

# Comparability of Individual Students' Scores on the "Same Test"

Charles DePascale and Brian Gong,  
*National Center for the Improvement of Educational Assessment*

## CONTENTS

INTRODUCTION .....	25
Interpreting an Individual Test Score: Comparability and Validity .....	26
Purposes for Comparing Test Scores .....	26
The Same Test .....	27
APPROACHES TO PRODUCING COMPARABLE TEST SCORES .....	29
Design Approaches to Producing Comparable Test Scores .....	30
Psychometric Approaches to Producing Comparable Test Scores .....	33
THREATS TO COMPARABILITY .....	38
Opportunity to Learn .....	38
Familiarity with Item Formats and Tools Used on the Assessment .....	41
Differences in Intended Uses of Test Results .....	41
Differences Among Assessment Contractors .....	43
CONCLUSION .....	45
REFERENCES .....	47

## INTRODUCTION

In large-scale assessments, individual student test scores on the same test are expected to be comparable, but meeting this goal is challenging. The challenge is exacerbated in large-scale K–12 testing because the term “same test” refers to various cases in which students may take different sets of items under different conditions. This chapter addresses how to evaluate whether comparability across conditions is sufficient to support a particular inference or test use. Common threats to comparability arise from a lack of attention to design decisions and psychometric procedures. There are also external threats that might affect the accuracy and/or interpretation of students’ scores. Students’ opportunity to learn (OTL) the content assessed and familiarity with the item formats and tools used on the assessment are two types of comparability threats related primarily to their prior experiences. The process of establishing the comparability of individual students’ scores

on the same test involves compiling sufficient evidence to support inferences and actions related to student performance based on those test scores.

### **Interpreting an Individual Test Score: Comparability and Validity**

Questions about the comparability of individual students' test scores begin not with the comparison of the scores of two or more students, but with the interpretation of a single test score of a single student.

With regard to an individual student's test score, a fundamental comparability question is whether the student would receive the same test score if they took a different form of the test or took the test under different conditions, that is, whether the test scores are interchangeable. If the test is computer based, would the student have received the same score with a different delivery platform, with a different response device, or on a paper-and-pencil form of the test? Would the student have received the same score with a different set of accommodations?

At its core, the question of comparability is a question of validity (Winter, 2010). What inferences about a student's knowledge and skills can we draw from the test score? What claims about a student's performance are supported by the score? What are the appropriate interpretations and uses of the score? These are the underlying questions that must be answered when we attempt to determine the comparability of individual students' scores.

The inferences that we draw about an individual student's performance may be different if we limit our consideration of comparability to only those cases in which all students take a test under the same testing conditions. We can consider the same test form administered under two different conditions. In scenario A, all students are required to take the test under strict timing with a limited set of accommodations; in scenario B, all students are allowed to complete the test and a wide selection of accommodations are available. Each scenario is likely to result in different test scores for many students.

It is likely true that test scores from scenario A are not comparable to test scores from scenario B. It is not true, however, that one scenario produces *comparable* test scores and the other does not. In both scenarios, whether an individual student's scores are comparable depends on the questions being asked about student performance and the interpretations and inferences made about student performance based on the test score.

### **Purposes for Comparing Test Scores**

When we move beyond the interpretation of an individual student's test score to make an external comparison, two types of comparisons are common. One involves comparing an individual score to a fixed standard or a fixed point on a scale. The other involves the direct comparison of two individual students' scores.

Over the past 20 years, perhaps the most common comparison made with individual students' test scores is the comparison of a student's score to the threshold scores (i.e., cut scores and passing scores) that delineate performance standards on state assessments. On the basis of such comparisons, student performance is classified into an achievement level. Achievement level classifications are used for a variety of purposes, including providing information to parents and students about student achievement of

state standards, school accountability, educator accountability, and student promotion or graduation.

Classification of student performance to an achievement level is not the only reason for comparing an individual student's test score to a point on a scale. In addition to criterion-based performance standards, a student's test score might be compared to a norm-referenced standard such as the 20th, 50th, or 95th percentile to inform placement decisions or eligibility for particular programs. A student's test score might also be compared to an expected score signifying an acceptable level of growth or progress from one test administration to the next to determine whether the student is on track to reach a specified goal.

The direct comparison of two or more test scores for the same student or different students may seem quite familiar; however, such comparisons of individual students' test scores are far less common than comparisons to a criterion- or norm-referenced standard. One example of such a direct comparison involves rank ordering of test scores to identify the highest or lowest performing students on a particular test. Another example is when a teacher regards a particular student's performance as a norm or criterion against which other students' performance is compared.

In general, however, the direct comparison of two or more test scores is much more common with aggregate school-, district-, or state-level scores, which are discussed in Chapter 3, *Comparability of Aggregated Group Scores on the "Same Test."*

### **The Same Test**

Any discussion of comparability issues associated with scores on the same test has to begin with a common understanding of what is meant by the term "same test" in the context of large-scale assessment. We are well beyond the point when the measurement community and most policy makers would consider any two tests with the same title (e.g., "Algebra I" or "Grade 3 Reading") to be the same test. Additionally, agreement is widespread that building two tests to the same set of content standards is a necessary but not sufficient criterion for considering two tests to be the same test. For example, few would consider the "Grade 3 Reading Tests" implemented in 2015 by Smarter Balanced, the Partnership for Assessment of Readiness for College and Careers (PARCC), and other state assessment programs in response to states' adoption of the Common Core State Standards to be the same test. One must go deeper than the same set of standards and consider factors such as the test blueprint, item specifications, and administration conditions.

It is rare, however, to find the one example that most people would agree fits the definition of the same test: the case where the same set of items is administered to all students at the same time under the same testing conditions. Even traditional, fixed-form, common assessments that were the norm for state assessments under the No Child Left Behind Act of 2001 (NCLB), such as the Grade 3 Reading Test, often included embedded field-test items that were matrix sampled and varied across students. Although the items from which student scores were produced were the same across students, it would be inaccurate to declare that the students' entire testing experiences were the same. The transition to computer-based testing introduced a plethora of supports and tools available to all students that may further alter the test experience across students (PARCC, 2017; Smarter Balanced, 2018). With the advent of computer-based adaptive

testing, it may be the case that no two students in a state who are taking the Grade 3 Reading Test complete the same set of items.

What then do we mean when we refer to student scores on the same test? Using a state’s Spring 2019 Grade 8 Reading Test as a reference, Box 2-1 describes what we consider to be the same test for the purposes of the discussion in this chapter.

Within the category labeled “same test” there is considerable variation in the set of test items that are completed by individual students. A defining characteristic of the test forms in this category, however, is that they are all constructed to meet the same test blueprint and test specifications. In practice, a state or testing company reports scores from each of these test forms on a single reporting scale and regards the individual student scores as interchangeable.

Test forms listed within the “gray area” have some similarities to the test forms that are considered to be the same test, but also have some key differences. One key similarity is that it is likely that results from these test forms are reported on the same scale as the original test and treated as interchangeable scores.

A “short form” of a test may be built to the same blueprint and even include the same item types, but it is likely less reliable than the original long form of the assessment. In many cases, however, the short form excludes certain item types or alters the distribution of items across item types. As we discuss in the section on psychometric

**BOX 2-1**  
**Examples of Tests Considered to Be the Same as a**  
**State’s Spring 2019 Grade 8 Reading Test**

**Same Test**

- All standard operational forms of the Spring 2019 Grade 8 Reading Test
  - Fixed form with embedded field-test items
  - Multiple fixed operational forms
  - Matrix-sampled operational forms
  - Computer adaptive test
- Standard forms administered with accommodations and nonaccommodation tools and supports
- Standard forms administered in different formats (e.g., paper based, computer based)
- Standard forms with items administered in random order
- Alternate forms of the standard Spring 2019 Grade 8 Reading Test to be administered at different times of the year (e.g., summer retest, fall administration if there was block scheduling)

**Gray Area**

- “Short form” of the Spring 2019 Grade 8 Reading Test
- “Focused retest” of the Spring 2019 Grade 8 Reading Test designed to determine only whether a student meets the mastery or proficient cut score
- Spring 2018 Grade 8 Reading Test
- Spring 2020 Grade 8 Reading Test

**Different Test**

- Spring 2019 Grade 7 Reading Test
- Grade 8 Reading Interim/Benchmark Assessment
- Spring 2019 Grade 8 Reading Test, released and administered by local choice after 2019

approaches to producing comparable scores, differences in reliability can affect score comparability and our inferences about individual student performance.

A "focused retest" such as the type administered in Massachusetts in the early 2000s to high school students attempting to meet a graduation requirement is designed to measure student performance at a more precise point on the reporting scale (i.e., at a significant achievement level cut point) and produce a comparable decision about student achievement at that point on the reporting scale. Even if the same test blueprint is applied to construct the test, it is likely that the original test and focused retest will differ in difficulty and reliability at various points along the reporting scale.

Versions of the Grade 8 Reading Test from the previous year or following year are included in the gray area category for a different reason. In many cases, those test forms can be considered alternate versions of the operational form administered within the same year. In other cases, however, key changes in the testing program from one year to the next have the potential to affect score comparability. Among these are changes in use of the test or stakes associated with individual student scores, transition to a new assessment contractor, and changes in achievement standards.

The three examples of test forms within the "different test" category have critical differences from the original test form that make it impossible to label them as the same test for our purposes. As was the case with the test forms in the gray area, it is possible that results from test forms in this category will be reported on the same scale as the original test form (including the use of a vertical scale). A test designed to measure performance at a different grade level (e.g., the Grade 7 Reading Test) though will have been aligned to different content standards and built to a different test blueprint. A released test form or an interim or benchmark version of the Grade 8 Reading Test may or may not be built to the same blueprint and test specifications as the original test form, but differences in the timing and conditions of the test administration will result in the same test score leading to different inferences about student performance. For example, it may be the case that an interim assessment built to the same blueprint as the original test form supports the same inferences about student performance at the time the interim test was administered (i.e., the student has met the proficient cut score), but it may provide different inferences about how the student is performing or will perform at the end of the school year.

## APPROACHES TO PRODUCING COMPARABLE TEST SCORES

Achieving comparability of individual students' scores on the same test requires very thoughtful and careful planning, execution, monitoring, and evaluation. Comparability is the result of a process that involves a combination of design and psychometrics (i.e., statistics). Comparable scores cannot be achieved solely through the application of formal psychometric procedures, nor can comparable scores be achieved solely through the design of a large-scale assessment system. Most importantly, the psychometric and design approaches to achieving comparability are interdependent. The psychometric procedures, particularly in the case of state assessment programs, depend on strict design assumptions having been met. The design approaches rely on psychometric procedures to account precisely for differences in difficulty between forms of the test.

The psychometric approaches to achieve comparability include a range of procedures designed to enable direct comparisons of student performance on different sets

of items, regardless of whether those sets of items are intended to be the same test or different tests. In the case considered here, in which different sets of items are intended to be the same test, the level of desired comparability is that scores from any of the alternate forms of the same test can be treated as interchangeable. In other applications, it may be sufficient to establish that classifications of student proficiency on the two assessments are comparable (e.g., determining a common college-readiness cut score on two different assessments).

The design approaches to achieve comparability involve a series of decisions related to the overall design of the assessment system. They include decisions related to how the test is developed, administered, and scored. They also include decisions about who will be taking the test and how the results will be used. As with the psychometric procedures, the goal with large-scale assessment programs that include multiple forms of the same test is to make design decisions that result in scores that are interchangeable across test forms.

### **Design Approaches to Producing Comparable Test Scores**

Design approaches to producing comparable individual student scores across alternate forms of the same test begin with an understanding of the construct being assessed. An understanding of the construct and the type of evidence needed to support the claims and inferences made about student performance on the construct are at the core of principled approaches to assessment design.

On large-scale K–12 assessments, the starting point for understanding the construct is usually the set of college- and career-readiness content and performance standards adopted by the state. These standards define the knowledge and skills that students are expected to have achieved at the end of a grade level, grade span, or course. Efforts to ensure alignment often begin with the development of evidence models or statements that describe in detail the aspects of student responses that would provide evidence needed to support the claims being made about student achievement based on performance on the test (Zieky, 2014). These evidence statements are supported by the development of detailed blueprints and test specifications that define what will be included on the test and how it will be measured. Test blueprints may contain information about the total number of items and points on the test, how those will be distributed across test sessions, and, most importantly, how the content standards to be assessed will be distributed across those items and points. Test specifications include additional information about the design of the assessment: details regarding item types, cognitive complexity (e.g., depth of knowledge), mode of administration (i.e., computer based or paper based), timing, and the nature and use of accommodations. In K–12 state assessments, the finished product is subjected to formal alignment studies that evaluate the degree to which each individual test item measures the standard or standards it was designed to measure and also the degree to which the set of items (i.e., the test form) measures the complete set of standards it claims to measure.

From a comparability perspective, this level of understanding of the relationships among the construct, the standards, and the assessment is critical to understanding what deviations from the “same test” are likely to impact the measurement of the construct and, consequently, impact the comparability of individual student scores.

These deviations include changes to the blueprint across test forms, such as shortening the test or changing the balance of selected-response and constructed-response items. Deviations that may impact the comparability of individual test scores also apply to the level of flexibility allowed within the administration of a single test form or alternate test forms built to the same test blueprint and specifications. Changes to the test blueprint such as shortening the test are addressed in the next section on psychometric approaches to producing comparable test scores. In the remainder of this section our focus is on decisions related to standardization versus flexibility that have the potential to impact the comparability of individual student scores on the same test.

### *Standardization Versus Flexibility*

Standardization has been a cornerstone of large-scale assessment. Content, administration, and scoring are the three pillars of standardization that drove the construction of large-scale assessment throughout the 20th century and into the beginning of the 21st century. One cannot overstate the importance attached to students taking test forms that contain the "same" content, are administered to all test takers under the same testing conditions, and are scored using the same specified scoring procedures. Standardization was considered essential to ensuring accurate measurement of student performance by controlling error and minimizing the impact of factors that are irrelevant to the construct being assessed or the purpose of testing and that might distort inferences about student performance. Standardization was also considered an essential requirement for making direct comparisons of individual students' scores.

With the Improving America's Schools Act of 1994 and NCLB, reauthorizations of the Elementary and Secondary School Act of 1965, federal assessment requirements affected the conception of standardization in several ways. Specifically, the requirements for (1) standards-based assessments aligned to challenging academic content standards, and (2) the inclusion of all students in large-scale assessments, resulted in a need to rethink what was meant by standardization with regard to key aspects of test administration and scoring. Within the category of test administration, two areas where more flexibility was allowed were in the timing of the test and in the use of test accommodations. Within the category of scoring, the increased use of constructed-response items and other item types that could not be easily scored by machine on large-scale assessments created a need for new sets of scoring protocols and procedures to ensure standardization in scoring, that is, to increase the likelihood that a student's response would receive the same score regardless of who scored it and when it was scored.

In each of these cases, the comparability of individual students' scores rests on the argument that the flexibility introduced into the testing process actually was more suited than strict standardization to minimizing error, removing irrelevant factors that might affect test performance, and ultimately producing more accurate student test scores.

**Timing** Historically, large-scale standardized tests were administered with time limits and strictly timed sections. Whether the test was regarded as a speed test or a power test, strict time limits were regarded as necessary to support the claim that all students

took the test under the same conditions, thereby meeting a necessary condition to support direct comparison of scores.

With the shift to standards-based assessment there was a prevailing belief that speed of response was not a component of the constructs being measured and that students should be provided adequate time to complete all items on the test in order to provide a more accurate estimate of their level of proficiency. In some states, the result was a timing policy in which timed test periods were expanded to levels designed to ensure that all students were able to complete the test. Test sessions were scheduled for 50 or 100 percent longer than the time expected for the vast majority of students to complete the test. In Massachusetts, the tests of the Massachusetts Comprehensive Assessment System are untimed, with the limitation that a test session must be completed within a single school day (MA DESE, 2019).

**Use of accommodations** Prior to the requirements of the Improving America's Schools Act, the Individuals with Disabilities Education Act, and NCLB to include all students in testing, it was not uncommon for 15 to 20 percent of students to be excluded from large-scale state assessments, a group including most students with disabilities (Lehr & Thurlow, 2003). The requirement to include all students in state assessments brought with it the requirement to allow students with disabilities the use of appropriate test modifications (i.e., accommodations) during testing. The 2014 Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014), referred to hereafter as the Joint Standards, define these test accommodations as adjustments that do not alter the construct being assessed that are applied to test presentation, environment, content, format (including response format), or administration conditions for particular test takers; and specify that accommodated scores should be sufficiently comparable to unaccommodated scores that they can be aggregated together. In practice, states often group accommodations into four categories:

1. Presentation, where accommodations involve varying the manner in which the test is presented (e.g., large-print, Braille, translated, read aloud);
2. Response, where accommodations involve varying the manner in which the student responds (e.g., use of graphic organizers, nonuse of a separate answer sheet, scribe);
3. Setting, where accommodations involve varying the setting in which the test is administered (e.g., one-on-one, small group); and
4. Timing, where accommodations involve varying the timing of the test (e.g., extended time, additional scheduled breaks).

The argument for increased flexibility and the use of accommodations for students with disabilities was centered on accessibility to the test and leveling the playing field for students with disabilities. Again, the goal was to produce a more accurate measure of the construct being assessed. One method recommended for determining if the accommodation leveled the playing field without affecting the construct required the demonstration of a differential boost. A differential boost indicates that the use of the accommodation affected the performance of students with disabilities for whom



the accommodation was intended more than it increased the performance of students without the disability.

In most cases, there was little debate over the appropriateness of several widely used and established accommodations such as Braille or large-print test booklets, small-group test administration, or allowing extended time. Other accommodations, such as use of a read-aloud accommodation on a reading or English language arts test or the use of a calculator in the noncalculator section of a mathematics test were often the subject of intense debate within and across states. In the case of the read-aloud accommodation, the debate often focused on whether the construct being assessed and the inferences drawn about student performance were related to decoding (i.e., being able to read the text) or comprehension (i.e., being able to interpret and apply the contents of a text). For a fuller discussion about the uses of accommodations, see Chapter 7, *Comparability When Assessing Individuals with Disabilities*.

**Scoring procedures** The increased use of constructed-response items and other item formats that could not be easily machine scored introduced variability into the scoring process that did not exist when tests consisted almost exclusively of machine-scorable, multiple-choice items. Not only were the rules for scoring a student response more complex, but also training and monitoring were required to ensure that groups of human scorers were able to apply those scoring rules consistently and accurately. The infrastructure needed to support the development of scoring rubrics, training materials, training, scoring, and real-time monitoring of scoring had to be developed and implemented to support the argument for comparability of individual student scores.

Scoring the same items across multiple test administrations within and across years added additional strain to the process of ensuring standardization in scoring. In addition to ensuring that there had been no changes to the item and scoring rules and that training and training procedures were consistent across years, it became necessary to take additional steps to ensure that student responses were being scored consistently from one administration to the next.

### **Psychometric Approaches to Producing Comparable Test Scores**

In large-scale assessment programs offering alternate forms of the same test, the desired level of comparability is that the scores on the alternate forms of the test are interchangeable. That is the case within state assessment programs when students are administered multiple forms of the Grade 8 Reading Test within the same year or when students take a new form of the Grade 8 Reading Test each year. That is also the case when students take multiple administrations of the ACT or the SAT as part of their college admissions process. In each of these cases, the task is to link the alternate forms of the test in such a way that an individual student would receive the same score regardless of which form is administered.

The type of linking required to achieve scores that are interchangeable is called *equating*. Although it is quite common to hear all forms of linking tests referred to colloquially as equating, the term is reserved for linkages that meet a strict set of requirements and where the intended interpretation is that the scores from the two tests can

be considered interchangeable for all intended purposes and uses. Holland and Dorans (2006) contrast equating with linking done for two other purposes: *predicting* and *scale aligning*. They refer to predicting as the oldest form of linking and one that is often confused with other methods of score linking.

As the name implies, the goal of linking approaches that are classified as predicting is to predict a student's score on a particular test from other information that is available about the student. That information might be a score on another test, which is where confusion with other purposes might arise. The other information about the student that is used as a predictor, however, could just as easily be performance on multiple tests, grades, dispositions, background characteristics, or anything that would improve the prediction of the student's score on the test of interest. Although a better prediction might be achieved by a predictor test that has a high degree of similarity to the test on which the score is being predicted, there is no requirement that the tests are similar in any way or even measure the same general content area, let alone the same construct. Regression procedures are the common tool to accomplish linkages for predicting.

The purpose of scale aligning, also referred to simply as scaling, is to produce comparable scores on two tests by transforming the scores from two different tests onto a common scale (Holland & Dorans, 2006). Holland and Dorans describe six types of scaling, categorized by whether the two tests are designed to measure similar constructs and, if so, whether the two tests have similar reliability. In each of the six cases, the result of scale aligning is that the scores from the two tests are placed on a common scale to produce comparable scores. In none of the cases classified as scale aligning, however, are the scores on the two tests to be considered interchangeable. That interpretation is reserved for linking procedures that meet the strict requirements necessary to be classified as equating.

In this chapter, we do not address all of the examples of scale aligning, but there are two that fall under the heading of calibration that are particularly relevant to the topic of comparability, large-scale state assessment, and tests that some might consider the same test. Calibration is considered a strong form of linking two tests and applies to situations in which the two tests meet many of the same requirements as equating, such as two tests that are designed to measure the same construct and may even be designed according to the same general test specifications (Dorans, Moses, & Eignor, 2010; Linn, 1993; Mislevy, 1992). The first is the case in which calibration is used to link a shorter form of a test to the longer, original form of the same test. The second is the general case in which item response theory (IRT) procedures have been used to place test items on a common scale.

With increasing concerns about the time needed to complete tests measuring complex, college-readiness standards, states are facing demands to reduce testing time by shortening tests. In these cases, most states attempt to maintain the same test blueprint in terms of content, cognitive demand, and the types of tasks that students are required to perform; however, even when such conditions are met, it is likely that the shortened test will have a lower level of reliability than the original test. Calibration procedures can be conducted to link the shorter and longer tests in a way that provides scores that have sufficient comparability to allow comparisons of scores of individual students on the short and long forms. It is also possible to apply calibration procedures to link the two tests so that comparisons can be made of aggregate group performances such as estimates of the percentage of students scoring above the proficient benchmark. It

cannot be assumed, however, that one set of calibration procedures will provide the same level of accuracy for individual and group comparisons (Linn, 1993; Mislevy, 1992). When it is desirable or necessary to treat long and short forms of a test as the same test, it is important to ensure that the appropriate linking procedures are used to support the most important comparisons and to understand how well all intended or likely comparisons can be supported.

A second caution is to not confuse equating with the use of IRT procedures to place items on a common scale. With the increased use of computer-based adaptive testing and states making use of items from commercial or shared item banks, there is a great reliance on the use of IRT to place items from many different tests and testing situations onto a common scale. It is true that applying IRT to place items on a common scale is typically a first step in equating large-scale assessments, but simply building alternate test forms by selecting items from a pool of items on the same scale should not be regarded as equating two tests. Even if all of the assumptions regarding the use of IRT to place the items on a common scale have been met, to use that common scale and those item parameters to claim that two tests built with items from that scale have been equated, all of the design and psychometric approaches to producing comparable test scores and threats to comparability discussed in this chapter must be considered.

### *Equating*

The purpose of equating tests is to allow the scores from each test to be used interchangeably, as if they had come from the same test. (Holland & Dorans, 2006)

We have established that in the vast majority of cases in large-scale assessments in which there is interest in the comparability of individual students' scores on the same test it is unlikely that those students took the same test in the literal sense (i.e., completed the same set of test items). Administering test forms with different items is the desirable case when referring to alternate forms of the same test administered in different years. Administering test forms which were not identical, however, was also necessary within the same year with fixed-form large-scale state assessments that included embedded field test items or matrix-sampled equating items, which were the norm for nearly two decades. In recent years, administering different sets of items across students has become increasingly common with the emergence of various forms of computer adaptive testing and the renewed interest in and use of matrix sampling to assess complex standards such as the Next Generation Science Standards.

When the goal is to treat individual students' test scores from two test forms that contain some different items as interchangeable then it is necessary to make a direct link between the two tests through equating. To support the claim that the test scores are interchangeable, equating has the strongest set of assumptions (i.e., requirements) of all of the approaches to linking two tests. Holland and Dorans (2006) identified five requirements for two tests to be equated successfully (see Table 2-1).

Each of the five requirements reflects, to some degree, a theoretical concept or measurement ideal that cannot be fully met in practice with real tests administered under real testing conditions, and with real people taking those tests. That, however, does not excuse test developers and test users from the need to adhere to best practices with regard to the development of test forms and the interpretation and use of individual

**TABLE 2-1** Requirements for a Linking to Be Considered Equating

The equal construct requirement	The tests should measure the same constructs.
The equal reliability requirement	The tests should have the same reliability.
The symmetry requirement	The equating function for equating the scores of Y to those of X should be the inverse of the equating function for equating the scores of X to those of Y.
The equity requirement	It should be a matter of indifference to an examinee to be tested by either of two tests that have been equated.
The population invariance requirement	The choice of (sub)population used to estimate the equating function between the scores of tests X and Y should not matter; that is, the equating function used to link the scores of X and Y should be <i>population invariant</i> .

SOURCE: Holland and Dorans (2006).

students' test scores. Striving to meet the five requirements, demonstrating an attempt to meet the requirements, and providing evidence of the extent to which the requirements have been met are critical to supporting an argument that individual students' test scores are not only comparable but are, in fact, interchangeable across test forms.

The starting point for the comparability of individual students' scores on the same test are the equal construct and equal reliability requirements. Stated simply, the two tests should be built to the same test specifications. Those test specifications must include factors such as the balance of representation of items (or score points) across content and cognitive processes, the use and distribution of items (or score points) across item types, and the number of items (or score points) that will be included on the test. In short, the test specifications should include details on any factors that could affect whether the two tests measure the same constructs and have the same reliability. Recall that in the section "Design Approaches to Producing Comparable Test Scores" we discussed several design choices where decisions must be made regarding whether a particular factor affects the constructs or claims being made about student performance (e.g., testing time, inclusion of particular item types, and use of accommodations). For the sake of this discussion, it is assumed that those issues are resolved before psychometric procedures are applied to equate the two test forms.

The equity requirement states that it should be a matter of indifference to an examinee to be tested by either of two tests that have been equated. In practice, we know that an individual student's test score might vary based on the particular set of items that they encounter on their test form. For example, if all items except the final item were the same across two test forms, it would make a difference to a student if they were able to respond correctly to the final item on form A but not to the final item on form B. The impact of an individual student being more or less familiar with a particular item or items on a test form will be lessened if the items are sampled from the domain in the same way across test forms and as the number of items and total points increase on each test form.

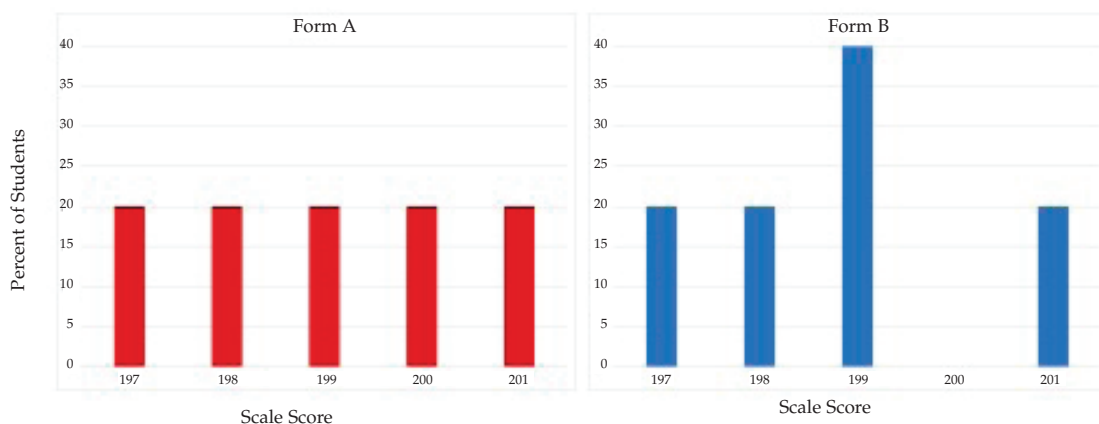
It is a greater concern to equity, however, if the likelihood of a student with a particular level of achievement having their performance classified as proficient or meeting a particular cut score varies across test forms. The purpose of equating is to adjust for

slight differences in difficulty across test forms so that an individual student will not be advantaged by receiving a test form that is slightly less difficult or penalized by receiving a test form that is slightly more difficult. However, equating, per se, cannot account for gaps in the reporting scale that might make it more or less difficult for students to attain a particular cut score based on the test form they receive.

Consider an example in which the proficient cut score on a state assessment is 200. Due to a combination of the difficulties of the particular items on form B and rounding rules adopted by the state it is possible for students to earn a scale score of 200 on form A, but it is not possible for students to earn a scale score of 200 on form B (see Figure 2-1). There is a gap in the reporting scale on form B such that students may earn a scale score of 199 or 201. In this example, the 20 students who would have earned a scale score of 200 on form A and a performance classification of proficient now receive a scale score of 199 and their performance is not classified as proficient. In this case, although the test forms are equated, it is clearly not a matter of indifference to students performing near the cut score of 200 which test form they receive.

Note that the example described here is not related to measurement error in the individual student's score. It will always be the case that there are students whose true score is near an achievement level cut who will score above the cut on one test administration and below the cut on a second administration—even if they were to take the same test form both times. This is more a case in which a student whose true score is near the achievement level cut is unable to earn a particular scale score because of the properties of the test form or the reporting scale.

The symmetry and population invariance requirements are relatively easy to evaluate in practice and the likelihood of meeting them is increased by meeting the equal construct and reliability requirements, that is, by developing test forms that measure the same constructs with the same level of reliability (Dorans et al., 2010). It is not likely that the equating functions produced to test either the symmetry or the population invariance requirements will be exactly the same based on individual test forms and samples of students, but it is important to understand where deviations from the requirements are occurring, to understand whether they fall within expected ranges



**FIGURE 2-1** Example of the difference in scale scores for two different forms of the same test.

given factors such as sample size and level or distribution of performance, and to attempt to determine why those deviations are occurring.

Assuming that the construct of the test forms is the same for all subgroups or samples of the population, demonstrating that the reliability of the test is consistent across subgroups is an important part of demonstrating that the test scores across forms are comparable for individual students.

There are a number of statistical approaches to conducting equating procedures and it is beyond the scope of this chapter to describe them or to discuss their advantages and disadvantages. With regard to the comparability of test scores, however, it is important to remember that all equating procedures begin with the requirements that the test forms being equated are measuring the same constructs with the same level of reliability. The statistical procedures will produce results even if those two requirements are not met, but those statistical results will not produce comparable test scores for individual students.

### THREATS TO COMPARABILITY

In describing design and psychometric approaches to producing comparable test scores for individual students, certain threats to score comparability are clearly implied. Test forms that are constructed to different test blueprints, include different item types, or are of significantly different length (i.e., differ in reliability) are unlikely to produce scores with a level of comparability needed to be considered interchangeable. Test forms that have significant differences in testing time, accommodations offered, and mode of administration (paper based versus computer based) are likely to have issues that must be examined and resolved before declaring that the scores they produce are comparable for individual students.

In addition to those internal threats to comparability (i.e., threats built into the assessment), there are additional threats to comparability that should also be understood when interpreting and using the results from large-scale assessments.

### Opportunity to Learn

OTL has long been acknowledged as a major instructional factor affecting student achievement (Kurz, 2011). At a gross level, OTL is defined in terms of the percentage of time in the school schedule allocated for instruction (Carroll, 1989). With regard to assessment, historically “OTL has been conceptualized as opportunity to learn what is tested” (Haertel, Moss, Pullin, & Gee, 2008).

More refined definitions include consideration of resources and other factors that affect the quality of instruction, students’ preparedness to participate in learning, and interactions among teachers, students, and standards (AERA et al., 2014; Banicky, 2000; Cooper & Liou, 2007; Friedlaender & Darling-Hammond, 2007; Shepard, Hannaway, & Baker, 2009). Among these factors are unequal or inequitable access to curriculum, instruction, and resources (including teacher and administrator quality); teaching to the test and other practices associated with high-stakes testing; adequacy of school finance; disciplinary and exclusionary practices; access to culturally responsive teaching and curriculum and school culture; access to evaluation for special needs; and implementation of 504 plans and individualized education plans.

With regard to large-scale assessments, a much more mundane and more easily correctable factor affecting OTL is the scheduling of the assessment with respect to completion of instruction. Historically, most individual state assessment programs have had a prescribed testing window that was relatively narrow and consistent across the state, in the attempt to ensure that all schools had the same opportunity to instruct students prior to the assessment. When multistate assessment programs such as Smarter Balanced and PARCC were introduced, test administration guidelines were developed to maintain this same sense of equal opportunity to be exposed to instruction while accounting for differences in school year starting dates in states across the country. An example from the Smarter Balanced 2014–2015 technical report demonstrates how the percentage of a school's annual instructional days was used as a metric to minimize scheduling as an OTL factor that might affect the comparability of the test scores of individual students completing the Smarter Balanced assessment in different states (Smarter Balanced, 2016, p. 6-2) (see Box 2-2).

### **BOX 2-2 Test Administration**

Students in Smarter Balanced member states participated in the 2015 test administration once a specified percentage of the school year had occurred. Each state established a schedule for the administration of the Smarter Balanced summative assessments using a testing window as defined below:

#### **Grades 3–8**

- Testing shall not begin until at least sixty-six percent (66%) of a school's annual instructional days have been completed, and
- Testing may continue up to and including the last day of school.

#### **Grade 11**

- Testing shall not begin until at least eighty percent (80%) of a school's annual instructional days have been completed, and
- Testing may continue up to and including the last day of school.

States were allowed to establish more specific windows within the constraints of the consortium-defined windows described above.

SOURCE: Smarter Balanced (2014c, p. 25).

Given the connection between OTL and student achievement, it is no surprise that OTL is closely linked to large-scale assessments, specifically to the interpretation and use of results from large-scale assessments for high-stakes decisions (Pullin & Haertel, 2008). The Joint Standards define OTL in the context of testing as "the extent to which individuals have had exposure to instruction or knowledge that affords them the opportunity to learn the content and skills targeted by the test." The Joint Standards further state that opportunity to learn "has several implications for the fair and valid interpretation of test scores for their intended uses." The two testing

standards related to OTL directly address the interpretation and use of test scores for high-stakes decisions:

- **Standard 12.8.** When test results contribute substantially to decisions about student promotion or graduation, evidence should be provided that students have had an opportunity to learn the content and skills measured by the test.
- **Standard 3.19.** In settings where the same authority is responsible for both provision of curriculum and high-stakes decisions based on testing of examinees' curriculum mastery, examinees should not suffer permanent negative consequences if evidence indicates that they have not had the opportunity to learn the test content.

Although the connection among OTL, student achievement, and test scores is quite clear, the relationship between OTL and comparability is less straightforward. If a student has not had an opportunity to learn the content and skills measured by the test, the Joint Standards are clear that it would be inappropriate to use the student's performance on the test as the basis for a high-stakes decision such as promotion or high school graduation.

There is clearly a fairness issue with regard to the use of the individual student's test score. Judgment of the comparability of the student's test score may change, however, based on the assertions being made about student performance and/or the intended interpretation of student performance on the test. If the assertion is that the test score describes the student's current level of achievement in the content area then it is likely that the score is an accurate reflection of student achievement and could be considered comparable to the scores of other students taking the test.

Different assertions, however, could lead to different conclusions about comparability. Consider our original assertion, that the test score describes the student's current level of achievement in the content area, in relation to three alternate or additional assertions about student performance:

1. The test score reflects the achievement of students after they have received instruction in the content area being tested.
2. The test score reflects the level of achievement students can attain if they have had an adequate opportunity to learn the material.
3. The test score reflects what students could achieve at the next grade level or in college if provided an adequate opportunity to learn.

For each of these assertions, it would not be appropriate to consider the student's score comparable to the scores of other students who had the opportunity to learn the material.

In considering the original assertion and three alternative assertions presented above, nothing about the student's test score and its reflection of student achievement has changed. Our interpretation of the comparability of the test score due to OTL, however, did change, in that it is dependent upon the assertion being made or the expected interpretation of the score.



### Familiarity with Item Formats and Tools Used on the Assessment

An issue often conflated with a student's OTL is a student's opportunity to become familiar and comfortable with the item formats and tools that they will encounter on an assessment. Unlike OTL, however, where the impact on comparability is conditional on the claims and use of the test score, the lack of an adequate opportunity to become familiar with the item formats and tools used on the assessment is almost always a serious threat to comparability.

The distinction between OTL and familiarity with the assessment has to do with the expected relationship between a student's level of knowledge and skills, or achievement, and the score that the student will earn on the test. It is expected that two students with different levels of knowledge and skills due to differences in OTL will earn different scores on the assessment, that is, they will earn scores that accurately reflect their current level of achievement. However, if two students have equal levels of content knowledge and skills but differ in their familiarity with the item formats and tools used in the assessment, it is likely that the student who is more familiar and comfortable with the assessment will earn a higher test score. The observed difference in student performance would be attributed to construct-irrelevant variance (i.e., familiarity with the assessment) and the test scores would not be considered comparable.

The Joint Standards make it clear in several places that students should be provided an opportunity to become familiar and comfortable with the item formats that will be included on the assessment and also the tools, accommodations, and other supports that will be available or required for use during the test:

- **Standard 4.16.** The instructions presented to test takers should contain sufficient detail so that test takers can respond to a task in the manner that the test developer intended. When appropriate, sample materials, practice or sample questions, criteria for scoring, and a representative item identified with each item format or major area in the test's classification or domain should be provided to the test takers prior to the administration of the test, or should be included in the testing material as part of the standard administration instructions.
- **Standard 6.5.** Test takers should be provided appropriate instructions, practice, and other support necessary to reduce construct-irrelevant variance.
- **Standard 8.2.** Test takers should be provided in advance with as much information about the test, the testing process, the intended test use, test scoring criteria, testing policy, availability of accommodations, and confidentiality protection as is consistent with obtaining valid responses and making appropriate interpretations of test scores.

The recent transitions from paper-based to computer-based testing and from fixed-form to adaptive testing have introduced several examples of potential threats to comparability if appropriate steps are not taken in advance to ensure that students are familiar and comfortable with the item formats and tools they will encounter on the assessment. Examples follow of areas in which issues that are a potential threat to score comparability resulting from newly implemented computer-based tests have been encountered:

- **Familiarity with the test platform.** As a starting point, students must be comfortable working within an item and navigating through a test form. This includes comfort with the use of elements such as mouseover hover boxes, pop-up windows, or hyperlinks. Students must be familiar with the procedures for moving from one item to the next, skipping items (e.g., if an answer is required before moving on), and returning to previous items.
- **Familiarity with the student response device.** Students must be familiar with the response device, including the issues associated with the size of the screen and comfort with the use of a touch-screen, keyboard, or mouse, as required by the particular device.
- **Physical demands of responding to new item types.** In addition to the cognitive complexity of items associated with the new item types (e.g., multiple-select selected-response items, and technology-enhanced items such as drag-and-drop and hot spot), students also must possess the dexterity required to respond to the item.
- **Space available for written responses.** On a paper-based form students were provided with a fixed space to produce written responses (e.g., one page for constructed-response items and four pages for an essay). The computer-based version of the test used an expanding response box, meaning that students were given no visual cues about the expected length of a response.
- **Use of tools to respond to items on the mathematics assessment.** Students are required to use equation editors, graphing tools, and built-in calculators to respond on screen to mathematics items.
- **Presentation of reading passages and other stimuli.** Students are unable to view the test item and the reading passage at the same time.
- **Impact on common test-taking strategies.** When taking adaptive tests, students cannot apply strategies that they have been taught for other testing formats, such as previewing all of the items in advance and focusing first on items they can answer.
- **Familiarity with scoring rules for new item types.** When responding to multiple-select selected-response items, students may not be aware that selecting too few or too many options will result in the response being scored as totally incorrect.

Each of the potential threats described above can be mitigated by providing adequate opportunity for students to become familiar with the requirements of the computer-based test.

### Differences in Intended Uses of Test Results

Much like the case with OTL, it is widely accepted that differences in the intended uses of test results can affect the performance of individual students. Students taking a test that will be used for high-stakes decisions such as promotion to the next grade, high school graduation, or eligibility for a scholarship may perform differently than students who are taking the same test without such stakes attached to the results (Steedle & Grochowalski, 2017; Wise & Demars, 2005). Also similar to the case with OTL, the threat to comparability lies primarily in the interpretation of the student's test score.

The threat to comparability is greatest in cases in which there is a difference between the intended uses for the test when achievement level cut scores on the test are set and the intended uses when the test is administered. This would be the case within a state when achievement level cut scores are determined under one condition and then applied under a different condition. This would also be true across states when achievement level cut scores that are intended to be common across states are applied under low-stakes conditions in one state and high-stakes conditions in another.

The comparability of individual student scores is also threatened by the fact that it is virtually impossible to quantify and isolate the impact that content-based performance standards, intended uses, and outcomes (i.e., impact data) have on the process of identifying achievement level cut scores (i.e., standard-setting processes) for large-scale assessments.

As a concrete example of the threats to comparability described in this section, consider the common case of a judgment-based standard-setting process used to establish achievement level cut scores on a high school mathematics test. The central question asked of panelists during standard setting is some variation of the following: "Would a borderline-proficient student answer this item correctly?" A standard-setting panelist's response to that question is likely to be affected by the intended use of the test scores. Panelists will consider how likely it is that a student will persevere on an item that is complex, requires multiple steps to complete, or requires a written explanation to support a response. It is generally accepted by standard-setting panelists that students will be more motivated to persevere on such items when high stakes such as high school graduation are associated with performance on a test. Therefore, for a particular item, panelists might conclude that a borderline-proficient student is likely to answer the item correctly under high-stakes conditions, but unlikely to answer the item correctly if there are no student consequences attached to the test score.

When a decision is based solely on content and student motivation, standard-setting panelists are likely to set a higher achievement level cut score on a test that is used for high school graduation than on a test used for school accountability but with no stakes for students. However, panelists' overall judgments may be more influenced by student impact data on a high-stakes graduation test than on a test used for school accountability, resulting in a lower achievement level cut score on the high-stakes graduation test.

As stated above, because it is virtually impossible to isolate and quantify the various factors that might influence the location of an achievement level cut score, it is best practice to exercise caution when comparing scores on tests with different intended uses and stakes for students.

### **Differences Among Assessment Contractors**

By design or through the peculiarities of the procurement process, it is often the case that the "same" large-scale assessment is administered by different assessment contractors within or across years and/or within or across states. When the goal is to produce comparable test scores across assessment contractors, there are threats to comparability that might be hidden beneath the surface, which may affect score comparability even when it appears that all of the same procedures are being applied by each contractor: test delivery, scoring, and psychometrics. In each case, additional layers of specifications and tests to confirm outputs and outcomes may be necessary before declaring scores comparable.

### *Test Delivery*

Different assessment contractors are likely to employ different test delivery platforms, whether they are using their own proprietary platform or an open-source platform. When test forms are administered on different platforms within or across years (by the same or different assessment contractors), it is critical to confirm that the test platforms function in the same manner or to understand differences in how they function. Even with test forms administered on a single test platform, it is critical to confirm that the test platform functions in the same manner across local networks and allowable response devices. Answers to the following questions may affect student performance on the test and, therefore, affect score comparability (Way, Davis, Keng, & Strain-Seymour, 2016):

- Do test items and directions render the same way on each platform?
- Are the procedures students must follow to navigate through the test similar across platforms?
- Are the tools and supports provided to students as accommodations and/or required for use to respond to questions similar across platforms and accessed in equivalent ways?
- Is the speed with which items load and responses are submitted consistent across platforms?

### *Scoring*

When student responses are scored by two or more scoring contractors, it is critical to monitor scoring and ensure that scoring is consistent across contractors. This may seem obvious for items that are human scored, but it is also necessary for items that are machine scored. It is true even for multiple-choice items with a single correct response, which seem very easy to score consistently across contractors.

In large-scale testing programs that are still paper based, differences in how student responses are read and processed may lead to differences in how a student response is scored. Dependent on the specifications provided, settings applied, and equipment and materials used, marks that are recorded as a response or erasure by one system may be treated as a blank by another system. Such differences could result not only in different item scores, but also in different flags being applied to students or test administrators through data forensics procedures.

With regard to human-scored items, it is assumed that assessment contractors will apply the same scoring rubrics, use the same training materials, and attempt to apply the same training procedures (CCSSO & ATP, 2010). It may also be assumed that contractors will apply the same rules for recruiting and qualifying scorers and monitoring scorer consistency and accuracy throughout the scoring process. With all of those safeguards in place, score comparability could still be affected by differences in how scoring contractors arbitrate and resolve score differences or by the thresholds set for when to rescore items from a scorer flagged for scorer drift or inconsistent scoring.

When student responses to constructed-response items are scored by automated scoring engines, it is critical to verify to the extent possible that a student response will be scored the same way regardless of the scoring engine and scoring algorithm used

if the goal is to produce comparable individual test scores. In such situations, it is as important to verify that there is consistency in student responses to which an automated system will not assign a score as well as to verify consistency when scores are assigned.

### *Psychometrics*

When psychometric analyses such as item calibration, equating, scoring, and scaling are conducted by different assessment contractors, as with item scoring described in the previous section, it is assumed that certain procedures will be held constant (i.e., the same IRT model applied, and the same rules and procedures used for equating). To a greater extent than with test delivery and scoring, the software used and decisions made during psychometric analyses can be a threat to score comparability.

Across IRT software packages, there may be different ways the same IRT models are executed, producing differences in results for some individual students. Within and across software packages there are also many decisions that must be made during item calibration and scoring that could affect the comparability of results. There may also be differences across assessment contractors in the procedures for selecting samples of responses to use for item calibration. It is likely impossible to develop rules for all of the decisions that must be made during psychometric analyses. Best practice, however, requires a demonstration that psychometric analyses applied to the same set of student responses produce comparable results.

## CONCLUSION

We began with the assertion that most often the desired level of comparability when considering individual student scores on the same test is that test scores are interchangeable. On K–12 state assessments in particular, there is an expectation and assumption that if individual students had received a different form of the same test they would have received the same test score (within measurement error) and that the individual scores of two students taking the same test can be compared to the same achievement standard or to each other.

Next, we established that the term “same test” refers to a wide variety of cases in which students are taking tests involving different sets of items; those students have access to a range of supports, tools, or test variations, as needed; and the students may be taking the test under different modes of administration with different response devices. In fact, in practice, the least likely situation to be encountered would be one that fits the traditional perception of standardized testing or the colloquial definition of the term “same test”—students taking the same set of items under the same testing conditions.

Based on those parameters, we discussed approaches to achieving the desired level of comparability of individual test scores and threats to achieving that level of comparability. Ultimately, however, it will be necessary for a test developer to provide evidence to support a claim of test score comparability and for a test user (e.g., policy maker, school administrator, teacher, parent, or student) to be able to evaluate and accept or reject that claim. As with most things related to educational measurement and large-scale assessment there is not a single test to determine test score comparability

nor is there a simple yes-or-no determination of test score comparability that applies to all situations. The process of establishing the comparability of individual student scores on the same test involves compiling sufficient evidence to support the claims and inferences that will be made about student performance based on those test scores.

Our starting point for evaluating the comparability of individual test scores on the same test is differences. There may be differences in test items, administration conditions, scoring procedures, students' opportunity to learn the content or become familiar with the test itself, and/or the intended uses of the test results. Determining whether there is sufficient test score comparability involves determining the degree to which those differences individually or cumulatively affect student performance or the interpretation of student performance.

We discussed design and psychometric approaches to producing comparable test scores or, at least, minimizing threats to comparability. There are standards, guidelines, and established best practices in large-scale assessments to increase the likelihood that test scores will be comparable. At times, the definition of best practices requires trade-offs, compromise, and an uneasy coexistence of measurement principles and policy priorities or government mandates. A combination of qualitative judgments and quantitative analyses is needed to evaluate both the application of best practices and their impact in producing comparable test scores across test forms.

Informed, expert qualitative judgment, for example, may be sufficient to approve changes to a test design such as reducing the test length by one point, one passage, or one performance task. Those initial judgments, however, must be supported by quantitative analyses of the impact of the proposed changes on the reliability of the assessment, measurement error associated with individual students' scores, or the accuracy and consistency of achievement level classifications. Similarly, statistical analyses may indicate that alternative test forms have the same level of difficulty, reliability, and relationship with external variables. It is likely, however, that qualitative analyses are also needed to determine whether inferences from test scores and claims about student performance can still be supported.

Unfortunately, there is no simple answer to the question of whether individual students' scores on the same test are comparable. As concluded by the Committee on Equivalency and Linkage of Educational Test two decades ago, "Ultimately, policy makers and educators must take responsibility for determining the degree to which they can tolerate imprecision in testing and linking ... and responsible people may reach different conclusions about the minimally acceptable level of precision in linkages that are intended to serve various goals" (NRC, 1999, p. 4). In this chapter, we have attempted to provide the tools with which policy makers and educators can make informed decisions on the extent to which test forms and administration conditions have been designed to support the conclusion that individual students' test scores are comparable as well as the extent to which forces external to the test may affect inferences about the performance of particular students taking the test.

Within large-scale state assessment programs, standards and best practices have been identified to ensure that tests produce scores that are sufficiently comparable by tightly specifying factors such as test content, format, administration, scoring, and intended uses. Understanding the factors that enhance and threaten comparability and

evaluating those factors with regard to a particular test and inference about student performance on that test is the ongoing responsibility of all test users.

## REFERENCES

- AERA (American Educational Research Association), APA (American Psychological Association), & NCME (National Council on Measurement in Education). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Banicky, L. A. (2000). *Opportunity to learn (Policy Brief Vol. 7)*. Retrieved August 1, 2019, from <http://udspace.udel.edu/bitstream/handle/19716/2446/opp+to+learn.pdf?sequence=>
- Carroll, J. B. (1989). The Carroll model: A 25-year retrospective and prospective view. *Educational Researcher*, 18(1), 26–31.
- CCSSO (Council of Chief State School Officers) & ATP (The Association of Test Publishers). (2010). Scoring open-ended responses. In D. Eignor (Ed.), *Operational best practices* (Chapter 14). Washington, DC: CCSSO and ATP.
- Cooper, R., & Liou, D. D. (2007). The structure and culture of information pathways: Rethinking opportunity to learn in urban high schools during the ninth grade transition. *The High School Journal*, 91(1), 43–56.
- Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). *Principles and practices of test score equating*. Princeton, NJ: Educational Testing Service.
- Friedlaender, D., & Darling-Hammond, L. (with Andree, A., Lewis-Charp, H., McCloskey, L., Richardson, N., et al.). (2007). *High schools for equity: Policy supports for student learning in communities of color*. Stanford, CA: School Redesign Network at Stanford University. Retrieved from [www.srnleads.org/resources/publications/pdf/hsfe/hsfe\\_report.pdf](http://www.srnleads.org/resources/publications/pdf/hsfe/hsfe_report.pdf).
- Haertel, E., Moss, P., Pullin, D., & Gee, J. (2008). Introduction. In P. Moss, D. Pullin, J. Gee, E. Haertel, & L. Young (Eds.), *Learning in doing: Social, cognitive and computational perspectives. Assessment, equity, and opportunity to learn* (pp. 1–16). Cambridge, UK: Cambridge University Press. <http://doi.org/10.1017/CBO9780511802157.003>.
- Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: Praeger.
- Kurz, A. (2011). Access to what should be taught and will be tested: Students' opportunity to learn the intended curriculum. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *The handbook of accessible achievement tests for all students: Bridging the gaps between research, practice, and policy*. New York: Springer.
- Lehr, C., & Thurlow, M. (2003). *Putting it all together: Including students with disabilities in assessment and accountability systems* (Policy Directions No. 16). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved April 30, 2019, from <https://nceo.info/Resources/publications/OnlinePubs/Policy16.htm>.
- Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, 6(1), 83–102.
- MA DESE (Massachusetts Department of Elementary and Secondary Education). (2019). *Test administrator manual for computer-based testing, spring 2019*. Retrieved August 1, 2019, from <http://www.doe.mass.edu/mcas/testadmin/manual/TAM-CBT.pdf>.
- Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects* (Policy Information Rep.). Princeton, NJ: Educational Testing Service.
- NRC (National Research Council). (1999). *Uncommon measures: Equivalence and linkage among educational tests* (M. J. Feuer, P. W. Holland, B. F. Green, M. W. Bertenthal, & F. Cadelle Hemphill, Eds.). Committee on Equivalency and Linkage of Educational Tests. Board on Testing and Assessment. Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.
- PARCC (Partnership for Assessment of Readiness for College and Careers). (2017). *PARCC accessibility features and accommodations manual*. Retrieved April 30, 2019, from <https://parcc.pearson.com/manuals>.
- Pullin, D., & Haertel, E. (2008). Assessment through the lens of "opportunity to learn." In P. Moss, D. Pullin, J. Gee, E. Haertel, & L. Young (Eds.), *Learning in doing: Social, cognitive and computational perspectives. Assessment, equity, and opportunity to learn* (pp. 17–41). Cambridge, UK: Cambridge University Press. <http://doi.org/10.1017/CBO9780511802157.004>.
- Shepard, L., Hannaway, J., & Baker, E. (Eds.). (2009). *Education policy white paper on standards, assessments, and accountability*. Washington, DC: National Academy of Education.

- Smarter Balanced Assessment Consortium. (2016). Test administration. In *Smarter Balanced Assessment Consortium: 2014-15 technical report* (Chapter 6). Retrieved April 30, 2019, from <https://portal.smarterbalanced.org/library/en/2014-15-technical-report.pdf>.
- Smarter Balanced Assessment Consortium. (2018). *Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations guidelines*. Retrieved April 30, 2019, from <https://portal.smarterbalanced.org/library/en/usabilityaccessibility-and-accommodations-guidelines.pdf>.
- Steedle, J. T., & Grochowalski, J. (2017). The effect of stakes on accountability test scores and pass rates. *Educational Assessment, 22*(2), 111–123. <http://doi.org/10.1080/10627197.2017.1309276>.
- Way, W. D., Davis, L. L., Keng, L., & Strain-Seymour, E. (2016). From standardization to personalization. In F. Drasgow (Ed.), *Technology and testing*. New York: Routledge.
- Winter, P. C. (2010). Comparability and test variations. In P. C. Winter (Ed.), *Evaluating the comparability of scores from test variations*. Washington, DC: Council of Chief State School Officers.
- Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1–17. [http://doi.org/10.1207/s15326977ea1001\\_1](http://doi.org/10.1207/s15326977ea1001_1).
- Zieky, M. (2014). An introduction to the use of evidence-centered design in test development. *Psicologia Educativa, 20*(2), 79–87.