration of a state-mandated fourth-grade math test. To be eligible to take the English-Spanish version of this test, ELL students must have been enrolled in U.S. schools for less than three years, be slated for Spanish language instruction the following year, and read and write at or near grade level in Spanish. The dual-language version of the test contained the English version of the test questions on the right-facing pages with the Spanish translation of the same test questions on the left-facing page. Any extra testing material was provided in both languages and test administrators were provided with test instructions in both languages.

The statistical analyses focused on evaluating the consistency of the factor structure across the English and dual-language versions of the test and evaluation of DIF. For the DIF analyses, test booklet, rather than ELL status, was used as the grouping variable. Thus, it was assumed that students who took the dual-language version were essentially responding to the Spanish versions of the items.

There were 76,783 examinees who completed the English form of the test and 585 students who completed the dual-language form. Given that such a vast difference in sample size could conceal meaningful group differences or exaggerate trivial differences, Sireci and Khaliq randomly selected three groups of 585 students from the population of students who took the English version. After noting substantial differences in the distribution of test scores for these groups, they also selected three stratified random samples from this population, where the stratification was done so that the distributions of total test scores were essentially identical to the total test score distribution of the dual-language sample. Thus, they were able to disentangle differences potentially due to test form (English or dual-language) from differences due to non-overlap of the proficiency distributions.

Both structural analyses and analyses of DIF were conducted. The structural analyses involved exploratory factor analysis and multidimensional scaling (MDS). These analyses revealed that different dimensions were needed to account for the structure of the students' item response data across the two versions of the test. The differences in structure were reduced when the stratified random (i.e., matched) samples were used, but they were still evident.

These findings regarding test structure are illustrated in Figures 1 and 2, which reproduce the MDS results for the random and stratified analyses, respectively. The four data points displayed in Figure 1 are the dimension weights for four groups of students. These weights represent the relative importance of each dimension for accounting for the students' responses to the test items. Essentially, a higher weight on a dimension for a group means that dimension accounted for the difficulties of the items associated with that dimension.²⁴ In Figure 1, the three points clustered together represent the random samples of students who took the English version of the test. The three points clustered together in Figure 2 represent the stratified (matched) random samples of students who took the English version of the test. In both figures, the ELL group has a noticeably larger weight on dimension 2, and a noticeably smaller weight on dimension 1, relative to the non-ELL groups, which suggests the dimensions differentially account for the structure in each test form. However, the difference between the ELL and non-ELL groups is smaller in the analysis involving the matched groups. These results illustrate a general lack of factorial invariance across these two test forms, but it appears the difference is exacerbated by general differences in proficiency between the groups.

Sireci and Khaliq also found a relationship between the lack of structural equivalence and presence of DIF. Specifically, when items flagged for DIF were removed from the analysis, the dimensional structures became more similar across the English and dual-language groups. In addition to providing information about this specific test, this study also illustrates how MDS can provide a visual display of the similarity of factor structure across different-language versions of an assessment.

²⁴ Multi-group MDS analyses display the structure of data simultaneously, but reveal structural differences across groups in the form of weights for each group on each dimension. The larger the group weight on a dimension, the more relevant the dimension for accounting for the structure of the item response data for the group. See Davison and Sireci (2000) for a more complete description of MDS.

The third study that evaluated an English-Spanish version of an assessment was Duncan et al. (2005).²⁵ In this study, Spanish-speaking ELL students were randomly assigned to an experimental dual-language or English language eighth-grade NAEP math test to evaluate the psychometric equivalence of the dual-language and monolingual versions of the test. The study involved analysis of test data, as well as focus groups and cognitive (think-aloud) analysis of small numbers of students from each group.

Like the Anderson et al. (2000) study, Duncan et al. (2005) compared three groups—Spanish ELLs who were educated in the U.S. for less than three years and were not proficient in English, Spanish ELLs of moderate English proficiency who were educated in the U.S. for at least three years, and native English speakers. The test comprised 60 items—45 multiple-choice and 15 constructed-response items. The sample sizes obtained in the study²⁶ were too small for formal analysis of DIF or structural equivalence. Instead, the focus was on mean score differences across the groups defined by ELL and test form.

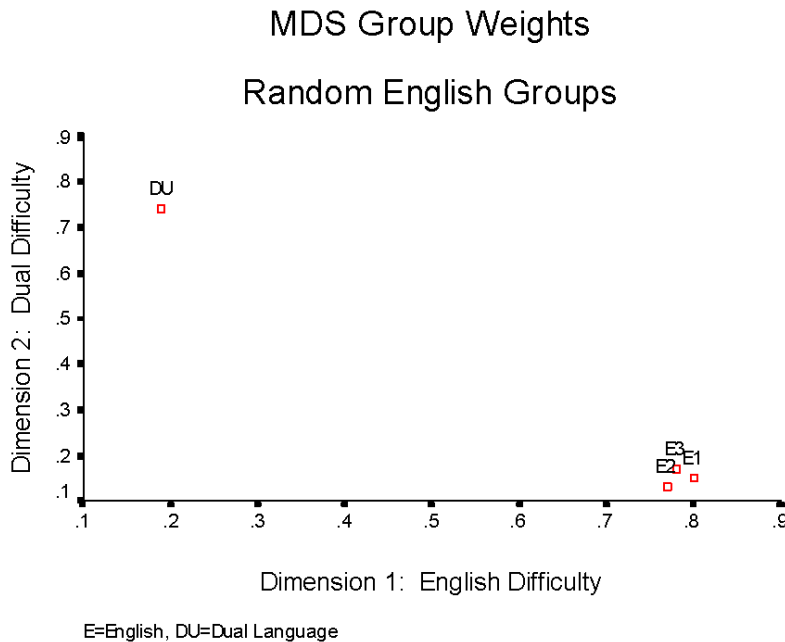


Figure 1. MDS Dimension Weights for Dual-language and English Test Samples (from Sireci & Khaliq, 2002)

²⁵ Duncan (2005) is the peer-reviewed journal publication of a more comprehensive report by Garcia et al. (2000).

²⁶ Sample sizes were n=127 for ELLs with less than three years of education in U.S. who took the dual-language test, n=74 for ELLs with at least three years of education in the U.S. who took the dual-language version, n=144 ELLs with at least three years of education in the U.S. who took the English language version, and n=144 native English speakers who took the English language version.

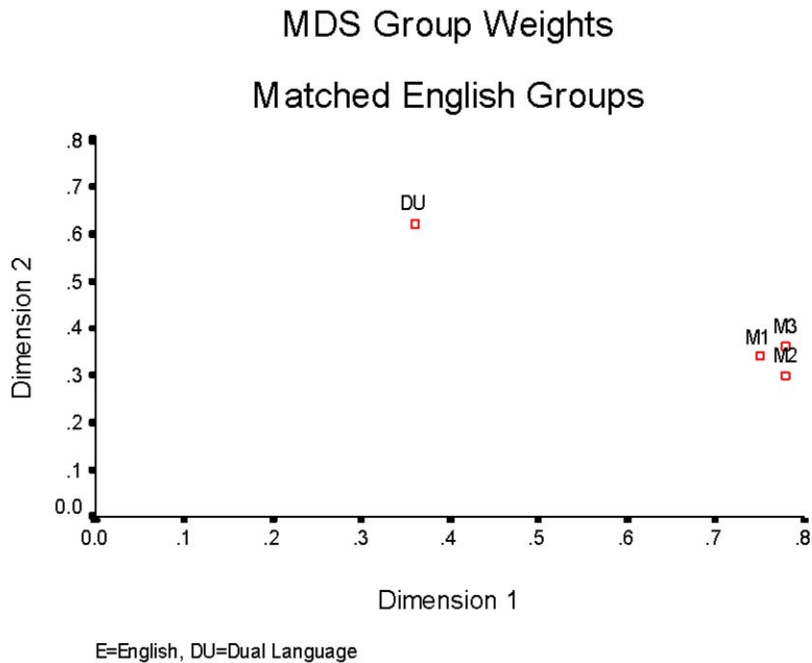


Figure 2. MDS Dimension Weights for Dual-language and English Test Samples Matched on Total Test Score (from Sireci & Khaliq, 2002)

Although a difference was found across the ELL and non-ELL group, and across ELLs with less than three years of education in the U.S. and ELLs with more than three years of education in the U.S., there was no difference between the dual-language and English language versions of the test for the students who could take either version (i.e., ELLs with more than 3 years of education in the U.S.). Duncan et al. interpreted these results as indicating that criteria other than education in the U.S. for more than three years should be used to determine which types of ELLs are most likely to benefit from dual-language test administrations.

Through the focus groups and cognitive interviews, Duncan et al. found that Spanish-speaking ELLs appreciated the dual-language booklets, with 85 percent of the students finding the booklets “useful” or “very useful.” However, they also found that students who had three years or less instruction in English predominantly read and responded only to the Spanish versions of the test items. They also found that ELLs with high levels of English proficiency scored slightly lower when taking the dual-language version of the test. These results suggest that the use of dual-language booklets is promising, but more research is needed to determine its utility.

As Duncan et al. (2005) noted, a limitation of their study was that they were unable to statistically evaluate the psychometric comparability of the dual-language and English booklets due to the small numbers of students who took the dual-language version. However, they concluded that the focus groups and cognitive interviews were helpful for confirming the quality of the translations.

Comparisons of Translated Assessments in Other Contexts

The Anderson et al. (2000) and Sireci and Khaliq (2002) were the only studies we found in the literature that focused on empirical analysis of translated versions of statewide assessments. However, there have been several published studies of translated versions of other educational tests. In this section, we review seven of these studies. These seven studies illustrate statistical methods that can be used to evaluate translated versions of tests and illustrate the general findings regarding the comparability of different-language versions of educational assessments. The studies reviewed are Baxter et al. (2007); Allalouf, Hambleton, and Sireci (1999); Gierl (2000); Gierl and Khaliq (2001); Sireci and Gonzalez (2003); Davis et al. (2006); and Ercikan and Koh (2005).

NAEP Trial Mathematics Assessments in Puerto Rico

There have been three trial mathematics assessments in Puerto Rico in grades four and eight as part of the National Assessment of Educational Progress (NAEP). The first was in 2003, the second in 2005, and the third just recently completed in 2007. Based on these trials, it is expected that Puerto Rico will be fully integrated into the 2009 NAEP assessment (Baxter et al., 2007).

The assessments in Puerto Rico were translated versions of the assessments in the U.S. that were administered in English. Baxter et al. (2007) describe the technical characteristics of the 2005 trial assessment and also discuss the characteristics of the 2003 assessment. The 2005 trial assessment featured an “enhanced translation” (p. 1) involving independent translators and English-Spanish bilingual educators who evaluated the translations. They also noted the results from the 2003 trial assessment were deemed unusable because of the large rate of students who did not respond to items and the large proportion of items that did not fit the item response theory (IRT) models. For example, in grade four in 2003, 25 percent of the item response data for Puerto Rico were missing, compared with only 7 percent in the U.S. For the extended constructed-response items, 59 percent of the data were missing in Puerto Rico, compared with 22 percent in the U.S. Revised test administration procedures were implemented in 2005 to address the nonresponse rates. These changes included adding 10 more minutes per test section, the aforementioned enhanced translation procedures, and revised directions for students.

To evaluate the 2005 trial assessment in Puerto Rico, Baxter et al. (2007) looked at missing response rates, item misfit, and DIF.²⁷ The percentages of missing responses in 2005 were lower (14 percent for grade four, 18 percent for grade eight), but were still generally double those seen in the U.S. With respect to items flagged for significant DIF or misfit, 30 percent of the items were identified. The authors concluded,

²⁷ Interestingly, the DIF analyses were conducted using the U.S., DC, Virgin Islands, American Samoa, and Puerto Rican students in the U.S. as reference groups. Unfortunately, details regarding the DIF results were not included in the report. The only finding reported was “Results indicated that items flagged as having significant DIF did not account for most or all of the item misfit” (p. 15). Nevertheless, the use of multiple reference groups of similar proficiency or language background is an interesting idea that deserves further study.

The high levels of item misfit in Puerto Rico may indicate curricular differences between Puerto Rico and the nation, or translation errors, or an assessment that is too difficult, or some combination of these issues. . . . Some of the items may have been at a different level of difficulty for students in Puerto Rico than for students in the nation. A battery of assessment items that is more difficult than the students' levels of proficiency can increase item nonresponse, guessing, and confound proficiency with speed. (p. 12)

Although the results suggested a lack of comparability of the English and Spanish versions of these NAEP assessments, the Baxter et al. (2007) study illustrates interesting methodology that can be used to evaluate translated assessments that are developed using IRT. For example, IRT model-fit (residual) analyses provide important information regarding item functioning across languages when the tests are scaled using IRT and the sample sizes are sufficient (sample sizes were about 3,000 for these trial assessments).

In addition to evaluating the statistical functioning of the items and nonresponse, Baxter et al. (2007) recomputed scale scores for Puerto Rico, the nation, and the other states and jurisdictions after eliminating the 30 percent of items that were flagged as problematic. They found that Puerto Rico's average scale scores increased 2–3 points, but their relative rank compared with other jurisdictions did not change. This finding led them to conclude,

Based on these analyses, it is the conclusion of NCES that the full set of NAEP items can and should be used for reporting the Puerto Rico results for both the 2003 and 2005 assessments. The mean scale score for Puerto Rico might be 2 to 3 points higher than reported, but this difference would not change Puerto Rico's ranking among other participating jurisdictions. Similarly for the achievement level results, the percentages of students at or above a particular achievement level may differ for the two scales by 1 to 2 percentage points. (p. 21)

They defended this conclusion in part by noting that the representation of the NAEP assessment, with respect to the proportions of items measuring each content area and representing each item format, was relatively similar before and after dropping problematic items. They summarized their findings by stating,

In sum, many of the NAEP items functioned differently in Puerto Rico than predicted. Item misfit primarily affects aggregate statistics by reducing the precision of the estimates. Item misfit may signal problems with the assessment and potentially biased results. However, the results may be an accurate reflection of the proficiency of the group being assessed. (p. 13)

The conclusions of Baxter et al. that scores from the 2005 trial NAEP Mathematics Assessment in Puerto Rico can be used in reporting and interpreting NAEP results indicates that scores from translated assessments can be useful, even when there is evidence the scores are not directly comparable to those from the original language version. Such a claim reminds us that when evaluating the validity of scores from translated assessments, we need to keep the purpose of the assessment and how scores are interpreted in mind. With respect to NAEP, there are no scores

reported for students nor any stakes at the student level. These factors help us understand the conclusions of Baxter et al., but the results also suggest the full set of items, and perhaps the entire assessment, may have limited Puerto Rican students' ability to demonstrate their true mathematics proficiencies.

Regardless of how the results of the Baxter et al. study are interpreted, its relevance to this paper is primarily methodological. The use of IRT misfit analysis and analysis of student nonresponse are useful ways for evaluating score comparability. The idea of reporting results for students who take translated assessments based only on items that functioned similarly to the original language versions is also interesting, and could be considered in statewide assessments where translated tests are used.

College Admissions Testing in Israel

Allalouf et al. (1999) summarized the results of studies conducted on the Hebrew and Russian versions of the verbal reasoning subtest of the Psychometric Entrance Test (PET), which is the primary admissions test used by colleges and universities in Israel. This test is developed in Hebrew and translated into Arabic, Russian, French, Spanish, and English. The study used DIF analyses and focus groups of bilingual content specialists and translators to derive reasons why items were thought to function differentially across languages. They noted that previous research using MDS confirmed that the factor structures of the Hebrew and Russian versions of the exam were similar, but that analyses of DIF had found that several analogy and sentence completion items were not statistically equivalent across languages. Specifically, they found that the Russian versions of the analogy items flagged for DIF tended to be differentially easier in Russian, but that the sentence completion items flagged for DIF were inconsistent in the direction of DIF.

To better understand the potential causes of DIF, Allalouf et al. convened a group of eight Hebrew-Russian bilingual content specialists and translators to hypothesize and discuss reasons why the items flagged for DIF were inconsistent across languages. Based on their analyses and discussions, four causes of DIF across these languages were identified:

- differences in the familiarity of specific words across languages due to frequency of usage (e.g., a difficult word in Hebrew translated into a trivial word in Russian)
- changes in the content of an item due to translation (incorrect translation or a word with a single meaning in Hebrew translated into a word with more than one meaning in Russian)
- changes in the format or appearance of an item (e.g., the stem was much longer in one language relative to the other)
- differences in the cultural relevance of an item

This study illustrated how statistical analyses of DIF can be followed up by qualitative analyses of test items to help interpret differences in test performance across examinees taking different-language versions of a test as well as the degree of comparability of the tests themselves. The results of the study are also useful for informing future test translation efforts.

Understanding Translation Differences on Achievement Tests in Canada

Gierl and Khaliq (2001) extended the Allalouf et al. study by first using focus groups to review DIF items to derive potential causes of DIF and then using these causes as hypotheses in the evaluation of subsequent different language versions of an assessment. Their study involved analysis of English and French versions of math and social tests administered in sixth and ninth grades in Alberta, Canada. There were three stages to their study. The first stage involved using a simultaneous item bias procedure (SIBTEST, Shealy, & Stout, 1993) to flag items for DIF across the two languages. The second stage involved convening a group of 11 bilingual content specialists to review the English and French versions of the items flagged for DIF and come up with consensus opinions regarding the likely sources of the DIF. The third stage involved a subsequent team of two translators to use the sources of DIF identified by the content specialists to categorize items on a subsequent assessment that were flagged for DIF into one of the source categories put forward by the previous committee.

The results of their study illustrated how an iterative DIF screening process could be used to identify items that function differentially across different-language versions and how the sources of DIF identified by bilingual content specialists could be used to explain subsequent items flagged for DIF. In fact, the sources of DIF identified by the content specialists were similar to those identified by the specialists in the Allalouf et al. (1999) study, even though the languages involved were very different. Furthermore, Gierl and Khaliq illustrated that sources of DIF could be used to evaluate the aggregate effect of translation and cultural relevance differences across items on students' test scores. That is, they found the substantive interpretations of DIF to hold up over subsequent test forms and to affect score differences across the English and French versions of the assessments. They concluded, "The next step is to develop more refined and detailed [differential test functioning] hypotheses where researchers identify content differences and predict group differences (as was done in the current study) but also study the content-by-group interactions" (p. 183).

Evaluating Translations on International Assessments

The Gierl and Khaliq study was similar to a study of different-language versions of a statewide assessment in the U.S., since the test data analyzed came from a specific province in Canada. However, other studies evaluating the comparability of different-language versions of tests have been done in the context of international assessments. Two of these studies involved analysis of different-language versions of tests associated with the Trends in International Math and Science Studies (TIMSS).

Sireci and Gonzalez (2003) investigated the consistency of the factor structure of several different-language versions of the 1999 TIMSS science assessment. Data were analyzed from Belgium, Canada, England, Hong Kong, Italy, Japan, Korea, Russia, and the United States. Both English and French language versions of the test were used in Canada. Thus, the analyses involved comparing the factor structure of the science assessment across 10 language groups: Flemish (Belgium), Canadian-English, Canadian-French, English in England, English in the United States, Chinese (Hong Kong), Italian, Japanese, Korean, and Russian. The sample sizes for these groups ranged from 2,437 (French Canadian) to 9,072 (United States). Like Sireci and

Khaliq (2003), they used both exploratory factor analysis and MDS to evaluate the equivalence of test structure across the different languages and cultural groups.

The exploratory factor analyses indicated a consistent, but somewhat weak, general factor across all language groups as well as several secondary factors that were difficult to interpret. The MDS analyses were more readily interpretable and revealed subtle differences in the factor structures across some of the groups due to differences in the item difficulties across languages.

The MDS results indicated that if a very general, two-dimensional solution were fit to the data, the two dimensions were similar across all 10 groups. However, a five-dimensional solution fit the data better, and revealed interesting but subtle differences in test structure across groups. In the five-dimensional solution, all 10 groups had weights above zero in the first dimension, but Japan, Korea, and Belgium had larger weights on the second dimension. There was a positive relationship between the group weights and the country-specific item difficulty estimates. The correlations between the item difficulties for each group and the dimension on which it had the largest weight suggested that the differences in dimensionality across the language groups was associated with differences in the rank ordering of the item difficulties across groups.

The group weights and the correlations among the MDS item coordinates and the group-specific item difficulties are presented in Table 3. A similar pattern of weights and correlations is evident for the English versions of the assessment (i.e., Canadian English, England, and U.S.) and some countries have idiosyncratically large weights on a single dimension (i.e., Hong Kong on Dimension 2, Japan on Dimension 5, Russia on Dimension 4, and Belgium on Dimension 3). Across the five dimensions, the correlations between the difficulty/coordinate correlations and the group weights ranged from .85 (Dimension 3) to .99 (Dimension 2). It is interesting to note that the Canadian English and Canadian French versions of the exam have very similar patterns of MDS weights and coordinate/item difficulty correlations.

Table 3. WMDS Weight Matrix for MDS Solution (from Sireci & Gonzalez, 2003)

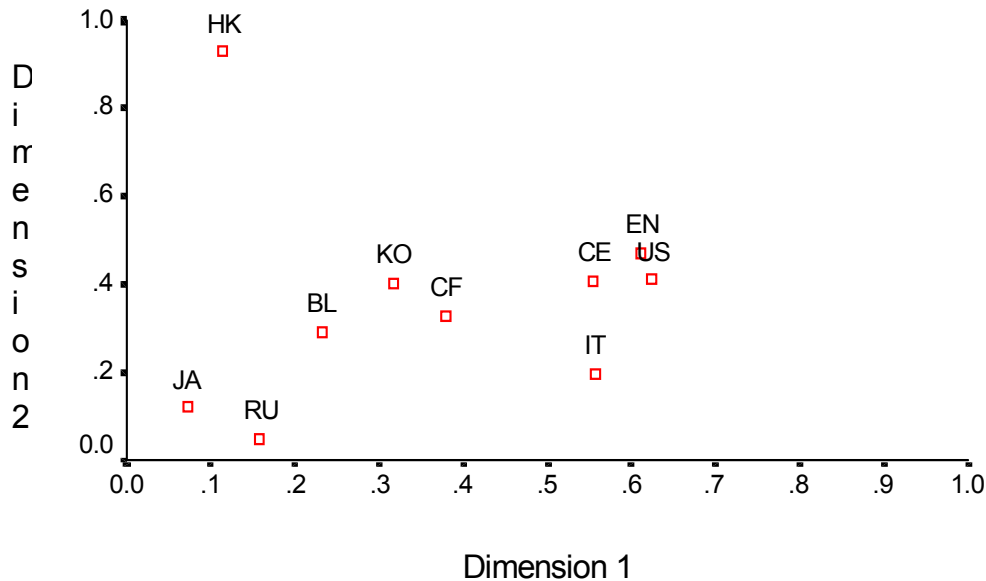
Group	Dimension									
	1		2		3		4		5	
	W	rb,c	W	rb,c	W	rb,c	W	rb,c	W	rb,c
United States	.62	-.77	.41	-.66	.23	-.42	.27	.58	.16	.50
England	.61	-.74	.47	-.70	.23	-.45	.15	.55	.24	.55
Canadian English	.55	-.71	.41	-.66	.34	-.50	.27	.61	.30	.58
Italy	.56	-.69	.20	-.50	.43	-.60	.29	.66	.09	.50
Canadian French	.38	-.61	.33	-.61	.40	-.50	.29	.61	.30	.56
Korea	.32	-.48	.40	-.63	.41	-.41	.27	.56	.23	.53
Belgium	.23	-.48	.29	-.58	.68	-.72	.27	.57	.23	.62
Russia	.16	-.42	.05	-.44	.12	-.46	.89	.89	.09	.47
Hong Kong	.11	-.44	.93	-.89	.05	-.33	.10	.41	.10	.50
Japan	.07	-.34	.12	-.48	.11	-.44	.08	.36	.91	.86
% VAF	.17		.18		.12		.13		.12	

W=weight for group on dimension

rb,c=correlation between IRT difficulty estimates and MDS coordinates

VAF= % of variance in item dissimilarities accounted for by coordinates

A two-dimensional sub-space of the five-dimensional MDS weight space is presented in Figure 3. This figure illustrates the importance of Dimension 2 to the Hong Kong data and the relative importance of Dimension 1 to the data from the English and Italian language versions of the test. However, it should be noted that these dimensions are “smaller” than those from the two-dimensional solution in that the inter-item variance accounted for by each dimension ranged 12–18 percent (see Table 3).



BL=Belgium CE=Can. Eng, CF=Can.Fr, EN=England, HK=Hong Kong
 IT=Italy, JA=Japan, KO=Korea, RU=Russia, US=USA

Figure 3. MDS Dimension Weights for TIMSS 1999 Science Assessment (from Sireci & Gonzalez, 2003)

In general, the MDS results suggested that the gross (two-dimensional) structure of the data is fairly consistent across groups, and the minor multidimensionality is related to differences in item difficulty across groups; and as more dimensions are fit to the data, more subtle structural differences are revealed and these differences also stem from differences in item difficulty across groups. Given the similar pattern of dimension weights across the three groups that took the English language version of the exam, item differences introduced through the test adaptation process should be investigated as a source of multidimensionality. However, instructional practices may also be similar across these countries, which could also be linked to multidimensionality. Most importantly, this study indicates how the consistency of the general and subtle aspects of test structure can be evaluated across multiple translated versions of an assessment via MDS analysis.

Ercikan and Koh (2005) also evaluated the comparability of different language versions of TIMSS assessments, but there were several differences from the Sireci and Gonzalez (2003) study. First, Ercikan and Koh looked at both math and science versions of TIMSS exams, but the data were from an older (1995) version of the assessment. Second, they looked at English and French versions of the exams administered in Canada, England, France, and the United States. Third, they looked at both DIF and structural equivalence. DIF was evaluated by looking at the consistency of item response theory (IRT) parameters estimated separately for each language group and by using the Linn-Harnisch procedure. Structural equivalence was evaluated using multi-group confirmatory factor analysis (CFA).

Their results indicated a lack of equivalence at both the structural and item levels. For example, they found substantial levels of DIF in some comparisons (e.g., 59 percent of the math items were flagged for DIF across England and France, 79 percent of the science items were flagged for DIF across France and the U.S.). The global fit indices associated with the CFA illustrated relatively worse fit of the models to the data in those situations where the greatest amount of DIF was observed. Thus, like, Sireci and Khaliq (2002), this study illustrates how DIF and structural analyses can be used in complementary fashion to evaluate the comparability of translated assessments. Ercikan and Koh warned that when substantial amounts of DIF and inconsistencies in test structure are observed across translated assessment instruments, comparisons of students who responded to different language versions of the items should not be made.

Linking Scores across Translated Tests Using Bilingual Students

Cascallar and Dorans (2005) evaluated the degree to which scores from a Spanish language version of the SAT used in Puerto Rico (the Prueba de Aptitud Academica [PAA]) could predict SAT performance. To accomplish that goal, they used a sample of Spanish-English bilingual Puerto Rican students who took both tests, as well as the English as a Second Language Achievement Test (ESLAT). The ESLAT was used to screen out students with low English proficiency and to serve as an additional variable for predicting SAT scores. They used equipercentile equating to establish concordance between the SAT and PAA, and multiple regression to predict math, verbal, and composite SAT scores. They found the prediction approach via multiple regression preferable to equating, due in large part to the contribution of the ESLAT in the prediction. They concluded that the PAA and ESLAT could be used to predict how well students in Puerto Rico might do in colleges in the U.S. This study illustrated how bilinguals could be used to evaluate score comparability, but it should be noted the PAA was not a translation of the SAT and so linkage of the two tests was the focus of the study, rather than evaluating score comparability.

A Social Moderation Approach to Score Equivalence

Davis et al. (2006) described a study in which pass/fail standards needed to be set on English and French versions of high school reading and writing tests. To set the standards, they convened separate panels of English and French reading and writing experts (teachers); however, they conducted the training simultaneously, using both an English-speaking and a bilingual (English-French) facilitator. The orientation and training was done first in English (with a French translation via headphones) and then in French (with an English translation via headphones). In

this way, the groups were introduced to the concepts of “borderline students” and the rating tasks (Angoff and Analytical Judgment methods for the reading and writing tests, respectively).

Following the common orientation and training, the groups were split into their language-specific panels and the same process was used to derive passing scores on each language version of each test. The differences between the standards set on each exam resulted in about 1–6 percent differences in the passing rate for each group of students, which was deemed acceptable by the authors. This study illustrates that parallel standard setting processes could be used to set defensible standards on different language versions of an assessment. However, the utility of setting the standards simultaneously deserves further study.

Discussion and Conclusions

In this report, we

- reviewed validity issues involved in test translations
- discussed methods for evaluating the comparability of different language versions of an assessment
- summarized the results of several studies that evaluated different language versions of educational assessments

From these discussions we can conclude that gathering evidence regarding the comparability of translated versions of educational assessments is important, and that statistical methods are available for evaluating the comparability of translated tests after they have been administered to groups of students who differ with respect to native language. With respect to the results of the empirical studies conducted thus far, it cannot be concluded that scores from these assessments can be considered comparable. In some cases, the lack of comparability can be traced to translation problems at the item level.

There are several interesting methodological findings from this review that may be useful for future research investigating the comparability of tests adapted for use across languages. First, two studies indicated the utility of following up statistical analyses of cross-lingual DIF by convening panels of bilingual content specialists. Using such specialists to identify and isolate sources of DIF across language versions will allow us to better ascertain the degree to which scores from these versions are comparable. Another interesting methodological finding was that analyses of test structure that used matched samples of examinees revealed more similar internal structures across original and adapted tests. Future research may want to use this strategy to distinguish between inconsistency in test structure due to translation problems and inconsistency due to a lack of overlap of the proficiency distributions across groups. The idea of conducting IRT fit analyses and recomputing scores after eliminating problematic items (Baxter et al., 2007) is also interesting and deserves further study.

One finding that was clear from this review is that the practice of translating tests is far outpacing research that evaluates the appropriateness of this practice. Increased use of translated educational tests is understandable for at least two reasons. First, the use of translated versions of educational tests promotes access to the assessment for students who are not proficient in the language in which the test is administered. Second, translating an existing test is more

economical than other options, considering the time and expense that goes into developing a statewide assessment in English.

Developing an alternate-language version of a test by translating the original version probably goes a long way in facilitating comparability of the translated version to the original, because substantial resources are typically invested in the development of the standard-language version. In most statewide educational assessments, substantial time and money has been invested into developing test specifications, developing test items, aligning the tests with curricula, and compiling validity evidence. Thus, the closer the translated assessment can be to the original, the greater likelihood the translated version will possess comparable validity.

That said, it is clear we cannot assume translated tests have sufficient psychometric quality, let alone psychometric properties that are similar to the standard-language version from which they are derived. Thus the lack of research on the comparability of original and translated tests is troubling. Perhaps states are conducting such research and the results are reported in technical manuals that are not publicly available. We hope that is the case because the use of translated tests requires evidence for the validity of the interpretations derived from the test scores.

When translated versions of educational tests are used, the specific purposes for which they are used must be considered when gathering validity evidence in support of use of the test. The current literature on evaluating translated versions of educational tests has primarily focused on evaluation of group differences (e.g., Anderson et al., 2000; Duncan et al., 2005) and evaluation of DIF and internal structure (e.g., Ercikan & Koh, 2005; Allalouf et al., 1999; Sireci & Khaliq, 2002). Evidence based on relations with external variables (e.g., differential predictive validity) and testing consequences should also be investigated to shed light on their appropriateness in specific situations. As Messick (1989) pointed out, “validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness of inferences* and *actions* based on test scores or other modes of assessment” (p. 13). We recommend that all sources of validity evidence be investigated whenever possible, and we also recommend that test developers consider ELLs throughout the test development process. Such consideration involves explicit inclusion of ELLs in pilot tests, norming studies, and analysis of DIF, and careful consideration of the various types of ELLs when conducting sensitivity reviews.

Empirical analyses are useful for evaluating the consistency of statistical characteristics of tests and items across original and translated versions of a test. However, good test translation begins with careful adaptation designs and quality control procedures to ensure consistency of meaning across original (source) and translated (target) versions of an assessment. One recent suggestion for improving cross-lingual assessment is simultaneous development of different-language versions of an assessment (Solano-Flores et al, 2002). This approach also deserves further research.

Our literature review pointed out that, aside from the comparisons of international comparisons of educational achievement, research on test translations in the U.S. has focused on translations of test material into Spanish. Given that tests are translated into many other languages within the U.S., research on assessments in these other languages is needed. The lack of research on non-

Spanish versions of translated tests may be due to relatively small sample sizes, so future research should also evaluate statistical methods that are best in small-sample situations.

One approach that has been used in statewide achievement testing is the presentation of test material in two languages within the same test booklet. Stansfield (2003) defended the use of dual-language assessments for ELLs by concluding, “When considered as a potential test accommodation, there is a growing belief that it ‘does no harm,’ while at the same time relieving the examinee and the test administrator from having to decide the language of the test. Since many examinees who take translated tests have some degree of bilingualism, making available the tests in both languages may reduce construct-irrelevant variance due to the influence of test language for such examinees” (p. 201). This point is well taken; further research on the strengths and limitations of dual-language test booklets is also needed.

In closing, it is important to remember, that like standardized testing in general, the specific accommodations granted to an ELL, such as providing a test in a different language, may not be suitable for all students within a particular linguistic group. As Anderson et al. (2000) pointed out,

Translations are not appropriate for every speaker of a particular language and not every student wants or will use an accommodation...on a large-scale assessment. A standardized method of determining which students are most likely to benefit from an accommodation is desirable. Results from native language proficiency and English language proficiency assessments would be indicators of the ability to benefit from a translation, but they should not be the only indicators. ...Students and families should have input in deciding which accommodations will be offered to an individual student. For the majority of language groups present in the [ELL] student population there may not be a standardized test of academic language proficiency in their first language. In these cases, other types of indicators of academic language proficiency in the native language need to be developed.

We consider this to be sage advice. Translated versions of educational tests are not likely to be panaceas for testing ELLs, but they will make the tests more accessible to many ELLs, and in many cases, are likely to provide valid information regarding their academic knowledge, skills, and abilities. Further research in this area will help inform test adaptation practices, and will help improve the validity of interpretations derived from translated assessments.

References

- Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the sources of differential item functioning in translated verbal items. *Journal of Educational Measurement, 36*, 185–198.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.
- American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, D.C.: American Psychological Association.
- Anderson, M., Liu, K., Swierzbin, B., Thurlow, M., & Bielinski, J. (2000). Bilingual accommodations for limited English proficient students on statewide reading tests: Phase 2 (*Minnesota Report No. 31*). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [March 23, 2003], from the World Wide Web: <http://education.umn.edu/NCEO/OnlinePubs/MnReport31.html>
- Angoff, W. H., & Cook, L. L. (1988). *Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test (Report No. 88-2)*. New York, NY: College Entrance Examination Board.
- Baxter, G.P., Ahmed, S., Sikali, E., Waits, T., Sloan, M., & Salvucci, S. (2007). *Technical Report of the NAEP Mathematics Assessment in Puerto Rico: Focus on Statistical Issues* (NCES 2007-462). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, D.C.
- Bechger, T. M., van den Wittenboer, G., Hox, J. J., & De Glopper, C. D. (1999). The validity of comparative educational studies. *Educational Measurement: Issues and Practices, 18*(3), 18–26.
- Beller, M, Gafni, N., & Hanani (2005). Constructing, adapting, and validating admissions tests in multiple languages: The Israeli case. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.) *Adapting educational and psychological tests for cross-cultural assessment* (pp. 297-320). Hillsdale, NJ: Lawrence Erlbaum.
- Boldt, R. F. (1969). *Concurrent validity of the PAA and SAT for bilingual Dade School County high school volunteers*. College Entrance Examination Board Research and Development Report 68–69, No. 3, Princeton, NJ: Educational Testing Service.
- Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-cultural psychology, 1*, 185–216.

- Campbell, J. R., Kelly, D. L., Mullis, I.V.S., Martin, M. O., & Sainsbury, M. (2001, March). *International Association for the Evaluation of Educational Achievement: Progress in International Reading Literacy Study*. Chestnut Hill, MA: PIRLS International Study Center, Lynch School of Education, Boston College.
- Cascallar, A. S., & Dorans, N. J. (2005). Linking scores from tests of similar content given in different languages: An illustration of methodological artifacts. *International Journal of Testing*, 5, 337–356.
- Cook, L. L., Schmitt-Cascallar, A. P. (2005). Establishing score comparability for tests given in different languages. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.) *Adapting educational and psychological tests for cross-cultural assessment* (pp. 139–170). Hillsdale, NJ: Lawrence Erlbaum.
- CTB/McGrawHill (1988). *Spanish assessment of basic education: Technical report*. Monterey, CA: McGraw Hill.
- Davis, S. L., Buckendahl, C. W., & Plake, B. S. (2006, April). *When adaptation is not an option: An application of cross-lingual standard setting*. Paper presented at the 5th Conference of the International Test Commission, Brussels.
- Davison, M.L., & Sireci, S. G. (2000). Multidimensional scaling. In H.E.A. Tinsley & S. Brown (Eds.), *Handbook of multivariate statistics and mathematical modeling* (pp. 325–349). Washington, DC: American Psychological Association.
- Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement*, 28, 227–246.
- Duncan, T. G., del Rio Parent, L., Wen-Hung Chen, Ferrara, S., Johnson, E., & Oppler, S., et al. (2005). Study of a dual-language test booklet in eighth-grade mathematics. *Applied Measurement in Education*, 18(2), 129–161.
- Ercikan, K. & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing*, 5(1), 23–35.
- Fitzgerald, C. (2005). Test adaptation in a large-scale certification program. In R.K. Hambleton, P. Merenda, & C. Spielberger (Eds.) *Adapting educational and psychological tests for cross-cultural assessment* (pp. 195–212). Hillsdale, NJ: Lawrence Erlbaum.
- Garcia, T., del Rio Paraent, L., Chen, L., Ferrara, S., Garavaglia, D, Johnson, E., Liang, J., Oppler, S., Searcy, C., Shieh, Y., & Ye, Y. (2000, November). *Study of a dual language test booklet in 8th grade mathematics: Final report*. Washington, DC: AIR.
- Gierl, M. J. (2000). Construct equivalence on translated achievement tests. *Canadian Journal of Education*, 25, 280–296.

- Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement*, 38, 164–187.
- Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, 10, 229–244.
- Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment*, 17, 164–172.
- Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R.K. Hambleton, P. Merenda, & C. Spielberger (Eds.) *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3–38). Hillsdale, NJ: Lawrence Erlbaum.
- Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological Assessment*, 11, 147–157.
- Hambleton, R. K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. *Social Indicators Research*, 45, 153–171.
- Hambleton, R. K., Sireci, S. G., & Robin, F. (1999). Adapting credentialing exams for use in multiple languages. *CLEAR Exam Review*, 10 (2), 24–28.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating: Methods and practices*. New York: Springer-Verlag.
- Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, 6, 83–102.
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement*, (3rd ed., pp. 13–100). Washington, D.C.: American Council on Education.
- Mislevy, R. J. (1992, December). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: Educational Testing Service.
- Organization for Economic Co-operation and Development (2000). *OECD Program for International Student Assessment: National Project Manager's Manual*. Available at <http://www.oecd.org//els/PISA/Docs/Downlaod/npmmanual110200.doc>.
- Robin, F., Sireci, S. G., & Hambleton, R. K. (2003). Evaluating the equivalence of different language versions of a credentialing exam. *International Journal of Testing*, 3, 1–20.

- Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika*, *58*, 159–194.
- Sireci, S. G. (1997). Problems and issues in linking tests across languages. *Educational Measurement: Issues and Practice*, *16*(1), 12–19.
- Sireci, S. G. (2005). Using bilinguals to evaluate the comparability of different language versions of a test. In R.K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 117–138). Hillsdale, NJ: Lawrence Erlbaum.
- Sireci, S. G. (in press). Evaluating test and survey items for bias across languages and cultures. In D. Matsumoto and F. van de Vijver (Eds.) *Cross-cultural research methods in psychology*. Oxford, UK: Oxford University Press.
- Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing*, *20*, 148–166.
- Sireci, S.G., Foster, D., Olsen, J.B., & Robin, F. (1997, March). *Comparing dual-language versions of international computerized certification exams*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.
- Sireci, S. G., & Gonzalez, E. J. (2003, April). *Evaluating the structural equivalence of tests used in international comparisons of educational achievement*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.
- Sireci, S. G., Harter, J., Yang, Y., & Bhola, D. (2003). Evaluating the equivalence of an employee attitude survey across languages, cultures, and administration formats. *International Journal of Testing*, *3*, 129–150.
- 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., & Parker, P. (2006). Validity on trial: Psychometric and legal conceptualizations of validity. *Educational Measurement: Issues and Practice*, *25* (3), 27–34.
- Sireci, S. G., Yang, Y., Harter, J., & Ehrlic, E. (2006). Evaluating guidelines for test adaptations: An empirical analysis of translation quality. *Journal of Cross-Cultural Psychology*.
- Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2002). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing*, *2*(2), 107–129.

- Stansfield, C. W. (2003). Test translation and adaptation in public education in the USA. *Language Testing, 20*, 189–207.
- Stansfield, C. W., & Bowles, M. (2006). Test translation and state assessment policies for English language learners. In C. Rivera & E. Collum (Eds.), *State assessment policy and practice for English language learners: a national perspective*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- van de Vijver, F., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1* (2), 89–99.
- van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13*, 29–37.
- van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R.K. Hambleton, P. Merenda, & C. Spielberger (Eds.) *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39-63). Hillsdale, NJ: Lawrence Erlbaum.
- van de Vijver, F. & Tanzer, N. K. (1998). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47*, 263–279.
- Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement, 30*, 1–21.

Chapter 9: Considerations for Developing and Implementing Translations of Standardized K–12 Assessments

Laura J. Wright
Center for Applied Linguistics

Introduction

Within the United States educational system, translating standardized tests into other languages has become increasingly popular in recent years. This is, in part, due to policy changes coinciding with the No Child Left Behind Act (NCLB), requiring that *all* students be tested yearly to demonstrate adequate yearly progress. Translations are specifically targeted as an accommodation for English language learners (ELLs) who, until recently, were often excused from participating in standardized testing until they became sufficiently proficient in English to demonstrate their academic knowledge in English (Rivera & Collum, 2006). However, because of current educational policy, the need to include ELLs meaningfully in testing has become a widespread concern. Translating standardized tests from English into other languages is one way that test administrators have sought to meaningfully include ELLs, but as straightforward as this process may appear to be, research has indicated that numerous factors can threaten the validity of translated tests (Hambleton, 2005). These factors should be accounted for in the translation process so that translated versions are valid and reliable, comparable to the original or source version, and measure the same construct.

In this chapter, we draw upon findings from recent research to propose a multi-step process for test translation that helps ensure the development of high-quality, valid, and reliable assessment instruments. We begin by outlining factors that should be considered before deciding whether or not test translation is the most appropriate accommodation strategy with regard to student demographics and state and district policy concerns. Next, we review the types and formats of translations that are appropriate under certain conditions. In the third section, we highlight steps in the actual translation process that can help ensure the production of a high-quality translation. Aspects such as choosing a qualified translation company and/or translator(s), the method used to verify the translation, and measures of quality control are discussed as they relate to the production of an accurate, cross-culturally appropriate assessment. In the final section, we discuss the need for piloting the translated version in order to check for various types of bias, and briefly list types of statistical analyses that may be conducted to examine outcome scores, ensure the validity of the test, and locate areas of the test that may need further refinement before going to scale (see Sireci 2009, this volume, for more detail about statistical analyses). In the conclusion, we discuss the importance of an iterative test development process that allows for collaboration among test developers, translators, test takers, and administrators. Overall, this chapter serves as a practical guide for those considering test translation as one possible way of accommodating ELLs and seeks to relate research in a meaningful way so that the potential of test bias can be minimized.

I. Student and Administrative Considerations

Before deciding to enter into the test translation process, a number of administrative concerns related to budget and policy should be discussed. One of the first considerations that testing officials ought to consider is the amount of time and money allocated to the project. Stansfield (2003) reports that states typically only allow two weeks to a few months for the development of a test translation. As seen from the process outlined in this chapter, this amount of time is likely to be insufficient to produce a carefully translated version of a test. While the actual translation of items may only take a few weeks, thoughtful decision making, translation, field testing and refinement, and training require a great deal of time and effort. Officials making decisions about the test should attempt to account for all these processes when devising a timeline for the production of a translation and be realistic about the amount of time and effort that should go into developing a valid test.

In addition to the time that it takes to produce a translation, all costs associated with the translation process should be considered. Typical costs associated with developing a translation include the cost for preparation (contracting a qualified translator or translation company); the cost for developing supplemental materials such as instructions, test answer sheets, test results, and training materials for test administrators; and the costs for scoring the translated test. Item formats such as multiple choice are less expensive to score, but constructed-response items require hiring bilinguals who know both the content area and have been trained in scoring test items. A full accounting of these financial obligations is important to consider so that the translation is developed as thoroughly as the original version of the test (Fortuny, et al., 2005).

A final administrative consideration is the state and local policy toward bilingualism. For example, some states that have large populations of ELLs such as California and Arizona have passed legislation that restricts bilingual education (Proposition 227 in California [1998], and Proposition 203 in Arizona [2000]). Developing translations of standardized tests may not be in keeping with such policies.

In addition to the administrative concerns of test development, there are a number of factors related to students' linguistic, cultural, and educational backgrounds that should be considered. Gathering information about the student group(s) will enable officials to decide whether translation is the most appropriate method of accommodation for a given population of students, and, if so, whether a written or oral translation is the ideal format. In addition, this background information will inform test developers during the process of the translation. Given the demographic factors, officials must determine whether translation is a cost-effective accommodation. For example, if a particular language group accounts for a significant majority of the ELL population, translation may be a cost-effective accommodation because the cost of test development will be spread out across numerous students. A written translation is less cost-effective for relatively small language groups, and in cases where smaller numbers of students must be tested, an oral translation or sight translation may be more cost-effective (these formats are discussed below).

Similarly, given the linguistic and educational background of a language group, a written translation may be more appropriate for students from one background, while an oral translation may be more appropriate for another. Consideration should be given to whether students have

mastery of the written code of their native language; if so, a written translation is often the most reliable and cost-effective type of test translation. If students have mastery of only the oral code of their native language, then an oral translation would be more appropriate. Stansfield and Bowles (2008) provide a practical example of why it is important to consider these factors before entering into the test translation process. They report that a translation project conducted for a standardized test in Rhode Island was only partially successful because two of the student groups for whom tests were translated, Lao and Khmer speakers, did not have high levels of reading and writing ability in their native language. While these students spoke Lao or Khmer as their home language, translating the test into a written form was not the most appropriate accommodation for them, and investing time and money into developing the written test translation did not yield successful results. As noted above, an oral translation may have been a more appropriate strategy.

A third aspect to consider before deciding to develop a test translation is the typical educational background of students from a given language group (Bowles & Stansfield, 2008). Students taking a test in their native language should have had access to education in their native language in the subject area in which they are being tested. That is to say, they either need to have had formal schooling in their home country, or have had access to bilingual education in the U.S. Many subject areas have specialized vocabulary and grammar not typically used in everyday conversation and because a translated test will necessarily rely upon these linguistic resources that are likely not learned outside of school contexts, it is important to account for whether and how much access students have had to formal education in their native language.

A fourth aspect to consider and identify before entering into the test translation process is the linguistic and cultural backgrounds of students (Hambleton, 2005; Solano-Flores, 2006; Solano-Flores, Speroni & Sexton, 2005). Before beginning a translation, test developers should initially investigate what dialect(s) their student population speaks. In addition, they should determine whether dialect groups of the same language have different cultural assumptions. For example, there are many Arabic speakers in the world, but dialects of Arabic differ greatly, with different lexical items (words), grammatical structures, idioms, and colloquialisms that may affect how a construct is represented. Hambleton (2005) advises that decisions about the dialect variety used for a test should be decided from the very beginning and that translators should be told how to treat aspects of dialect variation for the test. Dialects are systematic varieties of languages and therefore, specific features of the dialect should be identified for translators so that the linguistic modifications are applied systematically to the translation. Furthermore, the translator/translation company that is contracted should have insight into these issues and should choose qualified translators accordingly. (For example, a Castilian Spanish speaker may not be the best choice to translate a test for Salvadoran Spanish speakers unless s/he has had experience with that dialect and culture.)

A final aspect to consider with regard to the student population taking a translated test is student interest (Bowles & Stansfield, 2008). Not all students necessarily want to take a translated version of a test because of social pressures. Test developers should take into account student attitudes toward test translations to ensure that students are open to taking the test in their native language.

If, after considering administrative concerns and student backgrounds, test translation appears to be an appropriate accommodation, then a test translator or translation company must be selected. In the next section we provide background on important terminology and test translation theories that officials should understand before choosing a translation company. This will allow officials to decide what theoretical stance they wish to take toward the translation process and provide them the necessary information to instruct translators on how they would like the test to be translated.

II. Terminology and Processes of Developing Native Language Assessments (NLA)

To begin the discussion about the process of developing test translations, it is important to define some of the terminology typically used, and how this terminology, as it relates to the development processes, may be different from everyday understandings. First, there are three ways that test translations are referred to in research, test *translation*, *adaptation* (Hambleton, 2005), and *native language assessment (NLA)* (Bowles & Stansfield, 2008). The terms *translation* and *adaptation* reflect theoretical distinctions that affect the development process; however, they are often used interchangeably in test development literature (Hambleton, 2005). *NLA* attempts to avoid the theoretical distinctions taken by test developers and is inclusive of a third method of development called *parallel development* (described below). For the purpose of this chapter, we use these terms interchangeably to refer to the process of developing a test for speakers of other languages based on an original or source version of a test, or a test framework.

The terminology described above is indicative of the theoretical stance test developers take toward the *original* or *source* version of the test, and the methods they use for taking that version into a *target* language. Of these three, *adaptation* is the most popular and theoretically accepted form of development. Hambleton (2005) chronicles the favor of this process, noting that its theoretical underpinnings are a result of an International Test Commission conference held in 1999. *Adaptation* as a process reflects that items can be somewhat flexibly modified to account for linguistic differences between languages such as semantic differences and idiomatic expressions, and to account for cross-cultural construct comparability (Hambleton, 2005). When a test from one language is adapted (translated) to another, some of the items may need to be modified so that they are understandable cross-linguistically and cross-culturally. For example, an idiomatic expression in one language may not exist in another and therefore may need to be modified or represented with a comparable, but not literal, linguistic expression. Likewise, the construct of an item may be affected by its translation. For example, Solano-Flores (2006) reports that an item on a test translated from English into Spanish and Haitian-Creole posed a linguistic problem that potentially affected the construct. In English, the item stated that the height of a dinosaur was “rounded to the nearest 10 feet.” Those who were translating the test suggested that this may not make sense in Spanish or Haitian-Creole, and, as an alternative, they suggested using the wording “to the nearest tenth” even though this slight modification may have made the item more difficult. Solano-Flores (2006) states that making these kinds of decisions in test translation creates a tension between construct comparability and dialect and register. He writes, “even a correct translation that intends to address construct comparability may not necessarily be able to capture entirely the thinking associated with language and culture” (p. 2365). The term *adaptation* rather than *translation* reflects the practice of modifying test items so that they are equivalent or comparable, though not linguistically verbatim.

Translation, on the other hand, often suggests a word-for-word, or literal translation, which requires very little modification from one language to another (Bowles & Stansfield, 2008). Because the terms *adaptation* and *translation* reflect a theoretical difference in test development practices, most testing specialists prefer the term *adaptation* at the technical level with the idea in mind that construct equivalence is the primary objective when translating a test (Hambleton, 2005). Potentially, treating language somewhat flexibly, with appropriate construct-related constraints, could reduce the risk of item bias which stems from translation issues (Sireci, Patsula, & Hambleton 2005). Within testing literature, the technical distinctions of these words are discussed, but in actual practice, the terms *translation* and *adaptation* are often used interchangeably.

A third, less common process for developing tests in students' native languages is called *parallel development* (or *concurrent* or *simultaneous* development) (Bowles & Stansfield, 2008; Solano-Flores, Trumbull, & Nelson-Barber, 2002; Tanzer, 2005). This is a process in which two or more tests are developed at the same time, using the same blueprint, specifications, or test framework. For example, test developers are given a framework for developing a test in English and Spanish at the same time and work on both tests simultaneously, developing each test item by item in the two languages. The benefit of this type of development process is that if an item is not culturally appropriate or if it is awkward to translate, developers can be alerted to the potential problems during the development stage, rather than after a test has been fully developed. Ideally, parallel development leads to two (or more) tests that are comparable at the level of construct and, because they are developed simultaneously, the chance of item bias resulting from discrepancies in the translation is minimized.

The terminology used to describe the test development processes reflects important theoretical distinctions. Before entering into the test translation process, it is important for officials and test developers to decide which approach to take, and to convey the theoretical distinctions to those who will translate the test because these theoretical distinctions can ultimately affect the way in which items are treated throughout the translation process. Now that some basic terminology regarding native language assessments and the processes by which they are developed has been described, we would also like to explain two types of formats that are utilized for test translations and their benefits and potential drawbacks.

Written Translations

Bowles and Stansfield (2008) indicate that 12 states in the U.S. currently offer written translations of standardized assessments to students who speak languages other than English as their native or home language, and that written formats are the most commonly used type of translation. The primary decision when developing a written translation is whether to use a *dual-* or *single-*language format to administer the test. A *dual-*language format provides students with an item written in English on a page of a test booklet, with the equivalent item in the students' native language either on the same or the opposite page. This enables the student to see both versions of the item at the same time when taking the test. A *single-*language version, on the other hand, only provides items in the target language, or the language into which the test is translated, so the student or test administrator must decide before the student takes the test which version is most appropriate.

Some research has suggested that there are benefits to using a dual-language format (Anderson, Liu, Swierzbis, Thurlow, & Bielinski, 2000; Garcia et al., 2002; Liu, Anderson, Swierzbis, & Thurlow, 1999; Stansfield & Kahle, 1998) as opposed to a single-language format. First, many students who are being tested in their native language may have had exposure to certain content areas in English and may know school-related technical vocabulary in English, but not in their native language if they have not had content area instruction in their native language. Providing a dual-language version of the test provides the student with the opportunity to read both versions of the item and glean information from both. Work by Anderson et al. (2000) indicates that low and intermediate ELLs may do twice as well when given a dual-language version over a single language version, although this was not considered to be a statistically significant finding. In addition, both Anderson et al. (2000) and Duncan et al. (2005) found that ELLs appreciated having dual-language versions of written tests, suggesting that this format has affective benefits (see Sireci, 2009, this volume, for further information on these studies).

While a dual-language format has benefits for lower proficiency speakers, it may not have the same benefits for students who have developed higher language proficiency. Duncan et al. (2005) found that students who had reached high English proficiency and used a dual-language format scored lower than those who used a single-language format. While their sample size was not large enough to conduct statistical analyses, it does suggest that students' overall language proficiency is a factor that should be accounted for when deciding whether or not to develop and provide a dual-language format. A final consideration in providing a single-language format is who will decide which test format will be provided to students. In some schools, the teacher or student is allowed to make this choice, while in other areas this decision is made at the district level based on students' English language proficiency scores.

Oral Translations

Another less commonly used format of native language assessment is oral translation. Bowles and Stansfield (2008) indicate that oral translations are a good option for students who may be proficient in speaking their native language, but are less proficient at reading and writing it. There are two main types of oral translations provided to students in their native language: recorded or scripted oral translations, and sight translations (Bowles & Stansfield, 2008). A recorded version consists of a script being prepared from the source assessment (with checks for quality control much the same that a written form would undergo). Then, a native speaker who has also had professional voice training in the target language reads the test items aloud into a recording system using the script. After the recording has been made, another translator verifies the recording against the script to ensure that no words have been omitted and that the pronunciation of the items is comprehensible. If the recording is not verified, it is professionally re-mastered in an audio studio; if the recording is verified, it can be reproduced in a variety of formats including audio CDs, DVDs, and computerized forms, and distributed to test administrators. Stansfield (2008) notes that three states used recorded oral formats in 2006–2007, Michigan, Ohio, and Wisconsin, and each used a slightly different format. In all cases, however, the students saw the written English test as they heard the oral translation and answered in an English test booklet. In addition, students were able to pause and replay the recording as needed.

A benefit of recorded oral translations is that the delivery across testing contexts is standardized; all students hear exactly the same script, and the pacing, pronunciation, and wording is the same. However, this format is less flexible across contexts and aspects such as accent and dialect may not be modified according to student demographics. For example, if a test is translated into Spanish or Arabic, both of these languages have multiple dialects that vary significantly. It is important to be sure that the translation is sensitive to the dialect of the students who are taking the test and to minimize any confusion that could be caused by the speaker's accent or dialectal differences. Stansfield (2008) reports that an additional benefit of recorded oral translations is that they can be more cost effective than sight translations if there is a large enough population of students in a language group. He states that the state of Ohio actually determined that if a test is sight translated more than 59 times, it is less expensive to produce a recorded oral translation, even though it is expensive to produce the recording and distribute it to schools.

A second type of oral translation is a sight translation. This type of translation is performed for the student by a trained interpreter in person, item by item (Bowles & Stansfield, 2008; Stansfield, 2008). The benefit of this format is that it can be used with relatively small language groups and it can be more sensitive to students' dialects because an interpreter from a particular dialect background can be hired to match the students' background. In addition, sight translations are flexible in terms of delivery (volume, pacing, checking student understanding of questions). Typically in sight translations, an English script is provided to an interpreter a few days before the test is administered. This allows the interpreter to become familiar with the test and terminology that is used on it. On testing day, the interpreter reads and translates the directions for the test and then reads and translates each question from English into the target language. The student is allowed to ask the interpreter to repeat questions, but other types of questions that may help a student unfairly are not allowed. The student receives the English test booklet and answers the items in the English booklet either in English or his/her native language. Typically if a student answers constructed response items in his/her home language, the sight translator is expected to translate those answers for scoring purposes. Again, the interpreter is instructed not to assist students by correcting any answers that are incorrect.

Stansfield (2008) reports that there are several drawbacks to using sight translation as an accommodation for standardized tests. First, it is sometimes difficult to recruit certified interpreters for less commonly spoken languages, especially in rural settings. Not all bilinguals have an appropriate background for conducting sight translations and care needs to be taken to select qualified individuals, especially when test administrators are unfamiliar with the target language. Next, sight translation exposes test items more than any other form of translation. Therefore, all testing materials should be kept securely at the school and interpreters should be instructed that all test materials should be kept in strict confidence. Finally, the administration of sight translations cannot be standardized across implementations and contexts. Thus, differences in test scores may result from method bias (Sireci, et al., 2005), and yet remain unknown to the test developers. Because of the potential for differences in test administration, some states have policies requiring sight translation sessions to be recorded. While the above factors make sight translation less favorable than a scripted oral translation, its flexibility make it an appealing format for accommodating students who speak less commonly spoken languages and dialects, and for relatively small numbers of students requiring oral translations in a district or state.

In this section, we have described basic theoretical distinctions underlying translation processes and testing formats, and how these aspects should be considered before entering into the test translation process. It is important for officials and test developers to decide which format is most suitable and cost effective for the background of the students being tested. Both written and oral translations are appropriate given certain characteristics of students' linguistic and educational backgrounds. Now that we have described the basic formats used for test translations, we turn to a discussion of the test translation process. While this process is typically handled by the actual translators of the test, it is important for officials and test developers to provide guidance to the translators about the procedures to be used, and further, to have a plan in mind for verifying the translation that is produced to ensure that it is as accurate as possible. In the following section, we discuss factors affecting the initial quality of the translated test and the ways in which translations can be produced and checked for errors so that the overall quality of the test is high.

Issues That Affect Student Performance and Test Comparability

Before entering into the translation process, it is important to know about some of the typical problems associated with test translations so as to avoid them. Hambleton (2005) outlines four basic factors that may affect the assessment and interpretation of test results: cultural and linguistic differences, test administration differences, test formatting differences, and speededness. Because these factors may ultimately affect the interpretation of students' performance, it is important to understand these factors before undertaking the translation process. In this section, we discuss how each of these factors may affect the translated test and the considerations that can be made during the test development process to minimize their confounding effects, which lead to types of test bias.

Cultural and Linguistic Differences

When developing a test for speakers of other languages, it is inevitable that there will be cultural and linguistic differences between the original version of a test and a translation. Linguistic differences and the way in which they are translated may contribute to item bias and construct bias (Sireci, et al., 2005). All languages express ideas somewhat differently, and some ideas may be culturally irrelevant to different cultural groups. When translating a test, these linguistic and cultural aspects should be considered.

Linguistically speaking, early studies in second language acquisition called *contrastive analysis* hypothesized that the greater the linguistic difference between two languages, the greater the difficulty in learning the second language. While more recent studies have found that to be only one factor in language acquisition, the fact remains that some languages are more linguistically similar to others, and therefore easier to translate from one language to another. Linguistic differences can occur at many levels. For example, semantically, different languages can use different words and expressions to convey ideas, which can further be confounded by cultural understandings of words. For example, Valdes (1996) describes that for many Chicanos, the word *educato* as it relates to schooling means that someone is well-behaved. She further relates this to Chicano expectations of schooling, stating that many believe that the purpose of school is to help students learn to be well-behaved. Conversely the equivalent word in English, *educated*,

has a very different connotation, typically meaning someone who is intellectual and has learned a great deal of book knowledge. Culturally speaking, these words have very different meanings and it is only by probing the cultural understandings of these words that the differences can be understood. Therefore, it is especially important to find trained translators who are aware of these kinds of cultural differences.

Syntactic and grammatical differences also make test translations quite difficult, especially with longer stretches of text. Linguists have found that most languages, including English, follow a *subject, object, verb* (SOV) syntactic structure. However, some languages have different sentence structures that make translations more cumbersome. For example, Turkish sentences follow a *subject, verb, object* (SVO) syntactic structure, often with numerous embedded clauses. When developing a translation from a source language into a target language, the syntactic differences may make the translation more or less cumbersome and can affect the overall quality of the translation.

In addition to syntactic differences, there are also other grammatical aspects that can affect the overall quality of the translation. For example, many languages have different grammatical tenses that do not have equivalent meanings in another language. For example, Turkish contains two past tense markings to express different levels of certainty regarding the information conveyed (one marking first-hand knowledge, and one marking hearsay). In linguistics, these are called *evidential markers*. In English, however, these evidential markers are not grammatically structured and are typically expressed lexically. In scientific texts, evidential markers are particularly important to the meaning of the discourse (Viechnicki, 2002) and so a translator must be attuned to how these grammatical markers affect the expression of ideas and how to accurately translate the meanings across languages.

A final area of linguistic difference that should be considered when translating a test is sociolinguistic difference, including both dialect and register variation (Solano-Flores, 2006). Dialects refer to varieties of a language that people of different ethnic, regional, or socio-economic background speak. Even among native English speakers from the U.S., there is a great deal of variation between how people speak in the South versus how they speak in the North. The case may be even greater among varieties of a language spoken in different countries, for example the dialect of Arabic spoken in Iraq versus the dialect of Arabic spoken in Morocco, or the Spanish spoken in Mexico versus the Spanish spoken in Venezuela. While most languages have a “standard” variety to which a translator will try to adhere, children are often relatively unaware of the dialect they speak and potential dialect differences. These aspects of metalinguistic awareness are not typically developed until much later in life—if at all. Additionally, if a child has not had formal education in their native language, they may not be aware of the meanings or rules of “standard” forms of language. With regard to testing, Solano-Flores, et al., (2005) state that “in order to truly minimize language as a construct-irrelevant factor (a factor that is not intended to be measured) the process of test translation or adaptation must be sensitive to differences in which language is used by different communities” (p. 3). In other words, when translating a test, the specific dialect of language speakers should be considered and, if possible, modifications should be made for significant differences among dialects (Hambleton, 2005; Solano-Flores, et al., 2005).

A second area of sociolinguistic difference is that of register (Solano-Flores, 2006). Romaine (1994) defines a register as “variation in language conditioned by uses rather than users and involves consideration of the situation or context of use, the purpose, subject-matter, and content of the message, and the relationship between participants” (p. 20). Much recent work has examined the notion of academic language as a register that is specific to school settings, and even more so, the registers of subject areas such as the language of science and the language of math (Lemke, 1991, Roth, 2005, Schleppegrell, 2008). This work has shown that certain academic subjects have conventionalized ways of conveying meaning that are different from every day ways of talking, and that these new ways of using language are learned in school settings. Relatively little work has examined register differences across languages and whether similar kinds of grammatical conventions are used to convey meaning. Halliday’s (1975) seminal research on the mathematical register asserted that “*every language embodies some mathematical meanings in its semantic structure—ways of counting, measuring, classifying, and so on*” (p. 65). However, he noted that full mathematical expression is not completely developed in every language. Thus, translating a test from a language that has a highly conventionalized register may be difficult if the language into which it is translated does not have such a register. Furthermore, if translators are not familiar with the conventions of a register in the target language it may be difficult to capture nuanced meanings from the grammatical structure.

Research on the cultural and linguistic differences that may affect test translation demonstrates the need for the involvement of highly qualified translators in the process (Fortuny, et al., 2005). Stansfield and Bowles (2006) recommend that the translators chosen to develop a translation should have certification from the American Translators Association if possible (though the organization does not offer certification in all languages). Furthermore, the translators chosen to complete the task should also have relevant cultural knowledge of the group for which they are developing the translation, relevant knowledge of the domain being tested, as well as some knowledge of item writing. Involving highly qualified translators in the development process will help guard against item bias (Sireci, et al., 2005).

Administrative Differences

Another factor that can be confounding to the development of reliable translations is administrative differences (Hambleton, 2005), leading to method bias (Sireci, et al., 2005). Administrative differences include the communication that takes place between examiner and the examinee, and for this reason, among others, it is important that test directions are also translated as part of the test translation process. This will ensure that the same procedures for test administration and completion are followed across the different languages being tested. Many states provide written directions in students’ target language, even if they do not fully translate the items of the test (Stansfield & Bowles, 2006). This is one way that states have attempted to minimize the administrative differences that may affect student outcomes.

In addition to translating the test directions, it is also important to carefully select and train those who will administer the tests. Hambleton (2005) suggests that those administering the test should be selected from the target cultural group, that they should be familiar with both the language and dialect background of the students, and that they should have had adequate experience and familiarity with test administration procedures. Having two training sessions or meetings with

translators before the test and one after the test administration will help ensure that administration procedures are as uniform as possible across test groups, and if unexpected circumstances arose while the test was being administered, that test developers are made aware of potential causes of differences in outcome scores (Goldsmith, 2003). Differences in test administration can largely be avoided by being aware of potentially confounding effects and by planning for adequate translation of directions and training of administrators.

Test Formatting Differences

A third aspect that may affect test students' performance is their familiarity with item formats. Typical test and item formats differ cross-culturally; in some cultures, short answer and essay questions are most common, and multiple choice items are rare. Hambleton (2005) relates that the British educational system rarely uses multiple choice items in testing situations. Thus, many cultural groups who have had educational experiences in countries that base their tests on the British system may be unfamiliar with this test format. Likewise, essay and short answer questions are relatively unfamiliar to cultural groups that have had more educational experience with the American educational system. Lack of familiarity with test item formats may affect students' performance on a translated test, leading to method bias (Sireci, et al., 2005), so a general understanding of what test formats to which students may have previously been exposed should be considered. The potential performance differences caused by test formats can be addressed in two ways. First, if test developers know that students are less familiar with certain item formats, the item format of the test may be adapted during the translation process to one that is more appropriate for their cultural and educational background. Additionally, Kopriva (2008) stresses the importance of familiarizing students with test formats and addressing other cultural issues long before tests are given so that students have adequate time to become accustomed to different formats. Addressing these issues from both the test development standpoint and the implementation standpoint should minimize the risks of performance differences that could be caused by the format of the test.

Speededness

Although most criterion-referenced tests in K–12 are untimed, sections of a test may be timed and the expected test administration time is often specified. A final factor that may affect student performance is that of actual or perceived *speededness*. Hambleton (2005) relates that some cultural groups are more accustomed to considering time an essential component to test taking than others. Furthermore, research has shown that students who are not accustomed to taking tests quickly or those who are poor readers may be disadvantaged on certain tests. Because speed in test taking is not always a construct relevant factor, taking into account a cultural groups' orientation toward speededness ahead of the translation process may help to eliminate score bias. For example, Sireci (2009, this volume) relates that scores from students taking the National Assessment of Educational Progress (NAEP) in Puerto Rico dramatically improved when students were allotted more time to answer items. If speed is not a construct-relevant factor, then students' orientation to speed should be considered so that the number of items administered in one sitting included or the amount of time given to take a test can be adapted during the development process.

In this section, we have outlined several factors that may ultimately affect student performance on translated tests and how these factors can be considered before and during the development process to eliminate the potential of bias (Sireci, et al., 2005). In the next section, we turn to a discussion of the translation process itself and the guidance that officials should provide to the translator(s) for verifying the translation.

III. Translating and Verifying Test Translations

There are two primary approaches to translation and verification procedures, *forward* or *direct translation* and *back translation*. Both of these approaches can be done by one or more translators; however, when feasible, it is advantageous to involve a committee of translators to check and verify the accuracy of the completed test. The process of developing the translation should be treated as iterative, with potentially multiple revisions needed after the test has been initially translated. Additionally, whether a forward or backward translation process is used, a review process in which an outside translator examines the completed version should be incorporated. Finally, if the source version of the test is not yet in production, a process known as *decentering* (Hambleton, 2005), in which the source version is modified to smooth out discrepancies between the source and target version, may be helpful in facilitating the revision process. Allowing for decentering provides greater flexibility in building comparable items because if an equivalent term or expression cannot be agreed upon, it is still possible to choose a term that will suit both the original and translated versions adequately.

Direct/Forward Translation

Forward or *direct translation* is a process in which a translator, or group of translators, translate (or adapt) a test from a source language into a target language. Once the translation is completed, a different translator(s) compares the two tests to determine whether they are equivalent. The reviewer should identify any potential areas of difference and make recommendations on what and how items should be modified. Then, the test developers and original translator(s) read and discuss the suggestions and revisions are made to the test based on the judgment of the reviewer(s). This process is continued until there are few or no suggestions made about the target version of the test.

A final, optional, stage of forward translation involves hiring an additional translator once the reviewed translation has been completed to read and edit the target language test to ensure that it is well expressed. Hambleton (2005) reports that the language of some translations may be rough or unpolished, and having an independent translator look at the final version of the test for linguistic infelicities may help the test become more readable.

Forward translation is the most commonly used and cost effective translation process; however, a drawback to this method is that after a translation has been completed, a test developer may not know whether the target language test is equivalent in meaning unless s/he is also proficient in the target language (Hambleton, 2005; Stansfield, 2003). In order to obviate this problem, a process known as *back translation* is often used.

Back Translation

Back translation (Brislin, 1970) is a process whereby one or more bilinguals translate the original test into the target language and then a different bilingual individual translates the target language back into the source language (e.g., an English test is translated into Spanish and then the Spanish translation is re-translated back into English). Then, the two source language versions are compared against one another and points of discrepancy between the two tests are identified and analyzed. If a problem is found to be with the initial translation into the target language, then the translation is corrected. Typically, this process is repeated to draft, review, and revise the test.

Back translation has both benefits and drawbacks identified in the testing literature. First, the benefit of back translation is that test developers do not need to know the target language in order to identify potential translation issues because differences are identified by using two source language documents. However, test developers are left to determine whether the translation issue is due to the back translation or the translation into the target language. An additional drawback that has been identified is that if the person completing the translation knows that the translation will be back translated, s/he may be very conservative in producing the target language version or may retain inappropriate aspects from the source version (Hambleton, 2005). Rather than trying to capture the essence of the meaning of items, translators may be very literal so that it is easier to produce a translation that is equivalent to the source version. According to Stansfield (2003), being too literal may result in unnatural sounding language. Finally, the method of back translation comes from the field of cross-cultural psychology and typically uses bilinguals as translators rather than certified translators. Because of this, Stansfield (2003) notes that the people who typically complete back translations may not be as well equipped to handle some of the linguistic difficulties that arise in producing translations. Still, back translation is a popular method for verifying the quality of a test and may be an optimal way to allow test developers to understand what potential differences may exist between the source and target language versions of a test. Hambleton (2005) notes that researchers like this design precisely because it affords them the opportunity to make their own judgments about the translation without having to be fully proficient in the target language.

No matter what process is used for verifying the translation, translators should be fully briefed on the key theoretical factors that underlie the test translation, and assumptions about the treatment of language, such as how dialects should be regarded and what the revision process will be, should be discussed before the translation process begins. Overall, translators should be sensitive to the notion of construct equivalence across test versions and understand that producing a test translation may be different from the process of producing other kinds of translations. Finally, Stansfield and Bowles (2006) also recommend that if a translation is developed in a state where bilingual education programs exist, translators should be provided with the bilingual curriculum materials that are used in students' classes so that the terminology used on the target test is equivalent to that which students are expected to learn in their classes.

Once the translation has been produced and verified, it is not necessarily ready to be used for testing. Ideally, the test will first be piloted with students from the language background for which the test is intended and empirical checks on the comparability of the translation to the

English version will be conducted (see Sireci, 2009, this volume). In the next section, we describe two ways that tests can be piloted with students to ensure that there are not unforeseen issues with the ways the items perform.

IV. Piloting and Analyzing Results

The first way that items may be tested to ensure that students interact with them as intended involves interviewing a small number of students to gather their reactions to the items. These interviews are done fairly early in the test development process so that there is time to revise items if any issues are found with them. The interviews, called *cognitive labs*, typically use a *think-aloud* method in which students interact with test items while describing their interactions to an interviewer, or a *retrospective* method in which students answer items and then explain their answers to the interviewer (Willis, 2005). Based on student feedback, test developers can refine and revise items so that they are comprehensible to students. With test translations, this may be an important way of detecting differences in cultural understandings or linguistic differences.

Cognitive labs typically follow an interview protocol that is used to elicit information from 5 to 15 students about selected items from the test. Based on student responses, qualitative generalizations about student understanding are formed. If there are misunderstandings that are consistent across interviewees, the test developers can make modifications to the items before the test goes into production. Often, a second round of cognitive labs is done with items to verify that any issues have been resolved before they are tested with a wider selection of students. A drawback to cognitive labs is that they are labor intensive and only a small number of students are involved in responding to a relatively small number of items. Furthermore, conducting interviews cross-culturally may also be fraught with confusion if students are uncomfortable thinking aloud or admitting misunderstanding. Research suggests that many cultural groups have difficulty expressing disagreement directly, especially with those they perceive to be in a position of greater power (Fortuny et al., 2005; Pan, 2003). However, with proper training, many of these issues can be overcome by explaining the interviewees' role in the test development process, and helping them feel comfortable in the interview situation. The information provided by students in cognitive lab settings is potentially rich and may shed light on issues that may otherwise be overlooked.

Another way that tests can be piloted is in a *controlled trial*. Controlled trials require larger numbers of students who are given the test in the exact format in which the test is intended to be given. The tests are then rated and scored, and the scores across different demographic groups of students are compared. It is important to choose groups of students who will behave similarly to those who are intended to take the translated version. Sireci (1997) outlines three groups who may be matched in a controlled trial to compare scores. First, using both source and translated tests, bilingual students who are proficient in both languages could be tested and compared. Second, using a back translation of the source test, monolingual students from the source language group could take both the source version and the back translation version of the test and scores could be compared. Third, using both the source and target tests, two groups of monolingual students from the source and target language groups could be matched in terms of their proficiency with regard to the subject matter and tested; the scores would be compared to determine whether students of similar proficiency levels in a given domain do indeed perform

similarly. In the context of NCLB, adequate yearly progress calculations, the accurate determination of proficiency levels, are critical since these are the scores used in accountability.

If the translation is tested in a controlled trial, it is important that a qualified statistician or psychometrician performs the necessary statistical analyses (see Sireci, 2009, this volume for a further discussion of psychometric studies and concerns). According to Sireci, et al., (2005), all assessment instruments that are used cross-culturally should be examined to determine whether *construct*, *method*, or *item bias* may affect students' performance, ultimately affecting the comparability of tests. In this final section, we provide an overview of how these types of biases have been defined in testing literature, and identify the recommended types of statistical methods that may be used to identify these types of biases.

Construct bias is said to occur when a construct is not conceptually equivalent across cultural groups (Sireci, et al., 2005). That is to say, if constructs are defined differently cross-culturally, a test instrument should reflect the differences at the level of the construct. In the case of native language assessments used in U.S. K–12 schools, the construct is considered at the item level and is determined in accordance with state standards in each domain area. Because of this, there is less chance that construct bias will affect a test a K–12 standardized assessment (Sireci, 2009, this volume). While construct equivalence may be judged by a group of experts before field testing takes place, Sireci, et al., (2005) also identify a number of statistical methods that may be used after field testing including exploratory factor analysis, confirmatory factor analysis, multidimensional scaling, or comparing nomological networks. Of these, exploratory factor analysis is one of the most popular, if least rigorous, methods (van de Vijver & Poortinga, 1991), but both confirmatory factor analysis and multidimensional scaling also present benefits (Sireci, et al., 2005; Sireci, 2009).

Method bias occurs when there are biases in the conditions surrounding the administration of the test. According to van de Vijver and Tanzer (1997), there are three primary types of method bias. The first type of bias, *sample bias*, is related to differences in test scores due to the cultural or linguistic background of the students being tested such as their socio-economic status or their motivation to do well on the test. The second type of bias, *instrument bias*, relates to the way in which the instrument functions across groups and students' potential lack of familiarity with different testing formats. The final type of bias, *administration bias*, relates to the administration procedures and the way in which test instructions may be interpreted or delivered differently. As discussed earlier, with careful planning before a test is given, these types of bias can be minimized, but a statistical analysis after a pilot test will help test developers rule out the potential that they will adversely affect the test results. Sireci, et al., (2005) suggest that demographic information be collected when students take the test in order to be able to do analyses to rule out *sample bias*. This will allow the analyst(s) to divide students into meaningful demographic subgroups and disaggregate and interpret data accordingly. In order to rule out *instrument bias* or *administration bias*, Sireci, et al., (2005) recommend using monotrait-multimethod studies, collateral information (such as measuring the time it takes a student to respond to items on the test), and test-retest change.

A final type of bias that may affect the comparability of tests is *item bias*. Item bias occurs when there are issues with the translation itself. Sireci, et al., (2005) report that the statistical methods

used to investigate *item bias* vary from simple methods that employ visual analysis to more complex methods that rely upon modern measurement theory. The choice of a particular method depends on a number of factors on which a trained statistician or psychometrician can inform the test developers. Depending on the sample size, the total number of items on the test, how the items are rated, and the developers' access to statistical software, the statistical consultant can determine which approach is most appropriate. Sireci, et al., (2005) recommend using a process known as differential item functioning (DIF) analysis, to determine whether differences in item performance are due to student proficiency in the subject area or to item bias. Other popular methods for checking for item bias include the delta plot method, standardization, logistic regression and Lord's chi square (Sireci, et al., 2005). Finally, when conducting a statistical analysis of test results, Sireci, et al., (2005) recommend that construct and method bias be ruled out before examining results of an item level analysis.

Once a statistical analysis has shown that the test is not biased at any level, the test may go into production and be used at scale. Even after a test has gone into production, it is still important to get feedback from those administering the test, and their comments should be collected in case further revisions need to be made, and for future test development (Geisinger, 2003). Additionally, test developers should give thought as to whether the test results that are sent to students and their parents and interpretations of those results should also be translated. As noted previously, the translation process involves more than translating the test itself, and providing interpretations of the results are an important way of including ELLs and their families in the educational system.

Conclusion

This chapter has outlined the many steps that should be considered when attempting to develop a valid, reliable test translation that is comparable to the source version. As seen from the process that is outlined, gathering information on students' backgrounds, choosing a method of translation, translating and verifying a test, piloting items or the test itself and analyzing results is a labor intensive process. However, the benefits yielded from meticulous care to the process should ensure a test that will likely yield the most successful results possible. Developing a test translation is an iterative process that should involve collaboration among officials, test developers, translators, test takers, and administrators. The process outlined in this chapter accounts for the involvement of all of these groups of stakeholders and shows how each group can contribute to the development process. Involvement at all levels can lead to the development of a test that reflects the concerns of each group and their interests in high-stakes testing.

Translations are not always the most appropriate accommodation for students, therefore all alternatives should be considered before entering into this decision. In fact, some research (Kopriva, Emick, Hipolito-Delgado, & Cameron, 2007) suggests that determining what an appropriate accommodation is for individual ELLs is critical to their success, rather than adopting a one-size-fits-all approach. The other chapters of this handbook highlight other accommodations that officials may want to consider and weigh before entering into the test translation process.

References:

- Anderson, M., Liu, K., Swierzbis, B., Thurlow, M., & Bielinski, J. (2000). Bilingual accommodations for limited English proficient students on statewide reading tests: Phase 2 (*Minnesota Report No. 31*). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [April 2, 2009], from the World Wide Web: <http://education.umn.edu/NCEO/OnlinePubs/MnReport31.html>
- Bowles, M., & Stansfield, C. (2008). A practical guide to standards-based assessment in the native language. Retrieved 1/23/09, from the World Wide Web: http://www.ncela.gwu.edu/spotlight/LEP/2008/bowles_stansfield.pdf
- Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-cultural Psychology*, 1, 185–216.
- Duncan, T.G., del Rio Parent, L., Chen, W.H., Ferrara, S., Johnson, E., Oppler, S., et al. (2005). Study of a dual-language test booklet in eighth-grade mathematics. *Applied Measurement in Education*, 18(2), 129–161.
- Fortuny, L.A., Garolera, M., Hermsillo, R., Feldman, E., Fernandez Barillas, H., & Keefe, R. (2005). Research with Spanish-speaking populations in the United States: Lost in translation a commentary and a plea. *Journal of Clinical and Experimental Neuropsychology* 27, 555–564.
- Garcia, T., del Rio Parent, L., Chen, L., Ferrara, S., Garavaglia, D., & Johnson, E., (2000). Study of a dual language test booklet in 8th grade mathematics: Final report. Washington, DC: AIR.
- Geisinger, K. F. (2003). Testing and assessment in cross-cultural psychology. In J.R. Graham & J.A. Naglieri (Eds.) *Handbook of Psychology*. (pp. 95–117). Hoboken, NJ: Wiley & Sons.
- Goldsmith, S.M., (2004). Lost in translation: Issues in translating tests for non-English speaking, limited English proficient, and bilingual students. In J. Wall & G.R. Waltz, *Measuring up: Assessment issues for teachers, counselors, and administrators*. Greensboro, NC: CAPS Press.
- Halliday, M.A.K. (1975). Some aspects of sociolinguistics. *Interactions between mathematics and linguistics* (pp. 64–73). UNESCO.
- Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, and C.D.
- Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment*. (pp. 3–38). Mahwah, NJ: Lawrence Erlbaum Associates.

- Kopriva, R.J., Emick, J.E., Hipolito-Delgado, C.P., & Cameron, C. (2007). Do proper accommodation assignments make a difference? Examining the impact of improved decision making on scores for English language learners. *Educational Measurement: Issues and Practice*, 26(3), 11–20.
- Kopriva, R. J. (2008). *Improving testing for English language learners*. New York: Routledge.
- Lemke, J. (1990). *Talking science: Language learning and values*. Norwood, NJ: Ablex.
- Liu, K., Anderson, M., Swierzbis, B., & Thurlow, M. (1999). *Bilingual accommodations for limited English proficient students on statewide reading tests: Phase 1 (Minnesota Report 20)*. Minneapolis: University of Minnesota, National Center on Educational Outcomes.
- Pan, Y. (2003). The role of sociolinguistics in federal survey development. Proceedings of the Federal Committee on Statistical Methodology Research Conference. November, 2003, Arlington, VA.
- Proposition 227, English for the Children, California voter initiative (1998).
- Proposition 203, English for the Children, Arizona voter initiative (2000).
- Rivera, C. & Collum, E. (Eds.). (2006). *State assessment policy and practice for English language learners: A national perspective*. Mahwah, NJ: Lawrence Erlbaum and Associates.
- Roth, W. M. (2005). *Talking science: language and learning in science classrooms*. Lanham, MD: Rowman & Littlefield.
- Romaine, S. (1994). *Language in society: An introduction to sociolinguistics*. Oxford: Oxford University Press.
- Schleppegrell, M. J. (2007). The linguistic challenges of mathematics teaching and learning: A research review. *Reading and Writing Quarterly*, 23, 139–159.
- Sireci, S. G. (1997). Problems and issues in linking tests across languages. *Educational Measurement: Issues and Practice*, 16(1), 12–19.
- Sireci, S. G. (2009). Validity issues and empirical research on translating educational achievement tests: A review of the literature. In P. Winter (Ed.)...
- Sireci, S.G., Patsula, L., & Hambleton, R.K. (2005). Statistical methods for identifying flaws in the test adaptation process. In R. K. Hambleton, P. F. Merenda, and C.D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment*. (pp. 93–116). Mahwah, NJ: Lawrence Erlbaum Associates.

- Solano-Flores, G. (2006). Language, dialect, and register: Sociolinguistics and the estimation of measurement error in the testing of English language learners. *Teacher's College Record*, 108(11), 2354–2379.
- Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2002). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing*, 2(2), 107–129.
- Solano-Flores, G., Speroni, C., & Sexton, U. (2005). Test translation: Advantages and challenges of a socio-linguistic approach. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, April 11–15, 2005.
- Stansfield, C.W. (2003). Test translation and adaptation in public education in the USA. *Language Testing* 20(2), 187–207.
- Stansfield, C.W., (2008). A practical guide to sight translation of assessments. Retrieved 4/7/09, from the World Wide Web:
[http://www.eed.state.ak.us/tls/assessment/accomodations/Aguidetosighttranslationofassessments-ver35\(2\).pdf](http://www.eed.state.ak.us/tls/assessment/accomodations/Aguidetosighttranslationofassessments-ver35(2).pdf)
- Stansfield, C. W., & Bowles, M. (2006). Test translation and state assessment policies for English language learners. In C. Rivera & E. Collum (Eds.), *State assessment policy and practice for English language learners: a national perspective*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Stansfield C.W., & Kahl, S.R., (1998). Lessons learned from a tryout of Spanish and English versions of a state assessment. Paper presented at the annual meeting of the American Educational Research Association. San Diego, CA. ERIC Document Reproduction Service, ED 423–306.
- Tanzer, N.K. (2005). Developing tests for use in multiple languages and cultures: A plea for simultaneous development. In R.K. Hambleton, P.F. Merenda, & C.D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 235–263). Mahwah, NJ: Erlbaum.
- Valdés, G. (1996). *Con respeto: Bridging the distances between culturally diverse families and schools*. New York, NY: Teachers College Press.
- van deVijver, F., & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambleton & J.N. Zall (Eds.), *Advances in educational and psychological testing* (pp. 277–308). Boston: Kluwer Academic.
- van de Vijver, F. J. R. & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, and C.D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment*. (pp. 39–64). Mahwah, NJ: Lawrence Erlbaum Associates.

Viechnicki, G. (2003). Evidentiality in scientific discourse. Unpublished doctoral dissertation. University of Chicago.

Willis, G.B., (2005). Cognitive interviewing: A tool for questionnaire design. Thousand Oaks, CA: Sage.

Chapter 10: Impact of Language Complexity on the Assessment of ELL Students: A Focus on Linguistic Simplification of Assessment

Jamal Abedi
University of California at Davis

Abstract

Unnecessary linguistic complexity of assessments in content areas such as math or science may impact reliability and validity of assessments for ELL students. Based on the findings of studies on the effects of language on the assessment of ELL students, linguistic simplification of test items has been proposed. Under this approach, complex linguistic structure of the assessment that is judged by the content experts to be irrelevant to the content being assessed is identified to help develop a linguistically simplified version of the assessment. Results of some of the current studies using both original and simplified version of the assessments in math and science suggest that linguistic simplification of assessments improves reliability and validity of assessment for ELL students. This chapter presents a summary of studies on the impact of language on the assessment of ELL students and explains the concept of language simplification of test items.

Introduction

Inclusion of English language learner (ELL) students in national and state assessments is not only good assessment and accountability practice, it is mandated by law. The No Child Left Behind Act (NCLB), the most recent reauthorization of the Elementary and Secondary Act (ESEA) of 1965, holds states using federal funds accountable for academic achievement of all students, including ELL students. Thus in the past two decades the trend of an increased level of inclusion of ELL students in testing programs is clearly visible. Although this is a positive step toward equal education opportunity for all, it is certainly not sufficient. To truly provide equal educational opportunities for all students, appropriate instruction and reliable and valid assessments for ELLs should become major components of the nation's educational programs. Understanding assessment issues and improving the quality of measurement for ELL students is of utmost importance since ELLs are the fastest growing subpopulation in the United States. According to a recent report by the U.S. Government Accountability Office, about 5 million ELL students were enrolled in schools, representing approximately 10 percent of all public school students (GAO, 2006). Between 1990 and 1997, the number of United States residents born outside the United States increased by 30 percent, from 19.8 million to 25.8 million (Hakuta and Beatty, 2000). Consequently, fairness and validity issues relating to the assessment of ELL students are now among the top priorities in the national educational agenda.

Numerous interfering factors or nuisance variables affect the assessment of ELL students. Among these nuisance variables, the linguistic complexity of assessment is an important determination in the outcome of assessment of these students. Therefore, a common theme underlying assessment of ELL students is the language used in assessment.

Researchers and assessment experts argue that ELL students (as well as native English speakers with lower reading skills) may have difficulty with content-based assessments that have complex linguistic structures or contain language that may not be relevant to the content being assessed. For example, studies suggest that ELL students lag behind their non-ELL peers in almost all areas but particularly those areas with a high level of language demand (see, for example, Abedi, Leon, and Mirocha, 2003). However, it is difficult to explain such performance gaps. Would the low performance of an ELL student on a math test with complex linguistic structure be due to the lack of the student's content knowledge in math (instructional issues), or the student's difficulty understanding the question, or a combination of both (Abedi and Lord, 2001; Bielenberg and Wong Fillmore, 2004)?

To respond to this question and to test the impact of language factors on ELL students' performance outcomes, one may reduce the level of unnecessary (non-essential) linguistic complexity of the assessment—that is, complexity unrelated to the construct being assessed—and then examine the impact of such reduction of linguistic complexity on student performance. If the hypothesis that reducing linguistic complexity on assessment provides clearer interpretations of student performance is supported, then improvements on the outcome of assessments can be observed by using the less linguistically complex test items. Below we will first present a summary of studies on the concept of language simplification in general. Next, we will discuss studies that examine the impact of language complexity on the assessment of ELL students, and finally, we will explain how the language modification of assessment may help improve the validity of assessment for ELL students.

Language Simplification

The concept of language simplification applies to areas in which content other than language is being assessed (e.g., math or science) since the language construct may be irrelevant to the purpose of assessment. The main principle underlying language simplification is to reduce unnecessary linguistic complexity that is unrelated to the content of the test items. However, the judgment of whether language is related or unrelated to the target of assessment is arguable. Some researchers decide whether language is related or unrelated based on the judgment of content experts (see, for example, Abedi, Lord, and Plummer, 1997). To elaborate on the effects of complex language we present a summary of studies focusing on language simplification. However, it must be noted at this point that the findings of studies on language simplification are not quite conclusive; thus, future research should attempt to explain inconsistencies between findings of studies in this area.

A study by Thrush (2001) found that phrasal verbs may make texts less accessible to non-native speakers of English. The author indicated that some features in the English language are more challenging for non-native speakers of English or speakers of other English dialects. The study suggested that a modification of these features may make text more accessible to non-native speakers of English.

By comparing sections of an academic text with corresponding plain English "translations," Nevile (1990) showed that, although readability may be increased, linguistic simplification may affect the meaning of the test. Tweissi (1998) analyzed whether variations in amount and type of

linguistic simplification would create differences in the comprehension levels of Jordanian college students studying English. Students read different versions of a text and completed an achievement test. The results of Tweissi's study indicated that simplification positively affected students' reading comprehension. The findings of the study suggest that the type, not amount, of simplification affected comprehension. However, the author indicated that too much simplification was not necessarily helpful.

Leow (1997) indicates that the issue of the role and effects of simplification on learners' comprehension and intake remains controversial. For example, some see linguistic simplification as a positive intervention that could improve reliability and validity of assessments, and as a result it may help to reduce the performance gap between ELL and non-ELL students. However, others are concerned that the process of linguistic simplification may negatively affect the validity of assessments by changing the construct being measured since knowing the language of the content may be part of the construct. Shubert et al. (1995) studied the use of a restricted language, simplified English (SE), to write procedural documents for specific audiences. The study examined the effect of type (SE versus non-SE), passage (A versus B), and native language on the comprehensibility, identification of content location, and task completion of procedure documents for airplane maintenance. Results of this study suggested that SE significantly improves comprehension of more complex documents. It is not clear from the outcome of this study though, how strong the impact of SE was on the students' performance.

Thomas et al. (1992) found that training tools such as the simplified English analyses can lessen the reader's difficulties. Peterson (1990) provided a more operationally defined description of the concept of simplified English, mentioned several different techniques that have been developed, and then outlined a procedure by which any technical community can develop its own version for its specific "domain of discourse."

In a study on the impact of input modification, Oh (2001) found no major effect on student performance that could be attributed to language simplification. This finding was cross-validated with three test forms containing six English passages presented to 180 students who were divided into two proficiency levels. Linnell (1995) examined the extent to which linguistic modifications affect syntacticization, the reader's ability to understand the relationships between words to increase meaning. The results indicated that syntacticization was independent of the type of treatment given.

The summary presented above shows that the concept of language simplification has been the focus of attention in many different content areas and disciplines. We now present summaries on studies focusing particularly on the impact of language complexity on the assessment outcomes for non-native speakers of English.

Linguistic Complexity and Assessment

Recent studies on the assessment of ELL students have demonstrated that the unnecessary linguistic complexity of content-based assessments (e.g., math and science) is a likely source of measurement error, differentially impacting the reliability of assessment for the ELL subgroup. The linguistic complexity of test items as a source of construct-irrelevant variance may also influence the construct validity of the assessment for these students (Abedi, 2006). Results of

analyses of existing data from several locations nationwide show a substantial gap in reliability (internal consistency) and (concurrent) validity between ELL and non-ELL students on test items that are linguistically complex (Abedi, 2006; Abedi, Leon, and Mirocha, 2003).

Results of these analyses indicated that the gap in the reliability and validity coefficients reduces as the level of language demand of the assessment decreases. The reliability coefficients (alpha) for English-only (native speakers of English) students range from .898 for math to .805 for science and social science. For ELL students, however, alpha coefficients differ considerably across the content areas. In math, where language factors might not have much influence on performance, the alpha coefficient for ELL (.802) was slightly lower than the alpha for non-ELL students (.898). In English language arts, science, and social science though, the gap in alpha between non-ELL and ELL students was large. Averaging over English language arts, science and social science results, the alpha for native speakers of English was .808 as compared to an average alpha of .603 for ELL students. Thus, language factors introduce a source of measurement error affecting ELL students' test outcomes while they may not have much impact on students who are native or fluent speakers of English (for a more detailed description, see Abedi, 2006; Abedi, Leon, and Mirocha, 2003). As the level of linguistic complexity in science and social science tests decreased, the gap in the reliability coefficient was reduced substantially.

To examine the validity of assessments with complex linguistic structure, a multiple-group confirmatory factor analytical model was used. Several different hypotheses of invariance were tested. The study found major discrepancies in the structural relationships of assessment components between ELL and non-ELL students. For example, factor loadings of individual test items with the external criteria were lower for ELL students and the indices of fit of the model for ELL students were generally not as high as those for non-ELL students (see Abedi, Leon and Mirocha, 2003).

Findings from secondary analyses of data from the National Assessment of Educational Progress (NAEP) also clearly showed the impact of language on the assessment outcomes for ELL students. The results of this study (Abedi, Lord, and Plummer, 1997) suggested that ELL students had difficulty with the test items that were linguistically complex. The study also found that ELL students exhibited a substantially higher number of omitted/not-reached test items since it took them a much longer time to read and understand assessment questions. Results of this and similar studies led to the formation of the linguistic modification approach. Many linguistic features were identified that may slow reading speed, make misinterpretation more likely, and add to the reader's cognitive load, thus interfering with concurrent tasks. In one study, researchers found 48 linguistic features and grouped them into 14 general categories (Abedi, Lord, and Plummer, 1997). The impact of these linguistic features on the performance of ELL students in content-based areas (math and science) was then examined. A short description of each of these 14 categories along with research evidence of the impact of these features on assessment of ELL students is presented later in this chapter (see the section, "Features of Linguistic Complexity").

Impact of Language Factors

Below is a summary of some of the studies showing the impact of language factors on the assessment of ELL students, followed by a summary of studies demonstrating how linguistic simplification would help improve the accessibility of assessments for ELL students.

Abedi and Lord (2001) examined the effects of the linguistic modification approach with 1,031 eighth-grade students in Southern California. In this study, NAEP mathematics items were modified to reduce the complexity of sentence structures and to replace potentially unfamiliar vocabulary with more familiar words. Content-related terminologies (mathematical terms), were not changed. The results showed significant improvements in the scores of ELL students and also non-ELLs in low and average-level mathematics classes, but changes did not affect scores of higher performing non-ELL students. Among the linguistic features that appeared to contribute to the differences were low-frequency vocabulary and passive-voice verb constructions. These features contributed to the linguistic complexity of the text and made the assessment more linguistically complex for ELL students.

In another study, Abedi, Lord, and Hofstetter (1998) examined the impact of language simplification on the mathematics performance of English learners and non-English learners on a sample of 1,394 eighth graders in schools with high enrollments of Spanish speakers. Results showed that simplification of the language of items contributed to improved performance on 49 percent of the items; the ELL students generally scored higher on shorter/less linguistically complex problem statements. The results of this study also suggest that lower performing native speakers of English also benefited from the linguistic modification of assessment. A third study (Abedi, Lord, Hofstetter, and Baker, 2000) on a sample of 946 eighth graders found that, among four different accommodation strategies for ELL students, only the linguistically simplified English form narrowed the score gap between English learners and other students.

Another study (Abedi, Courtney, and Leon, 2003) examined 1,594 eighth-grade students using items from the NAEP and the Trends in International Math and Science Study (TIMSS). Students were given a customized English dictionary (words were selected directly from test items), a bilingual glossary, a linguistically modified test version, or the standard test items. Only the linguistically modified version improved the ELL students' scores without affecting the non-ELL students' scores.

Findings from studies by other researchers on different groups of students were consistent with the summaries presented above and suggest that linguistic simplification of assessment items provide a more valid and effective alternative to the conventional testing approach. Maihoff (2002) found linguistic simplification of content-based test items to be a valid and effective accommodation for ELL students. Kiplinger, Haug, and Abedi (2000) found linguistic modification of math items helped improve the performance of ELL students in math without affecting performance of non-ELL students. Rivera and Stansfield (2001) compared ELL performance on regular and simplified fourth- and sixth-grade science items. Although the small sample size in the Rivera and Stansfield study did not show significant differences in scores, the study did demonstrate that linguistic simplification did not affect the scores of English-proficient students, indicating that linguistic simplification is not a threat to score comparability.

However, researchers are not in agreement on the concept and application of language simplification of text used in the assessment and instruction of ELL students. For example, some researchers argue that to be successful academically, ELL students must be proficient in academic language, which is not necessarily the same as conversational fluency. Proficiency in academic language includes the knowledge of less frequent vocabulary and the ability to interpret and produce complex written language. Students should be able to understand complex linguistic structures in content areas such as science, social sciences, and mathematics (Bielenberg and Wong Fillmore, 2004; Celedon-Pattichis, 2003). According to these researchers, reducing the complexity of language that is required to perform such complex tasks may impede the goal of the assessment.

Conversely, Kopriva (2000) indicated that assessments designed for mainstream students may not allow ELL students to demonstrate what they know and can do. She provided a guide for improving large-scale academic assessments for ELL students (see also Kopriva, 2001a and 2001b). For example, she provides several recommendations to improve accessibility of test materials for ELL students. These recommendations include keeping item sentences or stems brief and straightforward; being consistent on paragraph structure; using the present tense and active rather than passive voice; avoiding rephrasing or rewording ideas to the extent possible; using pronouns in a very judicious manner; and using high-frequency rather than low-frequency words when possible (Kopriva, 2000, page 36).

Francis, Lesaux, Kieffer, and Rivera (2006) reviewed research on accommodations for English language learners, including studies on the validity and effectiveness of linguistic simplification of assessment as a form of accommodation for ELL students. By reviewing literature and comparing effect sizes of the linguistically modified version of the tests, Francis et al. indicated that the findings supporting the effectiveness of simplified English are weak. The authors also added that “While it is possible that the effects of Simplified English vary according to variables such as grade level, content area, and the nature of the assessment, the evidence does not currently support this conclusion” (p 26).

Findings from Francis et al.’s reviews raise serious concerns regarding methodological issues in the studies examining the impact of language simplification on the assessments for ELLs. There are several concerns with the design of some of the studies that were reviewed. Among these issues are

- small (and non-representative) samples of ELL students in the study
- lack of control of extraneous variables that could affect performance of ELL students, such as the ELLs’ level of English proficiency
- lack of an operational definition of the linguistic simplification approach

For example, in some of these studies the number of ELL students was so small that there were not enough subjects to validly assess the impact of linguistic simplification. More importantly, the various studies applied the concept of linguistic simplification of assessment quite differently. That is, there was no uniform approach to linguistic modifications across the studies. As indicated earlier, Abedi, Lord, and Plummer (1997) introduced 48 linguistic features (later combined into 14 categories) as indications of linguistic complexity in the assessments; others

used other linguistic features or even a simple editorial process in simplifying the language of assessments for ELL students.

Overall, the research evidence shows the impact of linguistic complexity as a major source of measurement error on the assessment results for ELL students. Research findings also suggest that reducing the level of unnecessary linguistic complexity of assessments may help improve the validity of assessment for these students. Improvement in the validity of assessment is due to reduction of unnecessary linguistic complexity of assessment as a source of construct-irrelevant variance in the assessment of ELL students. On the other hand, some people argue that reducing the level of complexity of academic contents may change the construct being taught and being assessed. However, the complex linguistic structures that are related to the content of assessment and instruction should be distinguished from the unnecessary linguistic complexity of text in both assessment and instruction. To illustrate our point, we use a few released test items to show how unnecessary linguistic complexity may hinder students' ability to provide a valid picture of what they know and can do. We first present these items in their original form and then propose some linguistic revisions that help facilitate students' understanding of the text. In these revisions, different linguistic features that contributed to the complexity of assessment questions were modified.

Example 1

Original

A certain reference file contains approximately six billion facts. About how many millions is that?

- (a) 6,000,000
- (b) 600,000
- (c) 60,000
- (d) 6,000
- (e) 600

Revised

Mack's company sold six billion hamburgers. How many millions is that?

- (a) 6,000,000
- (b) 600,000
- (c) 60,000
- (d) 6,000
- (e) 600

In this example, potentially unfamiliar, low-frequency lexical terms (certain, reference, file) were replaced with more familiar, higher frequency terms (company, hamburger).

Example 2

Original

Raymond must buy enough paper to print 28 copies of a report that contains 64 sheets of paper. Paper is only available in packages of 500 sheets. How many whole packages of paper will he need to buy to do the printing?

Revised

Raymond has to buy paper to print 28 copies of a report. He needs 64 sheets of paper for each report. There are 500 sheets of paper in each package. How many whole packages of paper must Raymond buy?

In this example, the original version contains information in a relative clause, while the revised item contains the same information in a separate, simple sentence.

Features of Linguistic Complexity

We now present a summary of the 14 linguistic features of test items that our research identified as slowing reading speed, making misinterpretation more likely and adding to the reader's cognitive load, thus interfering with concurrent tasks. Indexes of language difficulty include word frequency/familiarity, word length, and sentence length. Other linguistic features that may cause difficulty for readers include passive-voice constructions, comparative structures, prepositional phrases, sentence and discourse structure, subordinate clauses, conditional clauses, relative clauses, concrete versus abstract or impersonal presentations, and negation.

Word Frequency/Familiarity

Word frequency was an element in early formulas for readability (Dale and Chall, 1948; Klare, 1974). Words that are high on a general frequency list for English are likely to be familiar to most readers because they are encountered often. Readers who encounter a familiar word will be likely to interpret it quickly and correctly, spending less cognitive energy analyzing its phonological component (Adams, 1990; Chall, et al., 1990; Gathercole and Baddeley, 1993). On a test with math items of equivalent mathematical difficulty, eighth-grade students scored higher on the versions of items with vocabulary that was more frequent and familiar; the difference in score was particularly notable for all students (ELL and non-ELL) in low level math classes (Abedi, Lord, and Plummer, 1997).

Word Length

As frequency of occurrence decreases, words tend to be longer. Accordingly, word length can serve as an index of word familiarity (Zipf, 1949; Kucera and Francis, 1967). Additionally, longer words are more likely to be structurally complex. In one study, language minority students performed better on math test items with shorter word lengths than items with longer word lengths (Abedi, Lord, and Plummer, 1997).

Sentence Length

Sentence length serves as an index for syntactic complexity and can be used to predict comprehension difficulty. As we have already explained, to understand a sentence, one must be

able to understand the relationships between the words. The more words there are in a sentence, the more relationships must be understood. In addition, linguistic definitions of complexity based on the concept of word depth correlate with sentence length. Word depth is a measure of syntactic complexity based on a tree diagram of the linguistic structure of a sentence (Bormuth, 1966; MacGinitie and Tretiak, 1971; Wang, 1970)

Passive-voice Constructions

People find passive-voice constructions more difficult to process than active-voice constructions (Forster and Olbrei, 1973) and more difficult to remember (Savin and Perchonock, 1965; Slobin, 1968). Furthermore, passive constructions can pose a particular challenge for non-native speakers of English (Celce-Murcia and Larsen-Freeman, 1983) because a piece of information is missing—the doer of the action in the sentence. As a result passive-voice constructions are usually not considered the best form of expression in Standard Formal English. However, passive-voice constructions tend to be used less frequently in conversation than in some formal writing such as scientific writing (Celce-Murcia and Larsen-Freeman, 1983). In one study, eighth-grade students (native and non-native English speakers) were given equivalent math items with and without passive-voice constructions; students in average math classes scored higher in the versions without passive constructions (Abedi et al., 1997).

Long Noun Phrases

Noun phrases with several modifiers have been identified as potential sources of difficulty in test items (Spanos et al., 1988). Long noun phrases typically contain more semantic elements and are inherently syntactically ambiguous; accordingly, a reader's comprehension of a text may be impaired or delayed by problems in interpreting long noun phrases (Halliday and Martin, 1994; Just and Carpenter, 1980; King and Just, 1991; MacDonald, 1993). Romance languages such as Spanish, French, Italian, and Portuguese make less use of compounding than English does, and when they do employ such a device, the rules are different. Consequently, students whose first language is a Romance language may have difficulty interpreting compound nominals in English (Celce-Murcia and Larsen-Freeman, 1983).

Long Question Phrases

Longer question phrases occur with lower frequency than short question phrases, and low-frequency expressions are in general harder to read and understand (Adams, 1990).

Comparative Structures

Comparative constructions have been identified as potential sources of moderate-level difficulty for non-native speakers of English (Jones, 1982; Spanos, et al., 1988) and for speakers of non-mainstream dialects (Orr, 1987, but see also Baugh, 1988).

Prepositional Phrases

Students may find interpretation of prepositions difficult (Orr, 1987; Spanos et al., 1988). Languages such as English and Spanish may differ in the ways that motion concepts are encoded using verbs and prepositions (Slobin, 1968).

Sentence and Discourse Structure

Even shorter sentences can be problematic if their structure is complex. Two sentences may have the same number of words, but one may be more difficult than the other because of the syntactic structure or discourse relationships among sentences (Freeman, 1978; Finegan, 1978; Larsen, Parker, and Trenholme, 1978).

Subordinate Clauses

Subordinate clauses may contribute more to complexity than coordinate clauses (Hunt, 1965, 1977; Wang, 1970; Botel and Granowsky, 1974) because they require that one understand the hierarchy of meaning in the sentence. Separate sentences, rather than subordinate *if* clauses, may be easier for some students to understand (Spanos et al., 1988).

Conditional Clauses

Conditional clauses and initial adverbial clauses have been identified as contributing to difficulty (Spanos et al., 1988; Shuard and Rothery, 1984). The semantics of the various types of conditional clauses in English are subtle and hard to understand even for native speakers (Celce-Murcia and Larsen-Freeman, 1983). Non-native speakers may omit function words (such as *if*) and may employ separate clauses without function words. Separate sentences, rather than subordinate *if* clauses, may be easier for some students to understand (Spanos et al., 1988). In fact, some languages do not allow sentences with the conditional clause in the final position of the sentence (Haiman, 1985). Consequently, sentences with a sentence-final conditional clause may cause difficulty for some non-native speakers.

Relative Clauses

Since relative clauses are less frequently used in spoken English than in written English, some students may have had limited exposure to them. In fact, Pauley and Syder (1983) argue that the relative clauses in literature differ from those in spoken vernacular language (Schachter, 1983).

Concrete Versus Abstract or Impersonal Presentations

Studies show that students perform better when problem statements are presented in concrete rather than abstract terms (Cummins et al., 1988). Information presented in narrative structures tends to be understood and remembered better than information presented in expository text (Lemke, 1986).

Negation

Mestre (1988) observed that a considerable number of research studies indicate that sentences containing negations (e.g., *no*, *not*, *none*, *never*) are harder to comprehend than affirmative sentences. One of the reasons for this may be because there is a lack of parallelism in the use of negation between English and other languages. In Spanish, for example, double negative constructions retain a negative meaning instead of reverting to an affirmative meaning, as would be the case in grammatically correct English (Mestre, 1988, p. 212). Mestre (1988, p. 213) found that Spanish-speaking students processed negations from left to right, which works for natural discourse but does not always work for mathematics texts.

Determining the Comparability of Scores from Linguistically Simplified Assessments

The principle underlying linguistic modification of assessment is based on the premise that the unnecessary linguistic complexity of content-based assessment (e.g., math and science) is considered a construct-irrelevant source of variance which undermines the validity of assessment for ELL students. That is, scores for many ELL students under-represent these students' knowledge and skills. As elaborated in the studies using a linguistic modification approach (see for example, Abedi and Lord, 2001; Abedi, Lord, and Plummer, 1997), tests with simplified language are usually constructed with the advice of content experts to avoid changing the construct the items are intended to address, which is a first step in maintaining comparability between the simplified language test and the source (general) test. The increases in ELL student scores when linguistic complexity is reduced, as found in the studies cited above, and the relationship between linguistic complexity and reliability support the idea that ELL students' scores from linguistically modified assessments are more comparable, in terms of representing knowledge and skills, to English proficient students' scores on the general assessment than ELL students' scores on the general assessment are (Abedi, Lord, and Plummer, 1997; Abedi, Leon, and Mirocha, 2003). Several studies have been conducted to investigate the validity (comparability) of the inferences made from the linguistically simplified test scores with the scores from the original tests. In these studies the validity of linguistically simplified assessments has been thoroughly examined. These studies are based on the assumption that the linguistically simplified assessment is considered valid if it does not alter the construct being measured. Thus, any indication of a significant change in the performance of non-ELL students taking the simplified version of the assessment may jeopardize the validity of this approach.

For examining the validity (comparability) of linguistically modified assessments, in many studies both ELL and non-ELL students were tested under both original and linguistically simplified assessments. The results of these studies consistently indicated that performance of non-ELL students was not affected by linguistic simplification of assessments (see for example, Abedi and Lord, 2001; Abedi, Courtney, and Leon, 2003; Abedi, Lord, and Hofstetter, 1998; Abedi, Lord, and Plummer, 1997; Abedi, Lord, Hofstetter, and Baker, 2000; Maihoff, 2002; Kiplinger, Haug, and Abedi, 2000; Rivera and Stansfield, 2001). Findings of these studies that have been cross-validated clearly support the notion of comparability of scores from the linguistically simplified test scores and those from the original tests.

Summary and Discussion

Research on the assessment of ELL students clearly suggests that language factors affect performance outcomes of ELL students. More specifically, unnecessary linguistic complexity of assessments may undermine the validity of assessment for ELL students. Assessments with complex linguistic structures may confound students' performance in content areas with their language abilities. That is, students with lower levels of language proficiency may not understand assessment questions; thus, their performance may not, in spite of their knowledge in content areas, be accurately assessed and consequently the assessment may not provide a clear picture of what they know and can do.

To control for the impact of language factors as a source of construct irrelevance, researchers propose linguistic modification of assessment. This involves having the complex linguistic

structure judged by experts and the elements not relevant to the content simplified. This approach has been proposed for content areas such as math and science where language is not the target of measurement. Therefore, this approach may not quite apply to areas such as reading, where language is construct-relevant for the assessment.

Results of studies comparing original test items that are linguistically complex with the linguistically simplified version of the same assessment revealed that ELL students performed better under the simplified version of assessment. These findings suggest that the language of test items should be carefully examined for any sign of linguistic complexity that may not be related to the content. It must be noted at this point, however, that care must be taken to avoid any changes in the target construct during the process of language modifications.

References

- Abedi, J. (2006). Language Issues in Item-Development. In Downing, S. M. and Haladyna, T. M. *Handbook of Test Development* (Ed.). New Jersey: Lawrence Erlbaum Associates, Publishers.
- Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219–234.
- Abedi, J., Courtney, M., & Leon, S. (2003). *Effectiveness and validity of accommodations for English language learners in large-scale assessments* (CSE Tech. Rep. No. 608). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Abedi, J., Leon, S., Mirocha, J. (2003). *Impact of students' language background on content-based data: Analyses of extant data* (CSE Tech. Rep. No. 603). Los Angeles: University of California: Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.
- Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance* (CSE Tech. Rep. No. 478). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Abedi, J., Lord, C., & Plummer, J. (1997). *Language background as a variable in NAEP mathematics performance* (CSE Tech. Rep. No. 429). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice, 19*(3), 16–26.
- Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.
- Baugh, J. (1988, August). [Review of the article Twice as less: Black English and the performance of black students in mathematics and science]. *Harvard Educational Review, 58*(3), 395–404.
- Beilenberg, B & Wong Fillmore, L. (2004). ELLs and high stakes testing: Enabling students to make the grade. *Educational Leadership, 62* (4).
- Bormuth, J. R. (1966). Readability: A new approach. *Reading Research Quarterly, 1*(3), 79–132.
- Botel, M., & Granowsky, A. (1974). A formula for measuring syntactic complexity: A directional effort. *Elementary English, 1*, 513–516.

- Celce-Murcia, M., & Larsen-Freeman, D. (1983). *The grammar book: An ESL/EFL teacher's book*. Rowley, MA: Newbury House.
- Celedon-Pattichis, S. (2003). Construction meaning: Think-aloud protocols of ELLs on English and Spanish word Problems. *Education for Urban Minorities*, (2(2), 74–90.
- Chall, J. S., Jacobs, V. S., & Baldwin, L. E. (1990). *The reading crisis: Why poor children fall behind*. Cambridge, MA: Harvard University Press.
- Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology*, 20, 405–438.
- Dale, E., & Chall, J. S. (1948). A formula for predicting readability. *Educational Research Bulletin*, 27,11-20, 28, 37–54.
- Finegan, E. (1978, December). *The significance of syntactic arrangement for readability*. Paper presented to the Linguistic Society of America, Boston, MA.
- Forster, K. I., & Olbrei, I. (1973). Semantic heuristics and syntactic trial. *Cognition*, 2(3), 319–347.
- Francis, D. J., Lesaux N., Kieffer, M. Rivera, H. (2006). Practical guidelines for the education of English language learners. Research –based recommendations for the use of accommodations in large-scale assessments. Houston: Texas Institute for Measurement, Evaluation, and Statistics at the University of Houston for the Center on Instruction
- Freeman, G. G. (1978, June). *Interdisciplinary evaluation of children's primary language skills*. Paper presented at the World Congress on Future Special Education, First, Stirling, Scotland. (ERIC Document Reproduction Service No. ED157341)
- GAO (2006). No Child Left Behind Act. Assistance from education could help states better measure progress of students with limited English proficiency: Washington, DC: United States Government Accountability Office.
- Gathercole, S. E., & Baddeley, A. D. (1993). *Working memory and language*. Hillsdale, NJ: Erlbaum.
- Haiman, J. (1985). *Natural syntax: Iconicity and erosion*. New York: Cambridge University Press.
- Hakuta, K., & Beatty, A. (Eds.). (2000). *Testing English-language learners in U.S. schools*. Washington, DC: National Academy Press.
- Halliday, M. A. K., & Martin, J. R. (1994) *Writing science: Literacy and discursive power*. Pittsburgh, PA: University of Pittsburgh Press.

- Hunt, K. W. (1965). *Grammatical structures written at three grade levels* (Research Rep. No. 3). Urbana, IL: National Council of Teachers of English.
- Hunt, K. W. (1977). Early blooming and late blooming syntactic structures. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging*. Urbana, IL: National Council of Teachers of English.
- Jones, P. L. (1982). Learning mathematics in a second language: A problem with more and less. *Educational Studies in Mathematics*, 13, 269–87.
- Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixation to comprehension. *Psychological Review*, 87, 329–354.
- King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, 30, 580–602.
- Kiplinger, V. L., Haug, C. A., & Abedi, J. (2000, April). Measuring math – not reading – on a math assessment: A language accommodations study of English language learners and other special populations. Presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly*, 10, 62–102.
- Kopriva, R.J. (2000). *Ensuring Accuracy in Testing for English Language Learners: A Practical Guide for Assessment Development*. Council of Chief State School Officers Publications, Washington D.C.
- Kopriva, R.J. (2001a). Construction mechanisms influencing the appropriateness of achievement tests for students from diverse linguistic backgrounds. Educational Testing Service, Princeton, New Jersey.
- Kopriva, R.J. (2001b). Identification of salient selection review procedures that determine the adequacy of established testing instruments for language and cultural minority students. Educational Testing Service, Princeton, New Jersey.
- Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day English*. Providence, RI: Brown University Press.
- Larsen, S. C., Parker, R. M., & Trenholme, B. (1978). The effects of syntactic complexity upon arithmetic performance. *Educational Studies in Mathematics*, 21, 83–90.
- Lemke, J. L. (1986). *Using language in classrooms*. Victoria, Australia: Deakin University Press.

- Leow, Ronald P (1997). Simplification and Second Language Acquisition. *World Englishes*, v16 n2 p291–96
- Linnell, Julian (1995) Can Negotiation Provide a Context for Learning Syntax in a Second Language? *Working Papers in Educational Linguistics*, v11 n2 p83-103.
- MacDonald, M. C. (1993). The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language*, 32, 692–715.
- MacGinitie, W. H., & Tretiak, R. (1971). Sentence depth measures as predictors of reading difficulty. *Reading Research Quarterly*, 6, 364–377.
- Maihoff, N. A. (2002, June). *Using Delaware data in making decisions regarding the education of LEP students*. Paper presented at the Council of Chief State School Officers 32nd Annual National Conference on Large-Scale Assessment, Palm Desert, CA.
- Mestre, J. P. (1988). The role of language comprehension in mathematics and problem solving. In R. R. Cocking & J. P. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics* (pp. 200–220). Hillsdale, NJ: Erlbaum.
- Nevile, Maurice R (1990). Translating Texts into Plain English: The Cost of Increased Readability. *Open Letter: Australian Journal for Adult Literacy Research and Practice*, v1 n2 p27–38 1990
- No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).
- Oh, Sun-Young (2001). Two Types of Input Modification and EFL Reading Comprehension: Simplification versus Elaboration.
- Orr, E. W. (1987). *Twice as less: Black English and the performance of black students in mathematics and science*. New York: W. W. Norton.
- Pauley, A., & Syder, F. H. (1983). Natural selection in syntax: Notes on adaptive variation and change in vernacular and literary grammar. *Journal of Pragmatics*, 7, 551–579.
- Peterson, D. A. T. (1990). Developing a Simplified English Vocabulary. *Technical Communication*, v37 n2 p130–33
- Rivera, C., & Stansfield, C. W. (2001, April). The effects of linguistic simplification of science test items on performance of limited English proficient and monolingual English-speaking students. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.
- Savin, H. B., & Perchonock, E. (1965). Grammatical structure and the immediate recall of English sentences. *Journal of Verbal Learning and Verbal Behavior*, 4, 348–353.

- Schachter, P. (1983). *On syntactic categories*. Bloomington: Indiana University Linguistics Club.
- Shuard, H., & Rothery, A. (Eds.). (1984). *Children reading mathematics*. London: J. Murray.
- Shubert, Serena K.; And Others (1995). The Comprehensibility of Simplified English in Procedures. *Journal of Technical Writing and Communication*, v25 n4 p347–69
- Slobin, D. I. (1968). Recall of full and truncated passive sentences in connected discourse. *Journal of Verbal Learning and Verbal Behavior*, 7, 876–881.
- Spanos, G., Rhodes, N. C., Dale, T. C., & Crandall, J. (1988). Linguistic features of mathematical problem solving: Insights and applications. In R. R. Cocking & J. P. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics* (pp. 221–240). Hillsdale, NJ: Erlbaum.
- Thomas, Margaret; And Others (1992). Learning to Use Simplified English: A Preliminary Study. *Technical Communication*, v39 n1 p69–73
- Thrush, Emily A (2001). Plain English? A Study of Plain English Vocabulary and International Audiences. *Technical Communication: Journal of the Society for Technical Communication*, v48 n3 p289–96 Aug 2001
- Tweissi, Adel I. (1998). The Effects of the Amount and Type of Simplification on Foreign Language Reading Comprehension. *Reading in a Foreign Language*, v11 n2 p191–204
- Wang, M. D. (1970). The role of syntactic complexity as a determiner of comprehensibility. *Journal of Verbal Learning and Verbal Behavior*, 9, 398–404.
- Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley.

Chapter 11: Alternative Formats: A Review of the Literature

Karen E. Barton,
CTB/McGraw-Hill

Phoebe C. Winter
Pacific Metrics

Introduction

The inclusion of all students in statewide assessments and accountability systems has led states to create alternative options to meet diverse student needs. As discussed elsewhere in this handbook, some of those options include computer-based testing, translated assessments, and assessments that use clarified language. In addition to these, there are typical approaches most states utilize to insure wide-ranging participation of their student population. For example, students with or without disabilities or language barriers participate in the general assessments with or without accommodations, or students with severe cognitive disabilities participate in alternate assessments that are based on grade level standards but do not cover the breadth, depth, or complexity of those same grade level standards as the general assessments. More recently, some states have accepted the challenge of creating an alternate assessment based on modified achievement standards (AA-MAS) for those students who consistently do not perform well on the general assessment and are not eligible for the alternate assessment. As is clear to the purpose of this handbook, the challenge in utilizing various forms of assessment and various ways in which students can participate in the overarching accountability system is to measure a standard set of skills or criteria fairly, reliably, validly, and comparably regardless of form or approach.

What Is an Alternative Format?

One approach that is not widely used or typical is that of alternative formats. Further, there is little research/literature on evaluating the comparability of scores from alternative formats to those from the general test; this chapter includes the literature when it is available but relies more on literature re properties of good alternative formats (such as it is). In current U.S. Department of Education terminology, alternative formats are a type of alternate assessment based on grade level achievement standards (AA-GLAS).²⁸ These are not to be confused with assessments based on alternate or modified achievement standards (AA-AAS and AA-MAS). Alternative formats are sometimes used to allow students with disabilities or with English language barriers to participate in grade level measures of performance against the same grade level standards at the same breadth, depth, and complexity as their same-grade peers, and against the same (unmodified) achievement standards (or performance level descriptors). These approaches provide alternative participation options to students for whom all other typical assessment

²⁸ The idea of alternative formats of educational assessment instruments is not tied to current law and regulations; therefore we will use the more general term “alternative format” in this chapter rather than “AA-GLAS,” except when AA-GLAS is referred to by a document being cited.

approaches are not accessible by the student or that are not appropriate for the student given their individual needs, even with accommodations, translated forms, computer-based assessments, and so forth. Therefore, results from alternative formats are used in aggregate with the general assessment for school and district accountability purposes. It is not clear how many states offer the option of an alternative format. In a 2007 survey, of the 30 states responding, 4 indicated that they used an alternative format (Winter, May 2007) and at least 2 of the non-responding states used one at the time (personal communication, Winter, 2009).

Who Participates in Alternative Formats?

In the state described by Barton and Winter (2009, this document), most students who participated in the alternative formats are classified as English language learners. Their eligibility was based on their English language acquisition (performance on English language assessments, time in country, and so forth). Other participating students were eligible due to some inability to access paper-based assessments, such as a student recently blinded or with physical disabilities that require the student to have assistive technologies not readily available for testing. The following example of a state's selection criteria for inclusion in its AA-GLAS from Weiner (2006) illustrates categories of student disability that merit consideration of an alternative format:

- “Students unable to ‘maintain sufficient concentration’ to participate in standard testing, even with test accommodations, as a consequence of a severe emotional disability, traumatic brain injury, autism or Asperger’s Syndrome, or other disability or combination of disabilities.
- Students for whom the ‘demands of a prolonged test administration’ would present a significant challenge, as a consequence of a health-related, multiple physical, or other disability.
- Students who require ‘more time than is reasonable or available’ for testing, even with the allowance of extended time, as a consequence of cerebral palsy, deaf-blindness, or a significant motor, communication, or other disability.
- Students for whom the ‘format of the standard test is inappropriate,’ and the necessary accommodations are unavailable or would ‘give away’ or hint at the answers.
- Students who do not have significant cognitive disabilities, but whose disabilities result in other ‘unique and significant challenges’ to taking the standard test.” (p.3)

What Does an Alternative Format Look Like?

In general, alternative formats are any assessments built as *an alternative to the general test* in which the method of assessment delivery, scoring, and/or response is less standardized than on traditional, paper-and-pencil or computer-based tests. Most alternative formats are akin to performance assessments, which are not new in their approach to conducting assessments. Portfolios, “authentic” assessment, checklists, teacher observations—all have been used in various forms in various arenas, including outside education, as methods for collecting performance data. For example, teachers who wish to become certified by the National Board for Professional Teaching Standards (NBPTS) must submit a portfolio as part of their certification process (NBPTS, 2009). Performance tasks and portfolios are utilized for nursing and other medical licensure/certification and educational assessments (Meister, Heath, Andrews, and Tinggen, 2002). The emergence of digital or electronic portfolios for pre-service teacher

evaluations, high school graduation and other classroom-based assessment practice, to include higher education settings, is on the rise and may continue to grow as schools become more equipped digitally (Lambert, DePaepe, Lambert, and Anderson, 2007; Knight, Hakel, and Gromko, 2008; Love, McKean, and Gathercoal, 2004).

Research on performance assessment in K–12 education was relatively active until the mid 1990s, when states began expanding their testing programs using primarily paper-and-pencil tests. The literature was well debated, particularly when such formats are used for high-stakes purposes, such as graduation and accountability (validity: Linn and Baker, 1996; authenticity: Wiggins, 1989; equity: Darling-Hammond, 1994). In each type of alternative (e.g., portfolios, checklists, observations) the data are not standardized, so that the data, based on “evidence” from each skill for each student, scored by different judges, can vary almost infinitely. Such variations can impact reliability (Koretz, Stecher, Klein, McCaffrey, 1994), which is why the “scoring of portfolios and performance assessments requires disciplined judgments. It is the quality of these judgments, gauged in terms of a shared standard or level of aspiration implicit in those judgments, that is of paramount concern” (LeMahieu, Gitomer, and Eresh, 1995, p. 11). Alternative formats do not provide the luxury of all students receiving the same items in standard fashion as on a general assessment. Herein are the topics of debate: If the assessments or tasks are not standard, how are students’ scores compared? What of the reliability for such a small number of tasks or observations? How generalizable are the results and how can so few pieces of evidence capture the skills necessary to demonstrate progress against a set of standards? The assessments do, however, offer the flexibility to students to demonstrate and for teachers to evaluate, a pool of skills linked to state standards (as was found in the states participating in this study) in ways that are unobstructed by student access issues, such as language or physical limitations.

Establishing Comparability

A challenge for states that choose to use any assessment variation, including alternative formats, for high-stakes accountability purposes is to assure the scores remain comparable regardless of assessment approach (U.S. Department of Education, July 2007, revised January, 2009). In this section, the focus of comparability is between the alternative format and the general assessment. The alternative format may reflect different approaches, such as portfolios, observations of student work in the classroom, or checklists. The general assessment typically contains multiple-choice and often constructed-response items that are administered and scored in a standardized way across students. Both the general assessment and the alternative format are purposed to measure the *same* skills and standards at the *same* level of complexity and against the *same* level of achievement required for proficiency. However, with such variability in data from the alternative formats, the item level or task level analyses, score level analyses, and so forth that are often found in comparability studies of scores from other test variations (e.g., computer-based) will be limited by the design of the test and data availability. These limits might be improved when some standardization is imposed.

Design Issues

It is gravely important that the design of the alternative format is well aligned to and suited for the collection of supporting evidence for comparability. It is quite difficult to collect certain

types of evidence if the data are not readily available. However designing the assessment process with the data requirements incorporated is an essential first step. For example, to maximize the location along the continuum of content comparability for alternative and general tests, the alternative format should be built such that the content area, content standards, test blueprints, and variations in complexity match the general test as much as possible. This will provide data for alignment studies about the test design to the content standards, across assessments. Because the alternative format does not contain “test items,” it will be important for the process to allow collection of student work to serve as evidence against the alignment of the work to the standards and blueprint of the general test.

Such attention prior to “testing” will also help serve as an audit trail for administration issues and policy, scoring, rater validity, and so forth. For example, the training materials utilized for administration and scoring might specifically encourage data collection on the alternative format that targets aligned skills and rubrics for assuring similar complexity levels of standards assessed. The finer the grain of data collected, the more evidence can be collected, evaluated, and put forth toward comparability. Some states, such as Alabama, that use portfolio-based alternate assessments provide guidelines for the types of and minimum number of evidence or tasks to include in the portfolio, and examples of evidence that exemplify the standards to assure the tasks are “aligned” to or at least cover the standards to be assessed. Some states, such as Oregon, provide standard tasks from which teachers can select as evidence, rather than allowing teachers to choose any task for each student. Such parameters will only serve to support the comparability process necessary for alternative assessments used for adequate yearly progress (AYP).

Evaluating Comparability

The discussion of evaluating comparability is referenced to the four questions found in Chapter 1 (Winter, 2009).

Validity of inferences: In terms of accessibility, the purpose of the alternative format is to provide a more accessible assessment and therefore a more valid inference about student ability than can the general test (Question 1). The alternative format in and of itself should present a suitable method for students to demonstrate their ability. Accessibility is often addressed on the general assessment with accommodations and/or principles of universal design in testing (Thompson & Thurlow, 2002). Accessibility is therefore not only about how well students can access information or respond to and perform on an assessment, but it also includes how well the assessment is able to elicit student responses that have minimal measurement error (Barton, 2007). While a typical assessment can minimize error through attention to accommodations policies and provisions, as well as universal design elements, navigability, and so forth, these are not easily managed when the assessment is not standardized in the traditional way. With an alternative format that varies by student, it is hard to know how well each of the assessment components provides access in both senses—accessing ability well and providing student access to the information and ways to respond. For example, we can only assume that the choices that students and/or teachers make in creating and assigning tasks and the methods by which students respond represent the best way to evoke the student’s unobstructed (accessible) performance. Supporting the inference made from an alternative format is one reason that design principles,

criteria for inclusion in the assessment, solid training of administrators and scorers, and monitoring of implementation are part of establishing comparability (see section above).

Alignment of content standards: Similarly, if the alternative format is intended to measure the same skills aligned to the grade level standards in the same proportion as specified in the general test blueprint, the choices made on the alternative format components should be well aligned to the grade level content standards. In many cases, alternative formats have the same structural (although not content) characteristics of an AA-AAS. Recent work in establishing alignment for high-stakes assessments centered around alternate assessments based on alternate achievement standards which may be as varied as alternative formats may be helpful in structuring alignment review procedures for alternative formats (question 2). For example, two alignment methodologies have emerged recently to establish the alignment of alternate assessments: Links to Academic Learning (LAL) by the National Alternate Assessment Center (2007) and a variation on the Webb Alignment System by Tindal (2006). Both approaches provide avenues for establishing alignment across a variety of evidence or tasks within the alternate assessment to the grade level standards and can be modified to work in the context of an alternative format.

Comparability of achievement levels: Comparability at the achievement score level requires first that the cut scores associated with the achievement levels are set based on the construct-based requirements for meeting each achievement level on the general test. Widely used, technically sound standard setting processes can be used to set standards on the alternative format. For example, the Body of Work or Profile Sorting (Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006) methods would be appropriate for use with portfolios and other non-standardized assessments. In each method, either samples of portfolios of student work (Body of Work) or profiles of student performance across tasks (Profile Sorting) are evaluated and sorted by judges into proficiency or achievement levels. In any method, the standard setting process for alternative formats should assure the definition of performance at each level for the alternative and general assessments are comparable.

Similarly reliable results: Reliability in the results is a difficult criterion to establish for alternative formats. Reliability coefficients based on internal consistency assume independence within the data. Clearly, for alternative formats that require teacher selection of tasks and evaluation or scoring across tasks, the data from each individual task are not independent, given that the teacher is a constant across data. Traditional reliability indices, therefore, can only be used with caution. In addition, the reliability at the item or task level is difficult to collect when the scoring of each task varies both within and across student examinees, particularly given that there is no simple way to establish reliability due to the test (over all tasks) or to the scorer (Raju, 1991). However, the reliability of holistic ratings by the teacher can be estimated if the teacher is given a standard rubric for each task, as well as the inter-rater agreement for components of alternative formats that are scored by more than one evaluator or teacher.

Classification consistency can be estimated as well for students within achievement levels. It is first important to understand that most consistency measures depend on the reliability assumptions of independence of measures, as discussed. Further, such measures as the Subkoviak (1988), Livingston-Lewis (1995), and even Cronbach (1951) rely on those assumptions and reliability estimates. With those considerations in mind, some alternative format

assessments, if attention is paid to data collection requirements prior to collecting and scoring student work, can apply such statistics for evaluating classification consistency.

Using Social Moderation to Evaluate Comparability

Finally, comparability between two forms or variations of assessments is often established through various methods of linking. Mislevy (1992, p. 63) states, “The more assessments arouse different aspects of students’ knowledge, skills, and attitudes, the wider the door opens for students to perform differently in different settings.” (In-depth discussion of the variations in linking are found in Mislevy [1992] and Linn [1993].) Given the variability in assessment designs and statistical characteristics between alternative formats and general assessments, as well as the assumptions required for various linking methods, it seems the most viable solution to establishing comparability is via social moderation. According to Linn (1993), “social moderation substitutes requirements for developing professional consensus regarding standards and exemplars of performance meeting those standards for the more familiar measurement and statistical requirements associated with” statistically based linking methods. The use of social moderation for establishing links across grades in vertical scales has been well developed and utilized (Ferrara, 2003, Ferrara, et. al, 2005; Huynh, et. al, 2005; Lewis and Haug, 2005; Lissitz and Huynh, 2003). Because social moderation is a process that requires judgments and consensus, training is critical so that raters, judges, teachers (as in the case of alternative formats) have a consensual understanding of how to evaluate and score the tasks (portfolios, observations, and the like) (Linn, 1993).

Use of statistical techniques to establish comparability on alternative formats and general assessments might be considered, however cautiously. Even if scores on two assessments can be linked (from least to most rigorous: statistically moderated, calibrated, equated) resulting in a common metric across the assessments, the inferences of comparability may be far from interchangeable as one would expect when forms are equated, for example. Per Linn (1993), even with statistical moderation that does not even require that the same constructs be measured, “clearly, the inferences justified for equated scores or calibrated scores cannot be justified simply because the scores are reported on a common metric and adjustments have been made using statistical moderation procedures” (p.94). Therefore, unless the alternative format and general assessments can demonstrate they are measuring the same construct with the same reliability (equating), with differing reliabilities (calibration), or that there is some external measure by which the scores are adjusted (statistical moderation), it seems the best solution is comparability by social moderation—including all the preparatory consensus building procedures required.

References

- Barton, K. (2007). Validity and accommodations: The journey toward accessible assessments. In Cahalan Laitusis, C. & Cook, L. (Eds). Large-scale assessment and accommodations: What works? Council for Exceptional Children.
- Cizek, G. J., & Bunch, M. B. (2007). Standard setting: A guide to establishing and evaluating performance standards on tests. Thousand Oaks, CA: Sage.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review*, 64, 5–30.
- Ferrara, S. (2003, June). Linking performance standards: Examples of judgmental approaches and possible applications to linking to NAEP. In A. Kolstad (Moderator), Linking state assessment results to NAEP using statistical and judgmental methods. Symposium conducted at the National Conference on Large Scale Assessment, San Antonio, TX.
- Ferrara, S. , Johnson, E., Chen, W. (2005). . Vertically articulated performance standards: Logic, procedures, and likely classification accuracy. *Applied Measurement in Education*, 18(1), 35–59.
- Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 433–470). Westport, CT: Praeger.
- Huynh, H., Barton, K.E., Meyer, J. P., Porchea, S., & Gallant, D. (2005). Consistency and predictive nature of vertically moderated standards for South Carolina Palmetto Achievement Challenge Test 1999 assessments of English Language Arts and Mathematics. *Applied Measurement in Education*, 18(1), 115–128.
- Koretz, D., Stecher, B., Klein, S., and McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice*, 13(3), 5–16.
- Knight, W., Hakel, M., and Gromko, M. (2008). The relationship between electronic portfolio participation and student success. *Association for Institutional Research*, 107. Retrieved June 8, 2009 from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/43/7d/6f.pdf
- Lambert, C., DePaepe, J., Lambert, L., and Anderson, D. (Winter, 2007). *E-portfolios in action*. Kappa Delta Pi Record. Retrieved January 8, 2009, from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/28/57/02.pdf

- LeMahieu, P., Gitomer, D., and Eresh, J. (1995). Portfolios in large-scale assessment: Difficult but not impossible. *Educational Measurement: Issues and Practice*, 14(3), 11–28.
- Lewis, D., and Haug, C. (2005). Aligning policy and methodology to achieve consistent across-grade performance standards. *Applied Measurement in Education*, 18(1), 11–34.
- Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, 6(1), 83–102.
- Linn, R.L., and Baker, E.L. (1996). Can performance-based student assessments be psychometrically sound? In *Performance-based student assessment: Challenges and Possibilities*. Ninety-fifth Yearbook of the Society for the Study of Education, pp.84–103. Chicago: University of Chicago Press.
- Linn, R.L., Baker, E.L., and Dunbar, S.B. (1991). Complex performance-based assessment: expectations and validation criteria. *Educational Researcher*, 20, 15–21.
- Lissitz, R., & Huynh, H. (2003). Vertical equating for the Arkansas ACTAAP assessments: Issues and solutions in determination of adequate yearly progress and school accountability. Report submitted to the Arkansas Department of Education, Little Rock, AR.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32, 179–197.
- Love, D., McKean, G., and Gathercoal, P. (2004). Portfolios to webfolios and beyond: Levels of maturation. *Educause Quarterly*, 27(2). Retrieved June 8, 2009, from <http://www.educause.edu/EDUCAUSE+Quarterly/EDUCAUSEQuarterlyMagazineVolume/PortfoliostoWebfoliosandBeyond/157290>
- Meister, L., Heath, J., Andrews, J., and Tinggen, M. (2002). Professional nursing portfolios: A global perspective. *MedSurg Nursing*. Retrieved June 8, 2009, from http://findarticles.com/p/articles/mi_m0FSS/is_4_11/ai_n18613917/?tag=content:coll
- Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: Educational Testing Service.
- National Board of Professional Teaching Standards (2009). *2009 Guide to National Board Certification*. Retrieved June 8, 2009, from <http://www.nbpts.org/index.cfm?t=downloader.cfm&id=1134>
- Raju, N. (1991). Reliability of performance-based test scores. In Finch, F.L. (Ed.), *Educational performance assessment* (pp. 81–85). Chicago, IL: Riverside.

- Schaefer, W. D. (2005). Criteria for standard setting from the sponsor's perspective. *Applied Measurement in Education*, 18(1), 61–81.
- Subkoviak, M.J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement*, 25, 47–55.
- Thompson, S., & Thurlow, M. (2002). *Universally designed assessments: Better tests for everyone!* (Policy Directions No. 14). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [1.6.2009], from the World Wide Web: <http://education.umn.edu/NCEO/OnlinePubs/Policy14.htm>
- US Department of Education. (July 2007, revised January, 2009). Standards and Assessments Peer Review Guidance: Information and Examples for Meeting Requirements of the No Child Left Behind Act of 2001. Washington, DC: Author.
- Wiener, D. (2006). *Alternate assessments measured against grade-level achievement standards: The Massachusetts "Competency Portfolio"* (Synthesis Report 59). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [1.23.07], from the World Wide Web: <http://education.umn.edu/NCEO/OnlinePubs/Synthesis59.html>
- Winter, P.C. (May, 2007) Summary of Results: North Carolina Enhanced Assessment Grant State Survey. Unpublished report, Council of Chief State School Officers. Washington, DC.
- Wiggins, G. (1989, May). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan*, 703–713.
- Winter, P.C. (2009, in preparation). Introduction. In Winter, P.C. (Ed.), *Evaluating the comparability of scores from achievement test variations*. Washington, DC: Council of Chief State School Officers.

Section 4: Summary

Chapter 12: Where Are We and Where Could We Go Next?

Summary and Next Steps

Rebecca Kopriva
University of Wisconsin

This handbook has provided a useful template for thinking about measurement comparability and it has identified several useful methods for potentially demonstrating the extent of comparability between test variations and general large-scale tests. The present chapter will summarize selected points, placing them within a context of history and focus, what can be done today, and future possibilities and challenges.

As Winter mentioned in her introduction, over the last 15 to 20 years, rethinking how to best measure academic content led to rethinking what aspects of test-making, implementation, and analysis need to be involved in order to make large-scale test form variations comparable at some level of interchangeability. Specifically, in the 1990's advances in cognitive learning theory led to the identification of an expanded set of measurement approaches that seemed to be promising for use in large-scale assessment. The advances included re-considering what were valued constructs associated with academic content, and identifying novel approaches associated with how students might demonstrate their capabilities in large-scale academic settings. Such approaches included a) the use of open-ended writing prompts designed to measure the writing skills of students; b) short answer and more extended open-ended items meant to measure the students' sophistication in understanding content, such as mathematics, or history, through explaining their conceptualizations, reasoning or meta-cognitive skills; c) performance tasks where students manipulated materials and demonstrated their skills first-hand; and d) development and use of student work portfolios meant to chronicle academic maturation using multiple methods.

As these approaches were tried out, evidence was accumulated about what types of knowledge and skills were measured with various formats, how these methods might be developed to withstand use across classrooms, schools, and districts, how they might be administered in a variety of school contexts, and how results should be analyzed. Various challenges, some sizable, were also noted. Concurrently, measurement experts began to think about how to determine when and how different kinds of performances could be seen as 'comparable'. The focus in those days was on comparability of responses *within* approaches, for instance, when rubrics allowed for various ways for students to demonstrate their knowledge and skills at, say, a level of 2 or 3 out of 4 possible points. The focus was also on determining the kinds of standardized procedures that needed to be in place to build viable performance tasks or constructed response items that measured different content in different academic domains; determining how to construct defensible rubrics for various types of writing prompts, constructed- response items, performance tasks, or portfolio evaluation systems; and identifying standardized criteria that were needed to properly constrain the student work entries in portfolios while still allowing for variability in the types of entries. Because of competing costs and other

issues, work on the approaches has largely stalled for the last 10 years or so. However, the efforts left behind a legacy of work with direct implications for expanding how comparability might be considered, and what types of evidence need to be in place to make this claim.

Current Comparability Challenges

Recently, the clearest challenge to using one set of methods and form types for all students comes from broader inclusion requirements associated with today's statewide academic assessment systems. The 1994 reauthorization of the Elementary and Secondary Education Act required more inclusion of students with disabilities (SDs) and English learners (ELs) in their accountability reports. In 2001 the No Child Left Behind (NCLB) followed the lead of the earlier reauthorization and the wide-ranging students with disabilities legislation (IDEA) passed in 1997. It mandated specific accountability requirements aimed at not only inclusion but explicitly improving the academic performance of these two populations. NCLB also expanded the reviews of the statewide assessment systems to more closely evaluate whether or not the systems were defensibly measuring the academic achievement of SDs and ELs. The focus on special populations highlighted that, for some of these students, traditional large-scale testing methods were not sufficient or valid in assessing their knowledge and skills (i.e. see Kopriva, 2008). Not surprisingly, this focus resulted in an increase in the development of other methods that were supposed to be more appropriate for these students. The methods included adaptations of administration conditions involving the use of accommodations, construction of different kinds of forms, and alternative kinds of procedures aimed at collecting academic information in ways that vary from the general testing formats used for most students. Some questions of comparability involved determining the viability and interchangeability of performances within the testing system which were obtained under varying administration methods. Other considerations, including the work contained in this handbook, focused mostly on comparability among different form variations.

Within the last few years more attention has been paid to assessing in order to support learning as opposed to just what students have learned, doing so in a way that is responsive to identifying errors, while also being defensible, consistent over classrooms, and often over schools (for instance, see Pelligrino, Chudowsky & Glaser, 2001). This charge highlights the variability in learning processes over students and over time. As the field moves forward, viable learning-sensitive assessments will need to be receptive to different progressions as well and wrestle with questions of comparability within students at different points and across all students, not just English learners or students with disabilities.

Carlos Martinez, former associate superintendent in New Mexico, recently posed a pointed question: Do or should testing inferences change when the makeup of item types within a test radically changes from one year to the next? To afford the cost of large-scale academic assessments or to otherwise address changing standards or other pressures, in the last few years several states have altered the manner in how they assess their school populations. Most often, the systems have moved from more to less or no constructed-response items in their testing systems or in particular content areas. Some work has been done to consider how item type interacts with targeted constructs and therefore with inferences that can be made about student mastery (e.g. Li, Ruiz-Primo, & Shavelson, 2006). Over a series of papers, these authors and

others identified the four salient links between item aspects and valued knowledge or skills. They suggested that the first two item aspects, task demands and cognitive demands, established a link that differentiated question type by valued learning while the additional two, item openness and additional factors, served as mediators to the link. Considering comparability over time when testing systems change their item types, or when significant weighting changes occur within tests is to-date something that has not been researched. However, system alterations such as those suggested here appear to have comparability implications as well.

Issues of comparability are not ‘going away’. Rather, as a field, it seems that increasingly we are being held accountable for what we say are the test score inferences, as they pertain to students with particular profiles and, soon, students with different kinds of learning experiences. This handbook provides some guidance about how to think about comparability when variations exist in the system. Systemic frameworks such as Mislevy et al.’s (e.g. 2003) Evidence Centered Design could provide some direction about what elements in the development, implementation and analysis of assessments need to be present, which need to be standardized, and what evidence at each of these steps needs to look like or include. Comparability, like validity, will be an argument to be made and criteria of which arguments are suitable for which purposes will need to be identified and tested.

The Comparability Questions

To begin the conversation, Winter essentially identified three questions that need to be addressed as we consider comparability:

1. What do we want when we want score comparability?
2. What do we mean when we say comparability?
3. How can we evaluate comparability?

The first question seems to focus on the inferential achievement claims the test evidence can support. Evidence will come from elements in the development, implementation and analysis of performance data. Documentation of the procedures used to produce the evidence will need to pass scrutiny and should be evaluated through the lenses of appropriateness for capturing the knowledge and skills of particular students in particular situations—In other words the evidence is viable if the procedures address and minimize alternative explanations. It is probable that test score evidence will come, to a reasonably large degree, from viable evidence at the item level, or other kinds of criteria or constraints meant to focus the types of responses required about particular content and skills.

Winter’s grain size discussion seems to address the second question. That is, is comparability focused at the scale score, achievement score, or single cut-point level? This makes a difference for the kinds of evidence that need to be collected, with the overall expectation that scores from both the general test and variation should be considered ‘interchangeable enough’ and without flags. If the focus is one cut-off score (as in pass-fail), the whole assessment exercise should be focused on producing performances correctly identified on one side of the cut off or the other. If more than one but a discrete number of scores are of interest, then interchangeability documentation needs to address the same question at each of the cut points. When raw or scale

scores are the focus then evidence needs to demonstrate that multiple scores along a continuous range are measuring similar enough knowledge or skills for the students taking each form.

The third question focuses on how to analyze the evidence and make decisions about if the documentation is ‘good enough’. As Winter points out, there must be sufficient evidence of both construct equivalence and score equivalence. Basically, construct equivalence focuses on grounding the meaning of the score inferences resulting from students taking the test variation or the general test, and making sure the user can have confidence that the meanings are the same (or the same ‘enough’). This aspect of equivalence reflects the analysis of evidence produced to defend question 1. Score equivalence focuses on documenting that the scores from the variation and the general forms are behaving in the same way (or the same ‘enough’) for students with similar abilities. Evidence that will be analyzed for this aspect of equivalence comes from data which are appropriate to address question 2, that is, to defend the claims of interchangeability at the targeted grain size.

Examples of construct equivalent evidence that need to be evaluated include:

- same standards coverage
- similar criteria for inclusion
- similar judgments about relevant cognitive demands
- similar internal structure

Evidence of score equivalent evidence includes:

- similar proficiency percentages
- similar score distributions
- similar rank order

Properly used, evidence of score equivalence at the item level may be demonstrated by methods such as

- similar distractor distributions
- similar DIF results
- similar p-values.

It is important, particularly at the item level, that score procedures and group samples be suitably scrutinized and vetted to minimize the possibility of alternate inferential explanations for one or more groups under study.

On the whole it seems that, as variations are more divergent from the general test in terms of format and approach to collecting information, evidence of construct equivalence between the variation and the general test need to be more stringent at a particular grain size. Further, the notion of scores used for what purpose should probably also be considered here—higher stakes at whichever grain size suggests more defensible evidence that alternative explanations for the performances have been considered and found to be not tenable.

As a field we need to get better at linking inferential claims to the methods we are using to collect the data about students’ knowledge and skills. Currently research is lacking in when and how methods changes affect the targeted constructs ‘enough’, and if and when non-targeted cognitive demands are a factor and for whom. As the changes over years question posed by Martinez is debated, and as more variations in large-scale testing occur, it will be imperative that these distinctions become clear. Then we can confidently claim that one set of differences reflect

somewhat distinct, but similar, inferences where the same may not be true for another set. Finally, Dr. Winter has suggested that the criterion for making interchangeable inferences is to focus on the sparsest method used in the testing system that is part of the comparability argument. In other words, portfolios, if properly assembled and evaluated, would probably yield a great deal more information about student skills on targeted constructs than would a multiple choice test. However interchangeable inferences would be tied to the multiple choice form, perhaps with supplementary information available from the analysis of the portfolio contents.

Efforts We Can Make Today

The work on broadening what we mean by comparability started nearly two decades ago. From Messick's work (see Messick, 1989) and the research and development work in the 1990's we reminded ourselves of the importance and centrality of validity in determining the rigor of tests and their interpretations, and lately we have come back to the notion that comparability is essentially linked to validity. We rediscovered the item as a basic unit of analysis in validity, the importance of being clear and explicit about intended item-level targeted cognitive knowledge and skills, and the need for evidence to back up the intentions (not just vague notions at the subtest or test level). We also connected evidence arguments to validity and by extension comparability. Technological progress allows us to increase the standardization of all conditions and has shown promise in enhancing how some students can acquire meaning of items and be able to tell us what they know. These advances may be useful in bridging how students with different profiles can differentially take and respond to tests in ways that engender confidence in common inferences over different methods. All in all, several steps have been taken which can be directly applied today to improving how we consider and evaluate comparability among form and administration variations. Among them five are highlighted here.

1. Operationalizing the Evidence Centered Design Principles and Applying Them to Test Variations

Mislevy and others' Evidence Centered Design or ECD (e.g. 1996, 2003) presents a detailed framework for identifying the units and functions associated with developing assessments and specifying how test components must be synchronized to effectively support the interpretations of scores. Besides identifying particular tasks and procedures within a test and its parallel forms, ECD can also be used to build a comprehensive assessment system structure that concurrently plans for and considers design and evidence elements of all variations as well as the general test. One of the most central and important aspects of ECD is that the inferential claims expected from the test and subtests should be identified first, not last. Once these warrants are clarified, then the design of the system, the test frameworks, procedures and development of items are undertaken to be commensurate with producing data to satisfy these claims.

Another advantage of using ECD is to outline a string of evidence that demonstrates when a traditional system is robust. For instance, among other types of standardized elements, Evidence Centered Design suggests the use of task templates from which to build more traditional item versions. In these task templates contexts, specific elements associated with the targeted construct, characteristics of option choices, certain parameters, or other aspects are purposefully changed in order to 'grow' the task pool. These items may be used to build other forms in the

assessment system or used for other purposes. ECD can also be used to design a system where testing situations change because of student challenges, or it might be useful, potentially, when more than one learning map within a progression necessitates changes in tasks or relationships between tasks. Conceptually, ECD can handle systems where very different methods of collecting student performance data are included as well.

An important part of ECD, especially when test variations are part of the system, is the explanation of alternative arguments for certain students, and clearly specified options plainly linked to students with specific needs. Take for example the situation where variations are being proposed to improve accessibility. First, before options are identified, test developers need to consider arguments about why the general test (including the presentation of forms and general standardized administration and response conditions associated with it) isn't 'good enough' for particular groups of students. These arguments should propose alternative explanations about why students might be scoring as they do, explanations that are due to other factors than the intended constructs. As such, they call into question the meaningfulness and accuracy of the proposed general test score inferences for these students. For the alternative explanations to be effective they need to be tied to students with particular profiles, for example students with the lowest levels of English proficiency and little first language literacy. The precision at this step is essential so that testing solutions can be clearly identified which propose to minimize the alternative argument and bolster support for the overall inferences of the assessment system. A more thorough explanation of how alternative arguments need to be defined can be found in Chapter 12 of Kopriva (2008).

Second, procedural materials need to be proposed and evaluated to determine if they are viable and feasible candidates for minimizing the alternative arguments *and* supporting the general score inferences. As an example, solution options might include a signed version of the forms for deaf and hard-of-hearing students, or a translated script and an oral first language administration of forms for low English proficient students with little first language literacy. Third, evidence needs to be collected that the content constructs have not been altered by the options as compared to what was intended on the general test. Fourth, oversight documentation needs to be collected that the options were implemented in a standardized fashion—so, for instance, were steps taken to avoid cuing in the administration options by developing signing videos so students across the state would hear or see the same administration? Fifth, additional evidence should be collected reflecting that the variations performed as intended, including data that the options and the general test reflect specified similarities necessary if the administrations or forms are to be considered comparable. This type of evidence would probably include checks that both the general and option conditions produce similar structures of responses, including similar factor loadings or other types of dimensional scaling data.

Thus, for any type of variation, it is important to identify the target group who will benefit, identify the points within item, task and/or test development, administration, scoring or analysis that seem to be causing problems for the targeted group, and, within the points, what in particular is problematic. Once this is established proposed solutions that are salient for the targeted group need to be identified, and implemented, and appropriate follow-up needs to occur to provide evidence that the variation is behaving in a 'similar enough' fashion relative to the general population on the general test. As a general rule of thumb, whenever anything is changed at any

point within development or administration, there needs to be adequate evidence that the solution is working as expected for the targeted group.

All in all, Evidence Centered Design can provide construct equivalent documentation, and this type of ‘conscious’ design will also probably have an influence on producing ‘good enough’ evidence of score equivalence. It can be used to identify where evidence needs to be collected, and what types of evidence might be viable and possibly necessary in making a strong argument that the inferences are comparable for everyone taking a test within the assessment system.

2. Defining Suitable Elements

In most cases when form variations are proposed there are changes to the language approach and/or the method by which data are collected. Even when relatively straightforward ‘clarified language’, translated, Braille, or ASL signed forms (completed visually) are suggested, each of those approaches has its own challenges. Testing methods variations, for instance portfolios or performance events, including scripts and forms to complete that document how students demonstrated pre-specified skills or knowledge, present sizable changes to how questions or tasks are presented to students. Each of these sets of changes necessitates that suitable guidelines be specified to guide the development of the variations to produce forms of similar scope. The changes also necessitate that the developers design and implement a plan to independently evaluate and document if the completed variations and general forms are measuring ‘similar enough’ topics at similar levels of cognitive complexity. Additionally, identifying and evaluating form components such as these extend to the selection and use of supplementary tools. These might include, for example, bilingual glossaries, highlighters, or procedural guidelines associated with administration and response accommodations. They also extend to a robust and suitable explanation of how students with identified challenges are matched to accommodations they need, and to oversight procedures to ensure that the matching implementation went according to plan. As the federal peer review guidance has begun to evaluate, these types of evidence as well as evidence documenting fair use of accommodations assigned to those taking the general test vs. those who receive no accommodations are all part of building and demonstrating adequate construct equivalence documentation.

When methods differ substantially from general testing forms and procedures, they almost always involve additional cognitive demands. To date, there has been little work isolating which cognitive demands are salient and should be common across methods. The Barton and Winter chapter specifies several points during development and implementation that they suggest need to be evaluated, and last section below will briefly outline some of the challenges inherent in how clarifications and guidance about these topics might be specified.

Besides the identification of suitable components and procedures, the specification of suitable relationships between elements/procedures also seems to be essential. This extends to decisions of score equivalence (which scores or other quantifiable relational indexes need to be ‘similar enough’ across options), and also to decisions of what is considered suitable data to begin with. That is, for variations where data collection components are different than general test components (e.g. items) or vary substantively by type (e.g. type of items—multiple choice vs. constructed response), when can components be compensatory and for whom, and when do

conjunctive rules apply? When is it ok for students to demonstrate a skill one time? When should the instrument require more evidence? Initially these are judgment calls. Assuming common inferences across students taking different types of forms, methods, or types is the goal, when structured analyses of relationships of different salient variables and their magnitudes are completed, they will help inform the parameters we can comfortably work within. Today, it is suggested that we begin to frame the questions, isolate the components, elements, and relations of interest, and produce some rudimentary evaluations of our decisions.

3. Identifying the Grain Size of Targeted Components

In order to determine construct equivalence, one important aspect is to specify the data points in the variation and general test that will be compared. While the final focus of equivalence is usually at the test level (across what are generally recognized to be parallel forms), is it good enough to only address construct and score equivalence at the aggregate score test level? When item-by-item versions of general test items are made in a variation, several researchers suggest an evaluation by general-variation item pairs to determine if they are measuring the same intended content targets (for instance see De Pascale in this volume, and Sireci and Wells' reference to producing adequate translations). These independent judgments would be a necessary but not sufficient addition to post hoc analytic confirmation of similar test structures and other evaluations of score equivalence at both item and test levels (e.g. see Sireci and Wells in this volume).

But what about when the nature of the item or tasks is different over the general test and variation purported to measure the same content and cognitive skills? For portfolio entries Rigney and Pettit (see below) suggest clearly and specifically defined criteria. To evaluate the similarity of targeted content and cognitive complexity, should the set of entries corresponding to each of these criteria be the grain size of the portfolio variation that should be compared to the corresponding set of items on the general test? Who should perform this evaluation—is it 'good enough' for internal reviewers to document this crosswalk or should an independent judgment by a credible panel or external organization make the judgment calls? Since the general test and variation are supposedly measuring the same content standards, it would seem reasonable that both sets of data collection schemes would be evaluated against the intended constructs in the content standards. The result is essentially another type of alignment document.

The recent work on producing the alternate assessments for states' most cognitively disabled students suggest techniques for producing defensible evaluations of testing methods, although, they are pegged to different content and achievement standards. This means that the only comparison with the general test is at the achievement standards level. How might the comparative methods used in this context be extended to variations when the same content standards are also in play? The reason this question is important is that the alternates are expanding the kinds of defensible data points other variations might use, including, for instance, the use of observational protocols, performance events and other methods designed to directly collect information about students' knowledge and skills. How are these approaches constrained in development and implementation to produce responses the field is comfortable with, and what kinds of evidence needs to be documented to support the inferential judgments?

Initial answers to the questions posed above span a range of research that is currently available (for instance see the literature reviews associated with the studies cited in this volume). Many challenges remain as more development efforts share their findings, but some basic approaches are available now and can be more widely utilized.

4. Standardizing Development and Implementation Elements

In his chapter, DePascale gives examples of the type of development procedures testing systems might employ when clarified language form variations are being used as part of the assessment system. The Barton and Winter chapter summarizes several aspects of development, implementation and oversight which can and should be standardized and evaluated when test variations are substantially different than the general test. They also provide examples of the kinds of evidence that might be collected when these elements are in place. For instance, it is apparent that the target constructs needed to be well-defined, clear, and narrow in focus, and that clear ‘alignment’ techniques need to be put in place to determine if the general test and the test variations are measuring similar information. While the content standards provide the basis for this work, the burden of proof is on the variation and therefore more specific guidelines about how to evaluate data from this approach relative to the general test would seem to be warranted. Gong and Marion (2006) discussed the tradeoffs of building assessments where the items, form elements, and even accommodations are more standardized up front (as in traditional on-demand tests) versus the trade-offs when flexibility in entries is different across students. When different types of data collection methods are part of the overall system, these kinds of tradeoffs need to be considered when the comparability arguments are presented.

To provide credibility of the test variations themselves several elements need to be in place. Rigney and Pettit (1995) reported that identification of clearly defined characteristics of entries in portfolios was essential, which do not specify the nature of the tasks but require that specific representation elements need to be included in order for scorers to properly evaluate the work. For some approaches where teachers and students supply work, completion of a few anchor tasks may be required for all students. These common anchor tasks help to calibrate or anchor the scores across students. Additionally, strong, clear scoring rubrics, and in-depth rubric notes were key to building a successful variation, with training of scorers consistent with the standards of the field. Producing a defensible approach that could be systematically scored across students, teachers, and schools requires forethought and planning. Auditors of the statewide language arts portfolio system in Kentucky found that criteria could be effectively communicated throughout the state, and that teachers could learn to be accurate scorers, given proper training and rigorous oversight. The state of Maryland found that rigorous auditing with samplers and training of proper techniques was successful at minimizing variation and drift when locally evaluated elements were part of the testing system (Ferrara, 1999).

In Rigney and Pettit’s report portfolios were used for all students, rather than the use of portfolios as a variation of the general test and designed to be used for only some students. However, the document suggests the type of rigor that would need to be in place for the variation to confidently yield defensible data. Then, in addition, the use of portfolios as a variation would seem to require further criteria so evidence could be built to document if the variation was performing ‘similarly enough’ to the general test.

5. Using Simultaneous Item/Task Design

About 15 years ago Texas undertook an interesting approach to item development. Items in English were developed, reviewed etc. and then items in Spanish were translated from the English. While still in the pre-operational stage, if translators/ Spanish item writers found that they could not suitably translate the English versions, and changes suitable to both versions could not be found, the English items were discarded. In other words, items had to adequately convey the same meaning in both English and Spanish for this pair to be part of the state's large-scale test. This is a rudimentary example of simultaneous item development.

Using Evidence Centered Design, as variations become a stable part of testing systems (including the use of accommodations as well as form variations or other alternative types of data collection methods), approaching development and implementation in a systemic way involving all parts of the system seems reasonable. Rather than a “do the general test first and then ‘jerry-rig’ variations to address special populations” approach, simultaneous design considers the entire system from the beginning, builds various kinds of items/tasks measuring the same targets, and assembles a thoughtful set of procedures designed to convey the targeted meaning of the items/tasks to students with particular profiles and receive responses from them that are meaningful. A priori designs of analyses implemented in order to confirm the give and take of meaningful content knowledge and skills to and from the developers and students produce evidence of the level to which the methods are working as intended. For today's large-scale systems it has long been argued by some (for instance, Kopriva, 1999, 2008; Tindal & Fuchs, 1999) that actually there should probably not be one test (the ‘real’ test) with accommodations (not as good as the ‘real test’ but what can you do...). Instead, a better approach would be different but equal methods to measuring ‘similar enough’ content and skills, each with some sense of which students would benefit from each particular method without causing undue advantage. Further, as we gear up for the multi-dimensional benchmark tests and other systems designed to focus on supporting ongoing learning in a reliable way over classrooms and for student profiles with different learning maps, it would seem that simultaneous test design approach would be well suited. In either case, ECD provides the framework. As large-scale test implementation migrates to computers, algorithms designed to handle and direct various conditions to students who need them would support this type of development. To date it seems that the main drawback is lack of a cohesive vision and probably reluctance to spend the time upgrading well-worn development procedures to consider the full spectrum of work up front.

Looking Ahead

1. Unpacking Comparable Demands for Tests that Support Learning

As noted above, there is growing interest in developing interim or benchmark assessments. In balanced assessment systems these assessments sit at a middle ground between informal and formal classroom formative tasks occurring in an ongoing fashion throughout the school year, and summative tests focusing on what students have learned over usually a year-long period. Mark Wilson (2004, 2008) suggests that there are two types of possible benchmark tests. One

mimics a summative assessment and is focused on an evaluation of education, albeit over a quarter or semester as opposed to a year. The other type is designed to evaluate the end-status of education but also to provide guidance for continuing instruction based on demonstrated errors in conceptualization or level of skill development. It appears that most consumers would prefer the second type of assessment over the first, but limited understanding about how to build them has hindered their development on a large scale basis. In particular, clear and definitive learning maps keyed to construct grain sizes that are useful for this kind of work are in short supply, as developers struggle to avoid the minute progressions associated with learning research, and instead establish a manageable number of key indicators that demarcate levels of knowledge and skills maturing over time.

While levels along the continuum from novice to expert are difficult enough to identify, it is even more of a challenge to reasonably understand the pathways related to how students with different profiles might move through the levels. As the various pathways are conceptualized, tasks which differentiate levels and also various student progressions are being built. These tasks, and the assessments which embody them, will need to be scrutinized to determine if they are providing reasonable access to students from diverse profiles, and if common inferences across pathways at the assessment level can be supported by adequate evidence. In this case, the profiles, pathways and the access they require won't necessarily be by student subgroup (for instance, an EL or a student with a disability), but will be dependent, to a reasonable degree, on previous learning opportunities interacting with other schooling and personal experiences.

For these kinds of assessments, at least three types of comparability questions appear to be important. First is a key concepts question: "What are the salient levels of the learning maps for a particular scope of study and what are the key pathways between and across levels that adequately capture learning progressions for the full range of students?" Second is a substantive question: "How should different pathways be reflected in different versions of tasks and sets of tasks as necessary, concurrently retaining the integrity of the common constructs being measured across versions while possibly collecting different kinds of information or collecting information in different ways?" Third are two related evidence questions: "What are suitable types of evidence for documenting the veracity of the progressions and pathways, and for documenting access for the full range of students? What kinds of score equivalencies, and at what grain sizes, are acceptable for which purposes?"

It seems that the comparability work completed today will inform these complex questions tomorrow. Currently researchers are finding that access issues can be conceptualized with a discrete and parsimonious number of profiles, and that comparability concerns often can be addressed on the basis of these discrete profiles (for instance see Carr & Kopriva, 2009). While the profiles for the supportive assessments will be different, most likely, there will also be a reasonable and parsimonious number of them that can adequately capture the broad scope of student learning within a scope of topics and cognitive demands. Further, work on understanding under what conditions different methods can yield comparable information will inform this work as well. As a field, learning from the current comparability challenges, we will be much closer to understanding how to address the comparability needs associated with these new types of assessments.

2. Taking Advantage of Interactive Computer Capabilities

Every day we are becoming more technologically savvy, and expanding our use of computers for presenting and scoring a broader range of assessment tasks will be no exception. Already several states and large-scale testing systems have moved their traditional testing systems online, including Minnesota who has incorporated contextual sequences of animation into their discrete large-scale science items. In addition, there are notable examples of fuller assessment tasks in science, including Quellmalz and others' work (2007; 2008) with embedded interactive classroom tasks, and NAEPs experience with computer-based interactive extended tasks. Kopriva and others (2008; 2009) have demonstrated that using multi-semiotic interactive representations to replace language in items for low English proficient students are successful in producing scores on par with those of native English speakers taking either the same interactive items or their traditional item equivalents.

Using the capacity of computers expands how item contexts can more fully engage students in assessment as compared to how traditional tests typically present their questions. This capacity extends how tasks can build on and intertwine evolving contexts with test questions by using animated or simulated sequences, and video and/or audio clips. Computer-based assessment can also access the internet or selected archives of data or related sources, systematically broadening the scope of what information students might be able to use to solve problems or answer questions. Additionally, of course, using the computer and its various capabilities increases the kinds and complexity of algorithms we can easily use to score items for summative use, or elements of items within and across tasks for more formative purposes. All in all the use of computers as a delivery and response capturing tool opens up multiple avenues for enhancing large-scale testing, avenues formerly available only for classroom or research purposes.

Nonetheless, as we are pondering the potential for expanding how we might use computers to assess student knowledge and skills, it is also clear there are differences in how students negotiate these types of interactive items compared to their traditional static counterparts. Issues of comparability would seem to be important to consider in reconciling when data from these types of tasks are compared to or interrelated with data from traditional formats. Of these, four seem to be immediately relevant.

First, many interactive items use visual displays of contexts to provide the backdrop for target questions. Although these contexts are computer-simulated, often they are reminiscent of the 90's performance tasks where students would be presented with concrete materials and equipment with which to conduct an experiment in science or collect data to be analyzed in mathematics. As compared to traditional items which explain contexts using words or static drawings or photographs, these interactive items appear to engage the students more directly in the intended construct and the relevant surround, seemingly decreasing the distance between the latent construct targets associated with the tasks. This difference between the latent targets and the manifest ways they are being measured could have comparability implications. Kopriva, Gabel, and Bauman (2009) have referred to the distinction between static and interactive items as a difference in "stickiness", where it would seem that the greater cognitive engagement in the interactive items could have either a positive or facilitative effect, if communicated well, or a negative, perhaps confusing, effect when delivered poorly. To-date understanding when greater

and lesser distance between latent and manifest targets affects measurement and how has been largely unexplored.

Extending the response opportunities in assessment tasks is another capability eminently feasible as computers are used to deliver tests. To-date, other than extending the ability of students to complete tables, almost all of the projects still require that students respond to interactive tasks using traditional close-ended responses or typed-in constructed-response explanations. However, Kopriva and others have experimented with expanding how to capture performance responses directly, allowing students to manipulate stimuli within the tasks in order to demonstrate their reasoning and skills. For Kopriva et al., these advances are essential because their target groups of students with substantial language challenges find traditional methods of response to be language intensive and impossible to navigate, especially as more complex content is assessed. However, the implications for the work would clearly appear to extend to all students, because capturing performance responses would seem to broaden the range of latent cognitive schemas that could be accessed—schemas that the students' executive functions may not yet have had time to meta-cognitively process into language. Just as the impact of interactive contexts have not been determined or mapped, the same is largely true here. For what types of questions might performance responses be most useful or effective? If students are allowed to respond differently, when do differences matter and when don't they? How should these questions be answered when testing purposes are different, or when students have limited access to traditional methods of response and must use other avenues?

Third, these types of simulated and interactive tasks seem to involve greater amounts and kinds of cognitive demands than do static items. Cognitive psychologists point to the increase in the *density* of the cognitive demands, which appear to reflect increased connections of external and internal stimuli. As compared to static traditional items, in some cases density in interactive items appears to be related to minimizing the distance between the test questions and the latent constructs being measured; sometimes it may be associated with other elements within or across students' internal cognitive maps. The density phenomenon also seems to reflect quantity of input which, depending on students, the effectiveness of presentation, and other conditions, could either lead to a fuller, more complete picture of the task (think about more pixels per square centimeter), or to stimulus overload. To some extent the measurement field must not only become more aware of the distinctions between construct relevant and irrelevant elements of items, but also of how these distinctions are made manifest through greater or lesser density. While elements of static items often either were or weren't facilitative for various populations, density in these 'three dimensional' tasks probably involves continuums where even the most facilitative aspects might reach a threshold beyond which the demands are too imposing. As this is understood, how density interacts with both the targeted content and a broader range of student experiences and abilities will need to be considered as well. Advertising and other mixed-media fields have learned how to use density to their advantage; likewise, as test developers, we need to learn how and when density (or lack thereof) is useful or problematic in producing the targeted and facilitative effects we intend.

Fourth, how the relevant cognitive schemas are engaged in animated and interactive tasks appears to differ in some cases relative to how the schemas are engaged in static representations. To date most research on the relationships between schemas and test questions has focused on

using language (in text or orally) as the primary conveyer of meaning, although sometimes language might be supplemented with other elements such as static drawings or visual representations (Kopriva, Winter, & Wiley, 2004; Kress G. & van Leeuwen, T., 2006). However, research related to how students with language challenges (such as those with disabilities, English learners, and probably poor readers) learn as well as cognitive science advances (e.g. see National Research Council, 2001) makes clear that there are alternative avenues of comprehension and acquiring meaning. Based on this, it seems reasonable to assume that interactive assessment tasks probably do sometimes stimulate some other internal maps associated with meaning than the types of items we currently use. For many students, meaning could be activated through either language or other methods so the net effect may be the same. For those with substantial language difficulties, alternative methods activate essential meanings in the assessment task so students can access the item question where otherwise they couldn't. Further, if effective response possibilities are present, interactive elements in the items allow them to respond with what they know as well; if they aren't available the lack could negate the good intentions associated with other access improvements. What we don't really know is when or how we might be triggering different or even possibly conflicting methods of conveying meaning for different students and if or when this might be problematic. As one example, Carr (2006) found that certain elements in some types of static visuals were differentially 'read' by English learners in one way, by students with learning disabilities in another way, and by deaf and hard-of-hearing students in yet a third. Clearly, this implication needs further study.

Comparability questions in today's assessments seem daunting enough; challenges inherent in the considerations associated with future work appear overwhelming. While it may be tempting to revert back to "the good ol' days", a caution is that much of the work we are facing today and tomorrow is actually work that *should* have been done years ago. That is, although we blithely assigned inferences to test scores for all test takers, we understand now that we perhaps weren't actually as clear about what we were actually measuring as we purported to be. Yes, we had learned the importance of standardizing conditions, checking for reasonable consistency over time and within tests, and understanding how we might design items and evaluate the responses to generalize over students with various abilities. Yes, these attributes greatly improved the robustness of the score interpretations we made. However, we also knew of the high correlations between social economic status and total scores (and while there was certainly evidence to support that students with more resources learn more and better, they were probably too high); that our item development processes weren't always as cognizant as they should have been about the inferences we were making; and that our item type restrictions were sometimes limiting what we could actually measure. So, as we move forward in considering how to properly measure concepts and skills for the broadest range of students, it is time to catch up with the research in the learning sciences and consider how to more thoughtfully build items and tests to differentiate novice and mature learners. Overlaying technology on these findings could and should, then, allow us to reflect back to teachers more information about how to guide instruction, and at the same time use the computer capabilities to align more defensibly with the richness of constructs underlying content in the various subject areas. The test construction and awareness of test taker diversity that we address today will provide the necessary foundation for tomorrow's advances.

References

- Carr, T.G. (2006). Application of STELLA system and relevant findings. Paper presented at the annual National Conference on Large-Scale Assessment, San Francisco, CA., June.
- Carr, T.G. and Kopriva, R.J. (2009). It's about time: Matching English learners and the ways they take tests by using an online tool to properly address individual needs. Paper presented at the National Council of Measurement in Education, San Diego, CA., April.
- Ferrara, S. (May 28, 1999, personal communication).
- Gong, B. and Marion, S. (2006) *Dealing with flexibility in assessments for students with significant cognitive disabilities (Synthesis Report No. 60)*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Kopriva, R.J. (1999) Making State Tests inclusive for special populations: Training guidelines for developing and implementing Inclusive Title 1 Assessments. Washington, DC: Council of Chief State School Officers.
- Kopriva, R.J., Winter, P.C., and Wiley, D.E. (2004) Rethinking the role of individual differences in educational assessment. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA, April.
- Kopriva, R.J. (2008). *Improving Testing for English Language Learners: A Comprehensive Approach to Designing, Building, Implementing, and Interpreting Better Academic Assessments*, Routledge Publishers, NY, NY.
- Kopriva, R.J., Gabel, D. and Bauman, J. (2009). What happens when large-scale items actually use computer capabilities? Exploring issues and redefining challenges. Paper presented at the National Council of Measurement in Education, San Diego, CA., April.
- Kress, G. and van Leeuwen, T. (2006). *Reading Images: The Grammar of Visual Design*. Routledge Publishers, London, England.
- Li, M., Ruiz-Primo, M.A. and Shavelson, R.J. (2006). Towards a science achievement framework: The case of TIMSS 1999. In S.J. Howie and T. Plomp (Eds.), *Contexts of Learning Mathematics and Science: Lessons learned from TIMSS*, Routledge Publishers, NY, NY.
- Messick, S. (1989) Validity. In R.L. Linn (ed.), *Educational measurement* (3rd edn) (pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.
- Mislevy, R.J. (1996). Test theory reconceived. *Journal of Educational Measurement*, 33, pp. 379-416
- Mislevy, R.J., Steinberg, L. and Almond, R. (2003). On the structure of educational assessments, *Measurement: Interdisciplinary Research and Perspectives*, 2003, 1, 3-67

- Pellegrino, J.W., Chudowsky, N., and Glaser, R. (eds.) (2001) *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- Quallmalz, E. (2007). Calipers: Using Simulations to Assess Complex Science Learning, abstract received from <http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0814776>
- Quellmalz, E., DeBoer, G., and Timms, M. (2008). Foundations of 21st Century Science Assessments, abstract retrieved from <http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0814776>
- Rigney, S. and Pettit, M. (1995) Criteria for producing equivalent scores on portfolio assessments: Vermont's approach Presented at the annual meeting for the American Educational Research Association, San Francisco, CA, April.
- Tindal, G. and Fuchs, L.S. (1999) *A summary of research on test changes: An empirical basis for defining accommodations*. Lexington, KY: University of Kentucky, Mid-South Regional Center.
- Wilson, M. (Ed., 2004). *Towards Coherence Between Classroom Assessment and Accountability*. National Society for the Study of Education Press, Chicago, IL.
- Wilson, M. (2008). The nature of quality benchmark assessments. Keynote presentation at the Center for the Assessment and Evaluation of Student Learning Conference, San Francisco, CA, October.