

Creating and Evaluating Effective Educational Assessments

Chapter 3 Comparability



This digital workbook on educational assessment design and evaluation was developed by edCount, LLC, under Enhanced Assessment Grants Program, CFDA 84.368A.

1



Creating and Evaluating Effective Educational Assessments

Chapter 3: Comparability

Welcome to the third of five chapters in a digital workbook on educational assessment design and evaluation. This workbook is intended to help educators ensure that the assessments they use provide meaningful information about what students know and can do.

This digital workbook was developed by edCount, LLC, under the US Department of Education's Enhanced Assessment Grants Program, CFDA 84.368A.



Strengthening Claims-based Interpretations and Uses of Local and Large-scale Science Assessment Scores

2



The grant project is titled the [Strengthening Claims-based Interpretations and Uses of Local and Large-scale Science Assessment Scores...](#)



Strengthening
Claims-based
Interpretations and Uses of
Local and
Large-scale
Science Assessment
Scores

3

edCount²
Measuring up to the future

or its acronym, “SCILLSS.”

Chapter 3.1

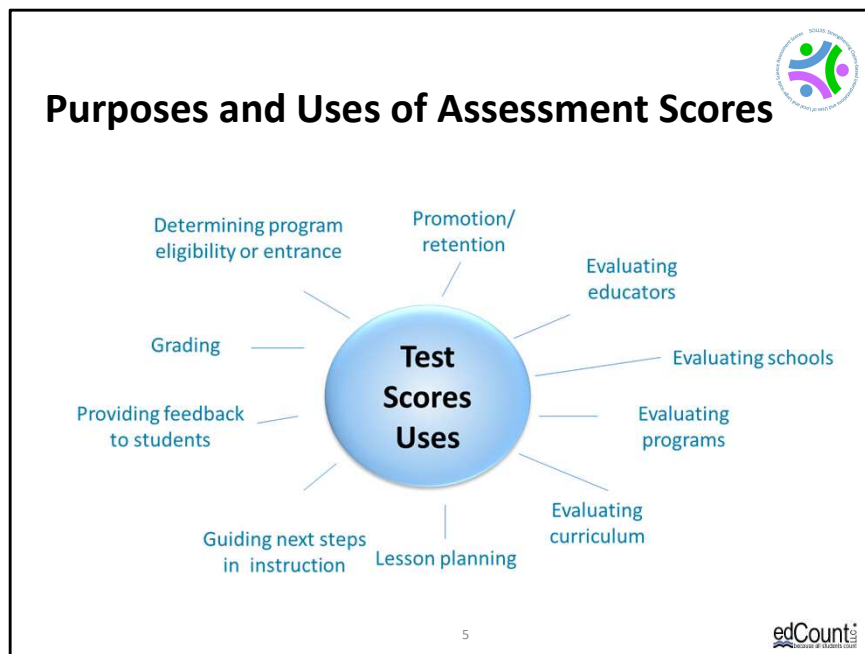


Review of Key Concepts from Chapters 1 and 2

4

edCount²
Measuring Up to the Future

Chapter 3.1. Review of Key Concepts from Chapters 1 and 2



Let's begin with a brief recap of the key concepts covered in the first two chapters of this series.

Chapter 1 focused on common reasons why we administer assessments of students' academic knowledge and skills and how we use those assessment scores. We learned that these purposes for administering assessments and the intended uses of assessment scores should drive all decisions about how assessments are designed, built, and evaluated.

Validity in Assessments



No test can be valid in and of itself.

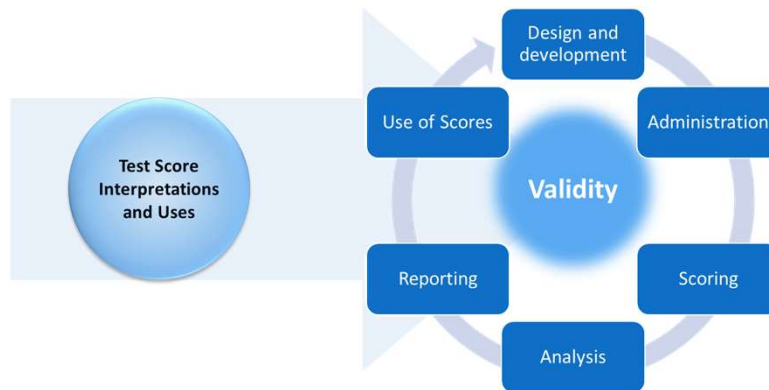
Validity depends on the strength of the evidence regarding what a test measures and how its scores can be interpreted and used.

6

edCount
Measuring up to the future

We learned in chapter 1 that validity relates to the interpretation and use of assessments scores and not to tests themselves. Validity is a judgment about the meaning of assessment scores and about how they are used.

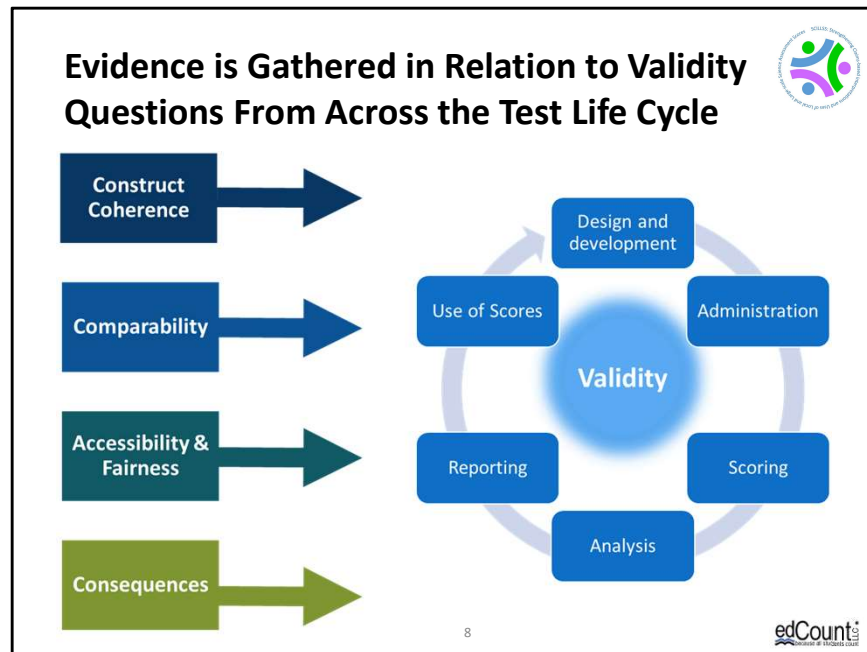
Purposes and Uses of Assessment Scores Drive All Decisions About Tests



7

edCount
Measuring up to the future

We evaluate validity by gathering and judging evidence. This validity evidence is gathered from across the entire life cycle of a test from design and development through score use. Judgments about validity are based upon the quality and adequacy of this evidence in relation to assessment score interpretations and uses. Depending upon the nature of the evidence, score interpretations can be judged as valid or not. Likewise, particular uses of those scores may or may not be supported depending upon the degree and quality of the validity evidence.



Chapter 1 also included a brief overview of four fundamental validity questions that provide a framework for how to think about validity evidence. These four questions represent broad categories and each subsumes many other questions.

The four validity question categories are:

- **Construct coherence:** To what extent do the test scores reflect the knowledge and skills we're intending to measure, for example, those defined in the academic content standards?
- **Comparability:** To what extent are the test scores reliable and consistent in meaning across all students, classes, schools, and time?
- **Accessibility and fairness:** To what extent does the test allow all students to demonstrate what they know and can do? And
- **Consequences:** To what extent are the test scores used appropriately to achieve specific goals?

Construct Coherence



1. What are you intending to measure with this test? We'll refer to the specific constructs we intend to measure as measurement targets.
2. How was the assessment developed to measure these measurement targets?
3. How were items reviewed and evaluated during the development process to ensure they appropriately address the intended measurement targets and not other content, skills, or irrelevant student characteristics?
4. How are items scored in ways that allow students to demonstrate, and scorers to recognize and evaluate, their knowledge and skills? How are the scoring processes evaluated to ensure they accurately capture and assign value to students' responses?
5. How are scores for individual items combined to yield a total test score? What evidence supports the meaning of this total score in relation to the measurement target(s)? How do items contribute to subscores and what evidence supports the meaning of these subscores?
6. What independent evidence supports the alignment of the assessment items and forms to the measurement targets?
7. How are scores reported in relation to the measurement targets? Do the reports provide adequate guidance for interpreting and using the scores?

9



Chapter 2 of this digital workbook focused on the first set of these questions, construct coherence. We addressed the types of evidence that relate to seven key construct coherence questions.

1. What are the measurement targets for this test? That is, what are you intending to measure with this test?
2. How was the assessment developed to measure these measurement targets?
3. How were items reviewed and evaluated during the development process to ensure they appropriately address the intended measurement targets and not other content, skills, or irrelevant student characteristics?
4. How are items scored in ways that allowed students to demonstrate, and scorers to recognize and evaluate, their knowledge and skills? How are the scoring processes evaluated to ensure they accurately capture and assign value to students' responses?
5. How are scores for individual items combined to yield a total test score? What

evidence supports the meaning of this total score in relation to the measurement target(s)? How do items contribute to subscores and what evidence supports the meaning of these subscores?

6. What independent evidence supports the alignment of the assessment items and forms to the measurement targets? And,
7. How are scores reported in relation to the measurement targets? Do the reports provide adequate guidance for interpreting and using the scores?

In this chapter, we turn our attention to the second set of validity questions, which relate to the notion of comparability.

Chapter 3.2



What is Comparability and Why is it Important?

10

edCount²
The Missouri Department of Education

Chapter 3.2: What is Comparability and Why is it Important?

Key Points in this Chapter:



- Most test score uses require some type of comparability evidence;
- Reliability/precision is necessary to support comparable meaning of scores across students, classes, schools, test forms and formats, and time; and
- Evidence of comparability can take different forms and the kinds of evidence that are most important depends on the intended meaning and use of the scores.

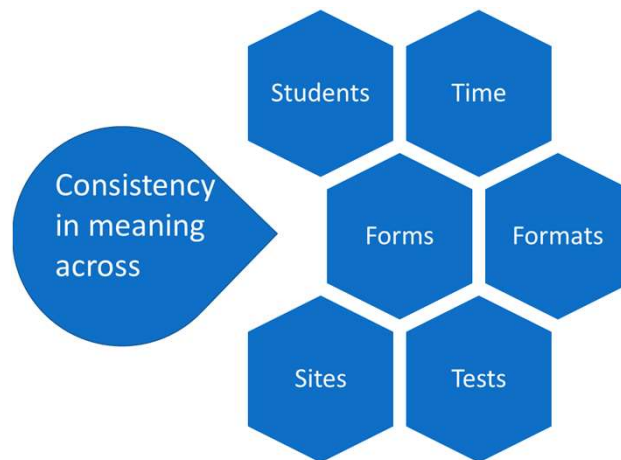
11



Our purposes in this chapter are to help educators strengthen their understanding of comparability by addressing several key points. These include:

- Most test score uses require some type of comparability evidence;
- Reliability/precision is necessary to support comparable meanings of scores across students, classes, schools, test forms and formats, and time; and
- Evidence of comparability can take different forms and the kinds of evidence that are most important depends on the intended meaning and use of the scores.

Comparability: Consistency in Meaning Across Variations



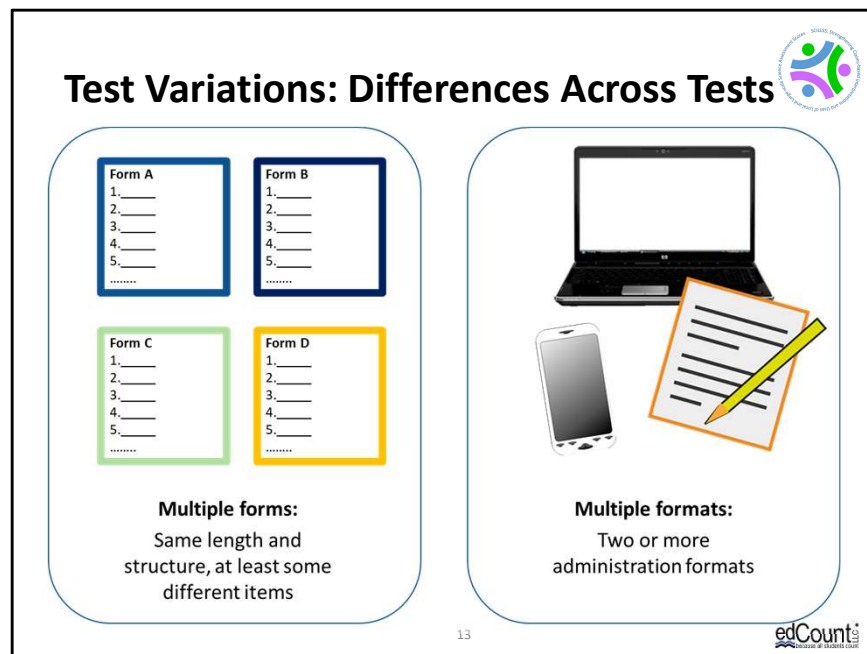
12

edCount
California's Data Center

Comparability for those building tests or using test scores relates to consistency in the meaning of test scores across variations including students, time, sites, forms or formats of the test, and different tests altogether.

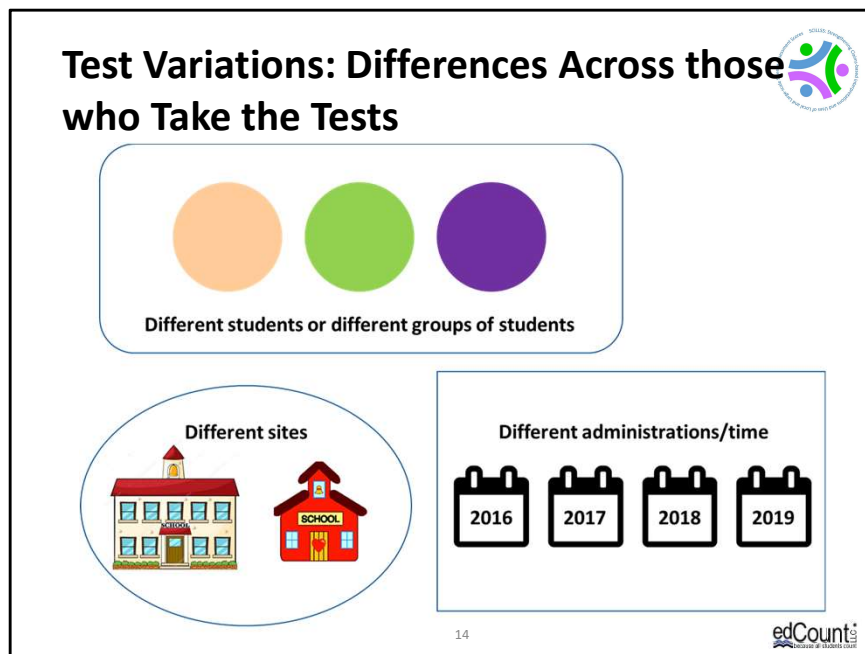
If scores vary in their meaning across forms or time or students or across any other dimension, those using these scores must understand how this variation affects score interpretation and the use of the scores for making decisions.

Evidence of comparability is important even when the scores are simply being combined, such as when one calculates an average for a class or a school, because such calculations rely on comparable interpretations of those scores across the students whose scores are being aggregated.



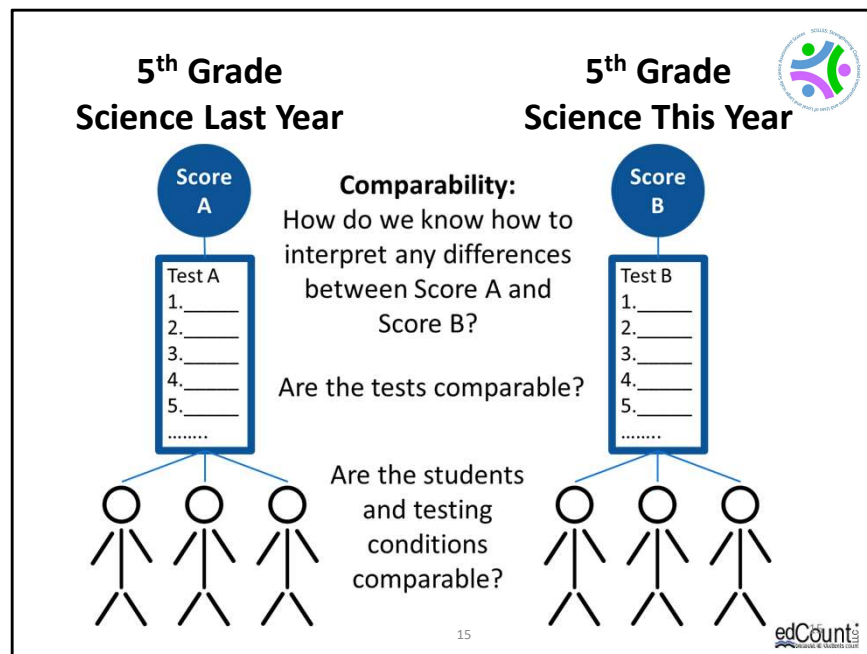
Test variations relate to differences or potential differences in the tests or those who take them. Differences in tests include:

- **Forms:** If two or more forms of a test are administered, which is very often the case for large-scale assessments, the vendor must provide evidence that the scores from the various forms are comparable. Different forms of a test typically have the same structure and length, as defined in the test blueprint, but include different test items.
- **Formats:** When tests are given in two or more formats, such as paper-and-pencil and also on computers or other devices, those creating these formats must provide evidence that the scores from the various formats are comparable.



Differences in those who take the tests include:

- **Students:** Any time two or more students take a test, there can be questions about comparability. These differences can relate to student characteristics such as disabilities and English proficiency as well as to motivation and opportunities to learn the material the test is covering. Differences across students also raise a number of accessibility and fairness issues and we will address those in chapter 4 of this series.
- **Sites:** Variations across classrooms, schools, districts, and states can disrupt score comparability, which is why there are often strong efforts to standardized testing conditions.
- **Administrations or time:** Students may be participating in the same or equivalent tests at different points in time, such as at the end of grade 5 each year, and if you want to compare scores across those administrations, you need evidence of comparability.



Given that there are several types of variation for almost every test, most tests should have multiple types of evidence of comparability.

Let's say you wish to compare performance of the 5th grade science students this year to the 5th grade science students from last year. To use test scores to make such a comparison would require evidence that the tests these two groups of students took yielded comparable scores.

Necessary comparability evidence would indicate whether the forms of the test each group of students took were equivalent, whether the conditions under which students took the test were identical or nearly so, and whether the groups of students were similar. In some cases, statisticians who specialize in working with assessment data, known as psychometricians, can account for variations statistically. That is, they can evaluate the differences across forms or students and make adjustments in the score scales that allow for comparable score interpretations. These statistics would also be considered evidence related to comparability.

We'll consider what some of this statistical and psychometric evidence might entail as we walk through the validity questions in this chapter. Before we get to those, we'll

turn our attention to the concept of reliability or precision because of its importance for all test scores under all circumstances.

Chapter 3.3



What is Reliability/Precision and Why is it Important?

16

edCount²
The Missouri Department of Education

Chapter 3.3: What is reliability/precision and why is it important?

Our Standards: Reliability/Precision



- **Reliability/Precision:** The degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and thence are inferred to be dependable and consistent for an individual test taker; the degree to which scores are free of random errors of measurement for a given group.

(AERA, APA, NCME, 2014, pp. 222-223)

- **Standard 2.0:** Appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use.

(AERA, APA, & NCME, 2014, p. 42)

17

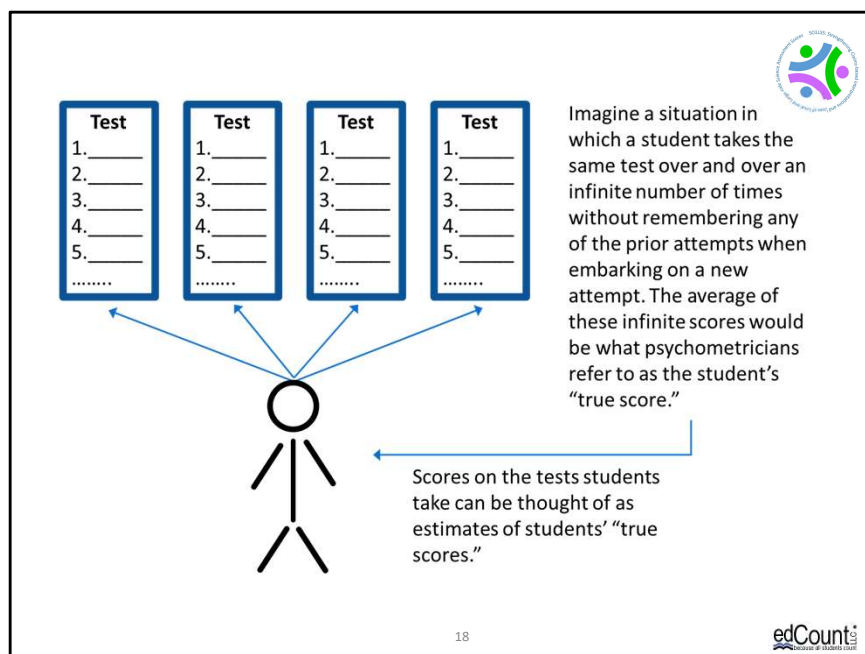


Reliability/precision is a necessary, but not sufficient condition, for comparability. We use the term “reliability/precision” rather than just “reliability” to indicate that we are not limiting our conceptualization of reliability as captured in correlation coefficients between scores on equivalent forms of a test.

The *Standards for Educational and Psychological Testing* define reliability/precision as the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and thence are inferred to be dependable and consistent for an individual test taker; the degree to which scores are free of random errors of measurement for a given group.

Standard 2.0, the first standard in the chapter on reliability in the *Standards*, says that, “appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use.”

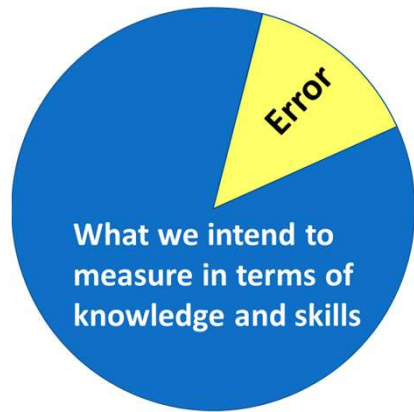
Like validity, reliability and precision relate to the scores and not to the tests themselves.



To understand reliability/precision, it's important to recognize that all test scores reflect some degree of error. No matter how carefully a test has been designed and administered, every score for every student and every group of students includes some error. In part, that's because every test is just a sample of a student's behavior. Let's think about this for a minute.

Imagine a situation in which a student takes the same test over and over an infinite number of times without remembering any of the prior attempts when embarking on a new attempt. The average of these infinite scores would be what psychometricians refer to as the student's true score. Of course, this is merely a thought experiment and not possible in practice. However, the point is that each of the attempts is a sample of evidence that results in a score and each of those scores is just an estimate of what the student actually knows and can do. In this way, every test score is just an estimate of actual knowledge and skills. We cannot know a student's true score, so we use statistics to estimate how close actual test scores – which always reflect some error – may be to that hypothetical, error-free, true score. These statistics allow us to estimate how reliable and precise test scores are as indicators of students' true scores.

Every Test Score includes Two Components

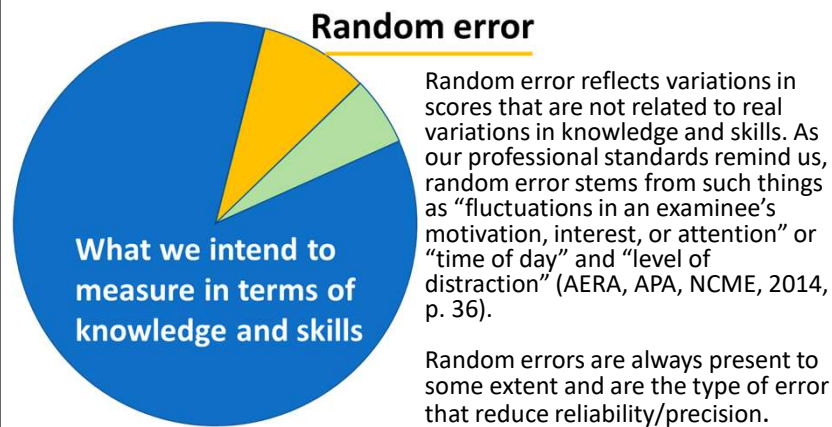


19

edCount²
Measuring what matters

Every test score, which we now know can only be an estimate of a student's true score, has two components. The first component reflects some degree of what we intend to measure in terms of knowledge and skills. The second component is error. All test scores and all item scores reflect a combination of what we intend to measure and error. Several factors contribute to error, including the sampling error that is always present because every student or group of students who takes a test is considered a sample of the student population. When building, administering, and scoring tests, we must take specific steps to maximize the part of the score that reflects knowledge and skills and to minimize the part that reflects error.

Two Types of Error in All Test Scores— Random Error

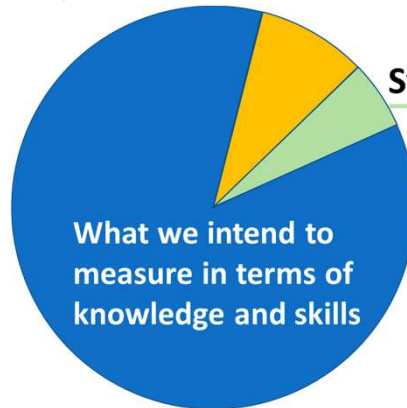


20

edCount
Measuring up to the future

In minimizing error, a key consideration is the source of the error. Some error is considered “random.” As our professional standards remind us, random error stems from such things as “fluctuations in an examinee’s motivation, interest, or attention” or “time of day” and “level of distraction” (AERA, APA, NCME, 2014, p. 36). Random errors are always present to some extent and are the component of error that reduces reliability/precision.

Two Types of Error in All Test Scores— Systematic Error



Systematic error

Systematic errors stem from sources that affect performance in a consistent manner across one or more groups of students or items or across test administrations.

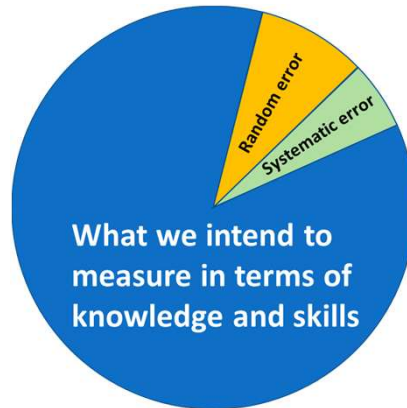
Systematic errors hinder validity because they represent what is called “construct-irrelevant variation” in what a test measures.

21

edCount
California's Data Center

The other part of error is systematic. Systematic errors stem from sources that affect performance in a consistent manner across one or more groups of students or items or across test administrations. An incorrect answer key, the use of language or formats that some students, such as students with disabilities or English learners, cannot access, and administrations of test forms that are not equivalent across students or time are examples of systematic errors. Unlike random errors, systematic errors do not reduce reliability/precision. Rather, systematic errors hinder validity because they represent what is called “construct-irrelevant variation” in what a test measures.

Reliability/Precision and Comparability



Comparability depends in part upon maximizing the component of each test that reflects what we intend to measure in terms of knowledge and skills and minimizing the component of each test that reflects error.

22

edCount
Measuring up to the future

Each of these sources of error is estimated in different ways and there are differences in what those building, administering, and scoring tests can do to minimize them. As our professional standards remind us, those building tests are obligated to evaluate and document error and provide test users with information about how error may affect score interpretations. We will provide some additional information about what to expect from test vendors in their reporting of score reliability and precision later in this chapter.

Here, we highlight that comparability depends in part upon maximizing the component of each test score that reflects what we intend to measure in terms of knowledge and skills and minimizing the component of each test score that reflects error. The greater the error in test scores, the less appropriate and meaningful it is to compare or combine those scores.

Chapter 3.4



Validity Questions Related to Comparability and Reliability/Precision

23

edCount²
The Missouri Department of Education

Chapter 3.4. Validity questions related to comparability and reliability/precision

Comparability Questions



1. How is the assessment designed to support comparability of scores across forms and formats?
2. How is the assessment designed and administered to support comparable score interpretations across students, sites (classrooms, schools, districts, states), and time?
3. How are student responses scored such that scores accurately reflect students' knowledge and skills across variations in test forms, formats, sites, scorers, and time?
4. How are score scales created and test forms equated to support appropriate comparisons of scores across forms, formats, and time?
5. To what extent are different groups of students who take a test in different sites or at different times comparable?
6. How are scores reported in ways that support appropriate interpretations about comparability and disrupt inappropriate comparability interpretations?
7. What evidence supports the appropriate use of the scores in making comparisons across students, sites, forms, formats, and time?

24



In this section, we will focus on seven validity questions related to comparability and reliability/precision that those building or adopting tests should consider carefully as they make their decisions. These include:

1. How is the assessment designed to support comparability of scores across forms and formats?
2. How is the assessment designed and administered to support comparable score interpretations across students, sites (such as classrooms, schools, districts, and states), and time?
3. How are student responses scored such that scores accurately reflect students' knowledge and skills across variations in test forms, formats, sites, scorers, and time?
4. How are score scales created and test forms equated to support appropriate comparisons of scores across forms, formats, and time?
5. To what extent are different groups of students who take a test in different sites

or at different times comparable?

6. How are scores reported in ways that appropriately support comparability in score interpretation and use?
7. What evidence supports the appropriate use of the scores involving comparisons across students, sites, forms, formats, and time?

Comparability Across Forms and Formats



1. How is the assessment designed to support comparability of scores across forms and formats?




25

edCount³
Measuring up to the future

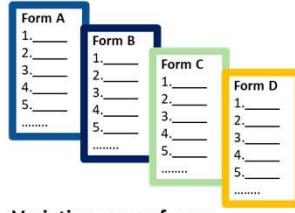
We'll start with the first of our validity questions for comparability:

1. How is the assessment designed to support comparability of scores across forms and formats?


Evidence related to this question would come primarily from the test design and development phase.




Test Variations: Differences Across Tests





Variation across forms of a paper-and-pencil, computer-based, or computer-adaptive test





Variation across formats in which a test is administered



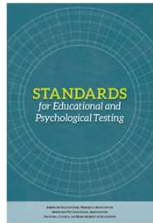
26


Recall that forms of test are different versions that have the same look and feel but include at least some items that are different on one version than on the other. Different forms of a test are meant to yield scores that have the same meaning even though the items on them are different. As long as the sets of items are meant to measure the same thing in the same manner, we consider these different forms rather than different tests altogether.

One common extension of the test form idea involves a computer-adaptive testing model, or CAT. While many tests are considered “linear” in that all students taking one form of the test see the same items in the same order, a CAT presents different items to different students depending on how students answer previous questions. This is an extreme version of multiple test forms with every student getting a different test form.

Different formats for a test means that the test is offered in, say, a paper-and-pencil version and on the computer. Or, on a PC and on a tablet. Even when different formats include exactly the same items, differences in the presentation modes may affect how students interact with the items. That would result in variations in the meaning of the scores across the formats.

Our Standards: Comparability Across Forms



- **Standard 5.12:** A clear rationale and supporting evidence should be provided for any claim that scale scores earned on alternate forms of a test may be used interchangeably.

(AERA, APA, NCME, 2014, p. 105)

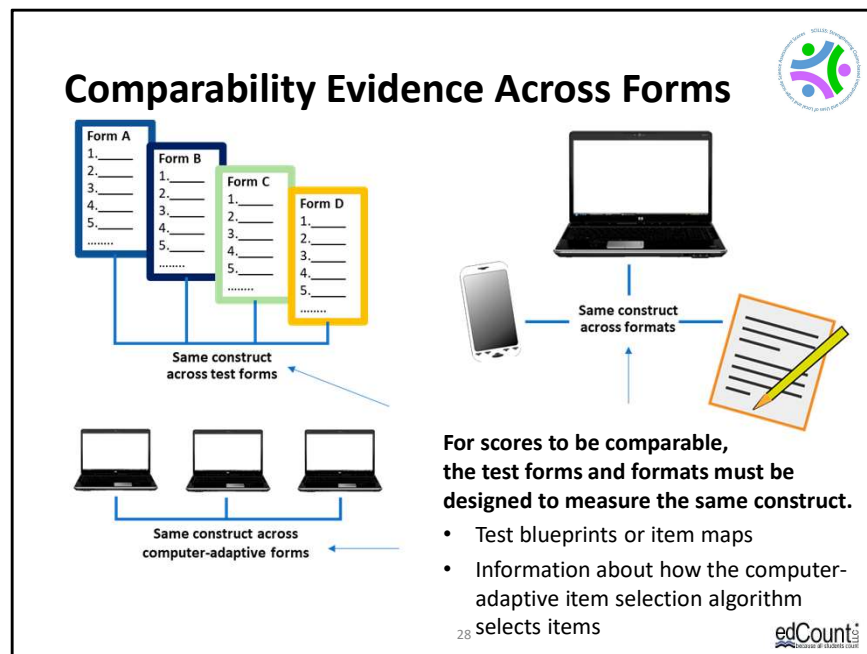
The testing publisher or vendor is responsible for providing this rationale and supporting evidence.

27



Our professional standards define many expectations related to comparability of forms and formats. For example, standard 5.12 states that, “A clear rationale and supporting evidence should be provided for any claim that scale scores earned on alternate forms of a test may be used interchangeably.”

The testing publisher or vendor is responsible for providing this rationale and supporting evidence.



At a minimum, for the scores from different forms of a test or from different testing formats to be considered comparable, the forms and formats must be designed to measure the same constructs. Therefore, the rationale and evidence should include information about the blueprint or test map or any other methods used to create comparable forms. Test blueprints or maps define the set of items that make-up the test in terms of how many items, what kinds of items, and what each item is supposed to measure. This information would be found in technical documentation describing test development and forms construction processes.

For CATs, the publisher or vendor must describe how the algorithm that selects the test items for each student does so in a way that covers the same content and skills for every administration. Even when a CAT algorithm is supposed to tailor its selection of items to meet a student's particular needs, if the scores are meant to be comparable across students then the content and skills covered on the test must be, too. This information would be found in technical documentation describing test development and item selection rules.

Comparability Evidence Across Forms



The technical documentation for a test must include:

- Test blueprints or item maps;
- Information about how the computer-adaptive item selection algorithm selects items;
- Results from studies of student performance on different forms and formats; and
- Information about the nature and quality of the scaling and equating procedures.

29



Evidence of comparability across formats should also include results from studies of student performance on administrations using those formats. Typically, these studies involve analysis of test performance for equivalent groups of students who take the test in different formats. In addition to considering differences in total test scores across these groups, the publisher must evaluate differences in how students perform on the items within the test. If there are differences in test scores or performance on items, the publisher must examine reasons for these differences and make adjustments to the tests if the scores are to be comparable. Information about these kinds of studies and how a publisher has ensured comparability across formats would be found in the technical documentation describing test development, scaling and equating methods, or validity studies.

Comparability Across Students, Site, and Time



- 2. How is the assessment designed and administered to support comparable score interpretations across students, sites (classrooms, schools, districts, states), and time?**



30

edCount²
Measuring up to the future

Our second validity question for comparability is:

2. How is the assessment designed and administered to support comparable score interpretations across students, sites (such as classrooms, schools, districts, or states), and time?

Evidence for this question would come primarily from the test design and development and the administration phases.

Our Standards: Comparability Across Administrations



- **Standard 6.0:** To support useful interpretation of score results, assessment instruments should have established procedures for test administration, scoring, reporting, and interpretation. Those responsible for administering, scoring, reporting, and interpreting should have sufficient training and supports to help them follow the established procedures. Adherence to the established procedures should be monitored, and any material errors should be documented and, if possible, corrected.
(AERA, APA, NCME, 2014, p. 114)
- **Standard 6.1:** Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user.
(AERA, APA, NCME, 2014, p. 115)
- **Standard 6.3:** Changes or disruptions to standardized test administration procedures or scoring should be documented and reported to the test user.
(AERA, APA, NCME, 2014, p. 115)

31



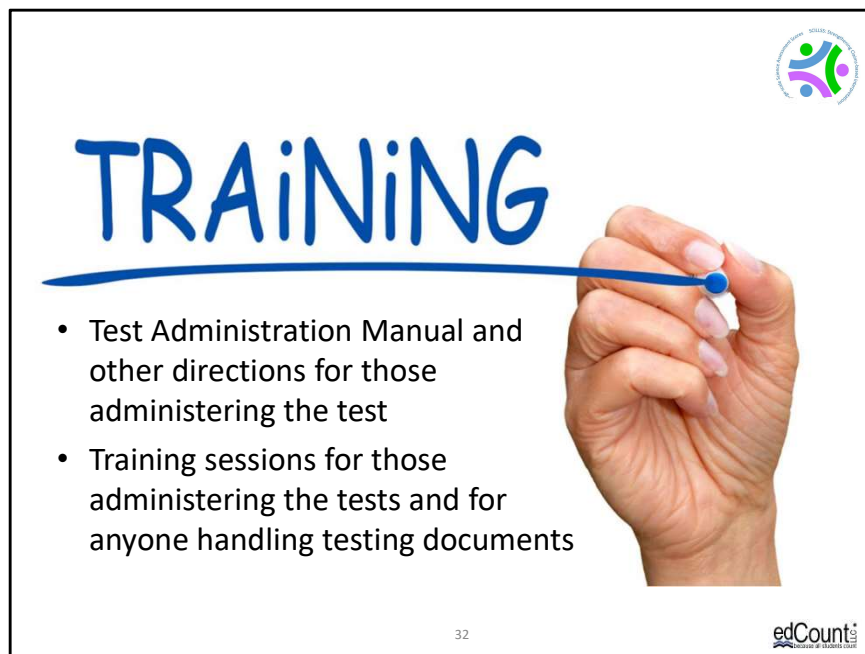
As was the case for the forms and formats question, a test publisher or vendor must provide evidence that its tests are measuring the same constructs even when they are administered at different times to different students in different places. Critical parts of the evidence for this question relate to the guidelines for how a test is administered and evidence that the test was actually administered according to these guidelines.

The *Standards for Educational and Psychological Testing* provide clear expectations regarding test administration and its contribution to the comparability of scores.

Standard 6.0 highlights the obligations of both test publishers and test users to establish and evaluate processes for supporting comparability of scores. “To support useful interpretation of score results, assessment instruments should have established procedures for test administration, scoring, reporting, and interpretation. Those responsible for administering, scoring, reporting, and interpreting should have sufficient training and supports to help them follow the established procedures. Adherence to the established procedures should be monitored, and any material errors should be documented and, if possible, corrected.”

Standards 6.1 and 6.3 further underscore these obligations. “Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user.” “Changes or disruptions to standardized test administration procedures or scoring should be documented and reported to the test user.”

Clearly, the test publisher or vendor has an obligation to set the guidelines for how a test is administered, but it is up to those using the test to follow those guidelines and document any deviations from them.



TRAINING

- Test Administration Manual and other directions for those administering the test
- Training sessions for those administering the tests and for anyone handling testing documents

32


edCount

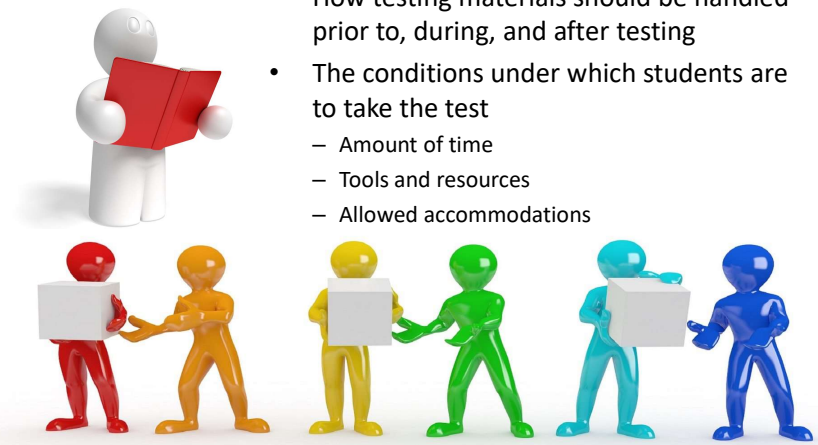
This shared responsibility means that evidence related to this question comes in part from the documentation and guidance the test publisher provides and in part from documentation gathered during the process of test administration.

In all cases, those using the tests should take great care in learning how to administer the tests. For large-scale assessments, such as annual statewide tests, the publisher or vendor will provide a test administration manual as well as means for training school staff to administer the tests. This is part of the evidence necessary to support comparability.


Local obligations related to comparability involve preparing for and administering the test. Someone at the local level is typically designated as the person responsible for coordinating the testing process by ensuring that the materials are kept secure, that those proctoring the test are appropriately trained, that the tests are given to the appropriate students during the specified testing period or window, and that the conditions under which students are taking the test conform to the specifications in the administration guidelines.

Test Administration Manual or Guidelines





- How testing materials should be handled prior to, during, and after testing
- The conditions under which students are to take the test
 - Amount of time
 - Tools and resources
 - Allowed accommodations



33


The administration guidelines, which are usually presented in the test administration manuals and other documents the publisher provides prior to the testing window, should always specify whether students are allowed to use resources such as dictionaries, formula sheets, or calculators while they are testing. Constraints on the use of these kinds of supports are necessary to ensure that all students take the test under the same conditions.

Of course, some students with disabilities and some English learners do require some adjustments to the standard administration procedures. Such adjustments are known as testing accommodations and we will address these in depth in the fourth chapter of this digital workbook.

Decisions about what supports students are and are not allowed to use while testing should depend in part on individual students' needs and also on what the test is meant to be measuring. For example, if a mathematics test is meant to be measuring students' computational skills, then it would not be appropriate for students to use calculators on the portions of the test with items measuring those skills. On the other hand, if a test is meant to be measuring other knowledge and skills, such as whether a student knows how to calculate mass or volume, then a calculator could be helpful


tool. A test administration manual should clearly specify what tools are allowed and it may be necessary to consult the technical manuals or the publisher directly for evidence to support decisions about which tools are and are not allowed and why. Using tools that are not allowed can change the meaning of the test scores and render them non-comparable.



Breaches of security dilute the scores for entire classrooms, grades, schools, districts, or even states. Scores from a breached test may no longer reflect students' knowledge and skills and cannot be combined with or compared to other scores.

Breaches in tests such as large-scale statewide assessments result in significant costs to the state and its taxpayers as well as a loss of students' time and information about them.

34



One of the biggest threats to the meaning of scores and their comparability across students, time, and locations is test security. Breaches of test security occur when, for example, individuals in a school or school district inappropriately assist students as they are taking a test, change students' answers on the test, or use information about the test questions to prepare students for the test.

When administering their own classroom tests, teachers generally take steps to prevent students from cheating by, for example, making sure students don't bring in "crib sheets" and can't copy answers from other students. The reason teachers want to prevent cheating is because they want each student's score to reflect that student's knowledge and skills. Cheating makes it impossible to know what a student knows and can do.

The same logic applies to large-scale tests, only the damage extends beyond an individual student. Breaches of security, such as those described above, can dilute the scores for entire classrooms, grades, schools, districts, or even states. Scores from a breached test may no longer reflect students' knowledge and skills and cannot be combined with or compared to other scores. A breach means that we cannot know if students' skills have improved from 4th grade to 5th grade, if the 5th graders are

improving in their science skills over time, or if a school or district program is effective in supporting student learning. Breaches in tests such as large-scale statewide assessments result in significant costs to the state and its taxpayers as well as a loss of students' time and information about them.

In summary, the evidence for our second comparability question comes from both the test publisher or vendor, in the form of guidance about test administration conditions, and from the adherence to that guidance from those using the test.

Comparability Across Scorers and Scoring Methods



- 3. How are student responses scored such that scores accurately reflect students' knowledge and skills across variations in test forms, formats, sites, scorers, and time?**



35

edCount
Measuring up to the future

Our third validity question related to score comparability is:

3. How are student responses scored such that scores accurately reflect students' knowledge and skills across variations in test forms, formats, sites, scorers, and time?

Evidence related to this question comes from the test design and development and scoring phases of the testing life cycle.

Accurate and reliable scoring of selected-response items requires that...



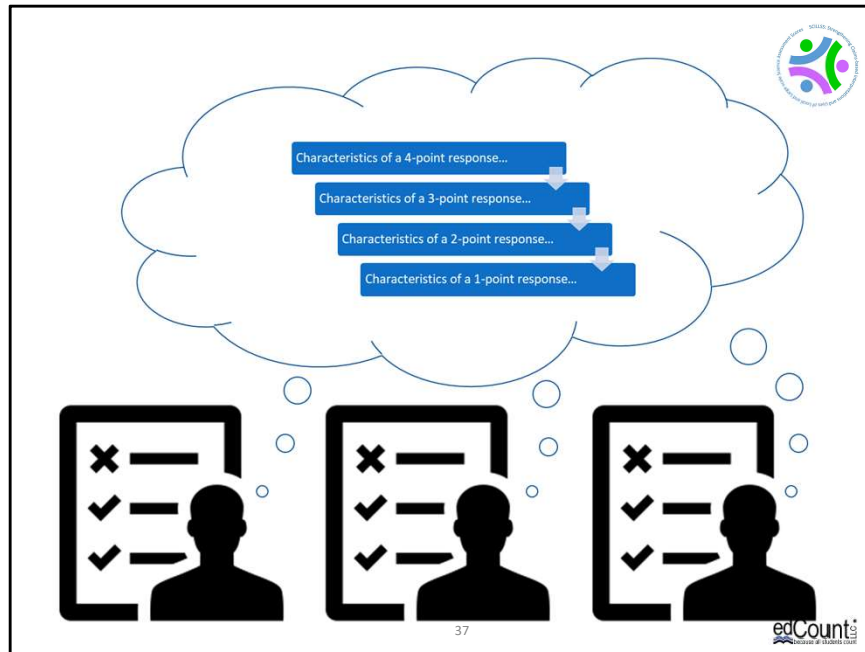
- The options meant to be scored as correct must actually be correct;
- The options meant to be scored as not correct must actually be incorrect;
- The answer key must be correct and be used accurately; and
- The rules for when students pick more options than allowed are clear and applied.



36

edCount
Measuring up to the future

Although students' answers to test questions cannot be scored until after the test has been administered, preparation for scoring must occur early in the test design and development phase. Item writers should craft items so that correct answers can be recognized and distinguished from answers that are not correct. Even for selected-response questions, where a student has to pick her answer from two or more options, test developers have to indicate which option is correct, be sure it actually is correct, be sure the other options are not also correct, and set rules for what happens if a student selects, for example, the correct option and also another option.



Rules for scoring constructed-response items are even more complicated. Scorers are responsible for translating students' responses into the language of the test and must not impose their personal beliefs or biases on that process. Their job is to apply the rules so that students' knowledge and skills are appropriately recognized.

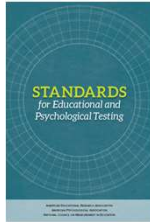
To support high quality scoring, item writers must develop the scoring rubrics when they develop the items and specify how all possible responses would be scored. What happens if a student's response includes some, but not all, characteristics of the third level of a four-level rubric? Does it get a 3 or a 4? What if the content of the response is correct, but there are spelling and grammatical errors? Does that matter to the score? What happens when responses are blank or off-topic or include profane language? Is each response to be scored by one person? Two? What happens if the two scorers don't agree?

Consider a local assessment context where all students in a grade are given a writing prompt and teachers in that grade score these responses. What would the teachers need to help ensure that their scores are accurate? Clear guidance, training, and probably a strategy where teachers score responses from students who are not their own and whose names they do not know. Perhaps each response would be scored

two or more times by different teachers so that the district could evaluate the reliability of the scoring process.

In large-scale assessment programs, some constructed-responses are machine-scored, which means that sophisticated computer programs analyze students' responses and assign scores. Machine-scoring can be used in lieu of human scoring, as the second scorer when all responses are also scored by a person, or as the only or second scorer for just some of the responses. In all cases, these decisions must be made well before test administration and should be based on research about what will yield the most valid and reliable scores in the circumstances at hand.

Our Standards: High Quality Scoring



- **Standard 7.7:** Test documents should specify user qualifications that are required to administer and score a test, as well as the user qualification needed to interpret the test scores accurately
- **Standard 7.8:** Test documentation should include detailed instructions on how a test is to be administered and scored.
(AERA, APA, NCME, 2014, p. 127)
- **Standard 6.8:** Those responsible for scoring should produce scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented.
- **Standard 6.9:** Those responsible for test scoring should establish and document quality control processes and criteria. Adequate training should be provided. The quality of scoring should be monitored and documented. Any systematic source of scoring errors should be documented and corrected.
(AERA, APA, NCME, 2014, p. 118)

38



These kinds of rules for scoring are critical to ensure that each student's test score reflects her knowledge and skills. Evidence that these rules have been implemented with fidelity is an important part of the technical documentation for a testing program, as well. Our professional standards include several expectations for high quality scoring, including:

Standard 7.7: "Test documents should specify user qualifications that are required to administer and score a test, as well as the user qualification needed to interpret the test scores accurately," and Standard 7.8: "Test documentation should include detailed instructions on how a test is to be administered and scored."

In addition, Standards 6.8 and 6.9 are very clear about other specific obligations related to the scoring process.

Standard 6.8: "Those responsible for scoring should produce scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented."

Standard 6.9: “Those responsible for test scoring should publish and document quality control processes and criteria. Adequate training should be provided. The quality of scoring should be monitored and documented. Any systematic source of scoring errors should be documented and corrected.”

This all means that the entity responsible for scoring should evaluate and provide information to test users about the reliability/precision of the scoring process. Depending upon how the scoring process is designed, that evidence could take different forms. For example, if two or more scorers always or sometimes rate each response, the scoring vendor must provide information about the rate of agreement among those scorers. In all cases, the scoring vendor is obligated to evaluate the scoring process and report the results of this evaluation to the test users.

Examples of evidence for high quality scoring to support score comparisons include:



- Information from the technical manual or reports on item development about:
 - How items are designed and developed to be scored accurately and consistently; and
 - Rubrics, criteria, or other guidance for scoring constructed-response items.
- Evaluation information in the technical manual after every administration about:
 - The scoring protocols and processes as designed;
 - The scoring protocols and processes as implemented, including the qualifications of those scoring constructed-response items and the accuracy of algorithms when items are machine-scored; and
 - Any errors that occurred during scoring and how these were resolved.

39



Examples of evidence for high quality scoring to support score comparisons include:

- Information from the technical manual or reports on item development about:
 - how items are designed and developed to be scored accurately and consistently
 - rubrics, criteria, or other guidance for scoring constructed-response items
- Evaluation information after every administration about:
 - The scoring protocols and processes as designed
 - The scoring protocols and processes as implemented, including the qualifications of those scoring constructed-response items and the accuracy of algorithms when items are machine-scored
 - Any errors that occurred during scoring and how these were resolved

Scaling and Equating to Support Comparability



- 4. How are score scales created and test forms equated to support appropriate comparisons of scores across forms, formats, and time?**



40

edCount
Measuring up to the future

This brings us to our fourth validity question for comparability:

4. How are score scales created and test forms equated to support appropriate comparisons of scores across forms, formats, and time?

Evidence for this question would come from the analysis phase of the testing life cycle.



People who study and use statistical models and methods with assessment data are known as psychometricians and their field is called psychometrics.



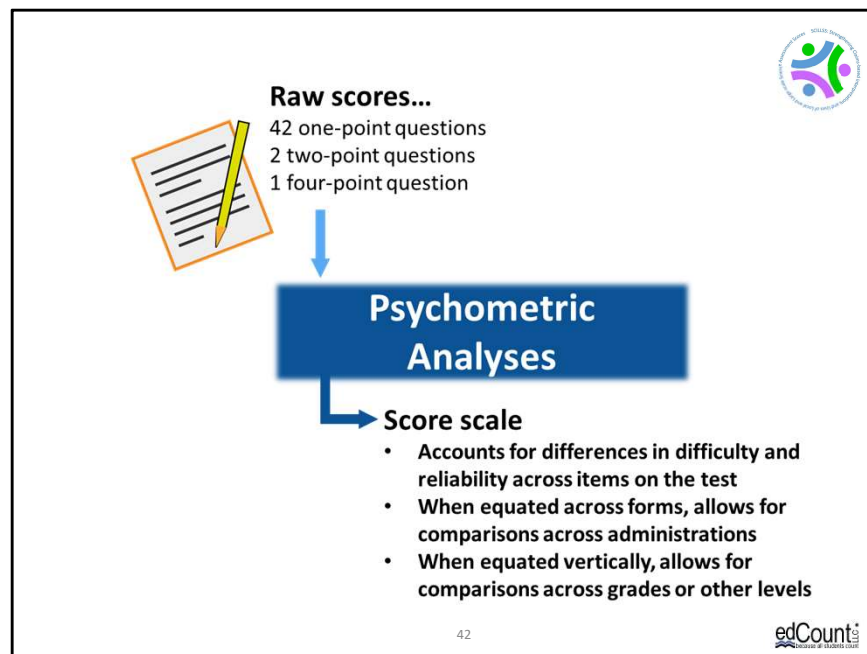
41

edCount
Measuring Up to the Future

Our first two questions in this chapter addressed several issues related to the comparability of scores across forms and formats and across students, sites, and time. In those sections, we pointed to the need for evidence that the forms and formats are measuring the same constructs and for clear guidance for administration as well as evidence that guidance has been well-implemented.

Our fourth question targets the measurement models necessary for generating scores for large-scale assessments. While classroom tests typically yield what are known as “raw” scores, which are simply the number or percent of items students answered correctly, large-scale assessments from outside the classroom almost always report scores using statistically-derived score scales.

People who study and use statistical models and methods with assessment data are known as psychometricians and their field is called psychometrics. There is no need for most educators in state or local education agencies or in schools to be trained in psychometrics, but it is helpful for educators to understand a bit about why score scales are used and why test equating is necessary for score comparability.



Think of a classroom test that includes 42 selected-response questions, two short constructed-response questions worth two points each, and one longer constructed-response question – perhaps requiring about three paragraphs of text describing and explaining a model – that is worth up to four points. That’s 45 questions and 50 raw scores points.

On this test, a student’s score could simply be the number of points he earned via a combination of the selected-response and constructed-response questions and that may be sufficient depending on how these scores are used, particularly if they are combined with other information about student performance.

However, we all know that not all test questions are the same. Some are more difficult, some are more discriminating between students who are well-versed in the skills being measured and those who are less well-prepared. Some items contribute more positively to test reliability than others. For these kinds of reasons, psychometricians who work with large-scale assessments create score scales that address variations in item characteristics and yield scores that can be interpreted in comparably ways across students, sites, and, with psychometric equating techniques, across test forms and formats and time. Without rigorous scaling and equating

methods, it may be completely inappropriate to compare scores across test forms or test administrations.

Our Standards: Scaling and Equating



- **Standard 5.12:** A clear rationale and supporting evidence should be provided for any claim that scale scores earned on alternate forms of a test may be used interchangeably.
- **Standard 5.13:** When claims of form-to-form equivalence are based on equating procedures, detailed technical information should be provided on the method by which equating functions were established and on the accuracy of the equating functions.
- **Standard 5.16:** When test scores are based on model-based psychometric procedures, such as those used in computerized adaptive or multistage testing, documentation should be provided to indicate that the scores have comparable meaning over alternate sets of test forms.

(AERA, APA, NCME, 2014, p. 118)

(AERA, APA, NCME, 2014, p. 106)

43

edCount

Our professional standards include several standards that define expectations for scaling and equating as those methods relate to score comparability. These include:

Standard 5.12: “A clear rationale and supporting evidence should be provided for any claim that scale scores earned on alternate forms of a test may be used interchangeably.”

Standard 5.13: “When claims of form-to-form equivalence are based on equating procedures, detailed technical information should be provided on the method by which equating functions were established and on the accuracy of the equating functions.”

Standard 5.16: “When test scores are based on model-based psychometric procedures, such as those used in computer adaptive or multistage testing, documentation should be provided to indicate that the scores have comparable meaning over alternate sets of test forms.”

Technical experts can assist state and local educators in interpreting technical documentation and making decisions about whether scaling and equating methods

are adequate for particular test score interpretations in general and in relation to comparability of scores.

Examples of evidence for scaling and equating to support score comparisons include:



- Information from the technical manual about:
 - How score scales were developed and evaluated to ensure that the scaled scores are accurate and meaningful; and
 - How score scales are equated across test administrations to support the comparison of scores across forms, sites, and time.

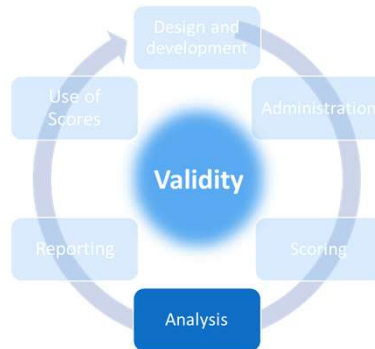
Education agencies that do not have psychometricians on staff may benefit from consultations with psychometricians to help them determine whether and how to compare scores from different tests or test administrations.

Although most state and local educators do not need any type of sophisticated psychometric knowledge to create tests or interpret test scores, they do need to know that the psychometricians who work on the tests they purchase are taking appropriate steps during scaling and equating to support comparable interpretations of test scores. All information about scaling and equating should be presented and summarized succinctly in technical manuals for a test so that test users can have some insights into those methods. To help them understand these methods and whether they are appropriate for a given test and a given interpretation of test scores, state departments of education usually convene and consult with technical experts. Some larger school districts may do the same and we would encourage any education agency that is considering the adoption of a test to seek advice from independent psychometric experts if they do not have such experts on staff.

Comparability of Student Groups



5. To what extent are different groups of students who take a test in different sites or at different times comparable?



45

edCount³
Measuring up to the future

Our fifth validity question related to comparability is:

5. To what extent are different groups of students who take a test in different sites or at different times comparable?

Evidence related to this question would come primarily from the analysis phase of the testing life cycle.

Common comparisons of test scores for groups include:



- **Year-to-year or cohort comparisons**, such as last year's 5th graders to this year's 5th graders; year-to-year comparisons of test scores are often used to help answer questions such as, "is this school or program doing a better job of serving students in science this year than it did last year?"
- **Site comparisons**, such as students in Orange High School to students in Pear High School; these comparisons are often used to answer questions such as, "Which school is doing the best job teaching science?"
- **Student group comparisons**, such as students who are classified as English learners and students who are not classified as English learners; these comparisons are often used to answer questions such as, "How well are schools serving students in their most challenged student subgroups?"
- **Time, growth, or progress comparisons**, such as last year's 4th graders to this year's 5th graders, with the assumption that they are for the most part the same students; these types of comparisons often relate to questions such as, "Are these students progressing in their mathematics knowledge and skills over time?"



46



edCount
FLORIDA DEPARTMENT OF EDUCATION

To this point, we've focused on characteristics of tests that are necessary to support combining or comparing scores across forms, formats, and administrations. This question relates to the people whose scores one wants to compare. We provide a brief overview of group comparability to serve as a caution against overinterpretation of apparent group differences and to highlight the responsibilities of test users, such as state and local education agencies, for appropriate score interpretation and use.

Common comparisons of test scores for groups include:

- Year-to-year or cohort comparisons, such as last year's 5th graders to this year's 5th graders; year-to-year comparisons of test scores are often used to help answer questions such as, "is this school or program doing a better job of serving students in science this year than it did last year?" Some also like to look at scores of students in the same grade across several years if they are interested in achievement trends.
- Site comparisons, such as students in Orange High School to students in Pear High School; these comparisons are often used to answer questions such as, "Which school is doing the best job teaching science?"

- Student group comparisons, such as students who are classified as English learners and students who are not classified as English learners; these comparisons are often used to answer questions such as, “How well are schools serving students in their most challenged student groups?”
- Time, growth, or progress comparisons, such as last year’s 4th graders to this year’s 5th graders, with the assumption that they are for the most part the same students; these types of comparisons often relate to questions such as, “Are these students progressing in their mathematics knowledge and skills over time?”

All of these kinds of comparisons rely on solid evidence for each of the preceding four validity questions; here we consider only the possible differences in the student groups. We strongly recommend that those hoping to use score comparisons to help answer questions seek the advice of an expert evaluator or statistician before doing so.

Test and Student Group Comparability for Different Types of Test Score Comparisons



Comparison	Test Comparability	Student Comparability
Year-to-year/cohort	Equivalent test forms	Equivalent representation of the student population in each year
Sites	Equivalent test forms	Equivalent representation of the student population at each site
Subgroups	Equivalent test forms	All students have equivalent opportunities to demonstrate what they know and can do
Time/progress or growth	Tests measure related knowledge and skills and score scales that are vertically articulated or equated	The same students in each year

47



To ensure that the test data one wishes to use in making comparisons yields comparable evidence, one must consider both the tests and the students. The year-to-year or cohort comparisons, as well as comparisons across sites and student groups, require evidence that the test forms are equivalent across the variations in years, sites, and groups. Comparisons across time that are meant to answer questions about growth or progress require evidence that the earlier tests measure knowledge and skills that relate or contribute to the knowledge and skills that the later tests measure. In addition, these growth or progress comparisons require evidence that the test score scales in the comparison years are statistically connected through an articulation or equating process.

In terms of the students in the groups one wishes to compare, the year-to-year and site comparisons require evidence that these groups reflect equivalent samples of the student population in those years or sites. That is, it would be inappropriate to compare test scores over time or across sites when the students who take the test in one year or at one site include a different subset of students than in the other year or site. If the sample at one site includes all or nearly all students in every group, such as students with disabilities, English learners, students who are eligible for free or reduced-price lunch and those who are not, and students in each racial/ethnic group,

then so must the sample in the comparison site or year if the comparisons are to yield meaningful, useful information.

In cases where one wants to compare scores across these types of student groups, one must first disaggregate the data. Disaggregation means that scores are calculated separately for each group and aggregation means scores are calculated based on all scores. When comparing disaggregated scores, one should establish evidence that the students in the comparison groups have had similar opportunities to learn what the test is measuring, have appropriate and adequate access to demonstrate their knowledge and skills on the test, and do not differ in terms of their motivation to perform well on the test. We will speak more to these issues in chapter 4 of this digital workbook, which focuses on Accessibility and Fairness.

Time comparisons that relate to progress questions and those that relate to growth questions require specific evidence regarding the students tested. Progress questions are about groups of students as in the example of looking at this year's 5th graders compared with last year's 4th graders at the same school. Comparisons at this group level require evidence that the groups are reasonably comparable and, ideally, include mostly the same students. Growth questions require "within-student" data; that is, Sally's 5th grade scores connected to Sally's 4th grade scores and Carlos' 5th grade scores connected to Carlos' 4th grade scores. This is a much more challenging type of calculation because these data requirements can be tricky over time.

Evidence to support score comparisons should include information about:



- Any variations across comparison groups in:
 - Policies about who is tested and included in reporting of results;
 - Students' opportunities to learn the material being tested;
 - The availability and use of testing accommodations; and
 - Students' motivation to take the test.



48

edCount
California's Data Center

In all cases, evidence to support score comparisons should include information about any variations across comparison groups in:

- policies about who is tested and included in reporting of results;
- students' opportunities to learn the material being tested;
- the availability and use of testing accommodations; and
- students' motivation to take the test

Any of these variations could diminish the interpretability of the score comparisons. Those who calculate and report score comparisons are always the ones obligated to provide evidence to support test score interpretations.

Reporting to Support Appropriate Comparisons



6. How are scores reported in ways that appropriately support comparability in score interpretation and use?



49

edCount²
Measuring up to the future

Our next question in this chapter is:

6. How are scores reported in ways that appropriately support comparability in score interpretation and use?

Evidence for this question comes from the analysis and reporting phases of the testing life cycle.

Our Standards: Reporting



- **Standard 6.10:** When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what the scores represent, the precision/reliability of the scores, and how the scores are intended to be used.

(AERA, APA, NCME, 2014, p. 119)

All reported test scores should be:

- As accurate as possible in reflecting the knowledge and skills the test is meant to measure;
- Accompanied by information about the reliability of each reported score;
- Accompanied by information about how the scores are to be interpreted and used and how they should not be interpreted and used; and
- Clear and accessible to those who are meant to interpret and use the scores, including students, parents, and educators.

50



Standard 6.10 of our professional standards states that “when test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what the scores represent, the precision/reliability of the scores, and how the scores are intended to be used.”

As suggested in this standard as well as in several other of our professional standards, test users have four specific types of obligations regarding score reporting. All reported scores should be:

- as accurate as possible in reflecting the knowledge and skills the test is meant to measure;
- accompanied by information about the reliability of each reported score;
- accompanied by information about how the scores are to be interpreted and used and how they should not be interpreted and used; and
- clear and accessible to those who are meant to interpret and use the scores,

including students, parents, and educators.

Our Standards: Reporting to Support Appropriate Score Comparisons



- **Standard 2.4:** When a test score interpretation emphasizes differences between two observed scores of an individual or two averages of a group, reliability/precision data, including standard errors, should be provided for such differences.

(AERA, APA, NCME, 2014, p. 43)

This means that:

- A test vendor/user must report reliability/precision information for each of the scores and for the observed differences between the scores.
- Anyone making claims about the meaning of score comparisons should establish and make available evidence to support such claims.

51



These obligations apply to all forms of scores, whether they are raw scores, scale scores, or scores reported as performance levels. It is never appropriate to simply report a score in the absence of any explanation about what the score means. The information accompanying scores should include explanations about the evidence that supports the meaning of the scores and the error associated with each score.

With regard to the comparison of scores, standard 2.4 states that when a test score interpretation emphasizes differences between two observed scores of an individual or two averages of a group, reliability/precision data, including standard errors, should be provided for such differences. That is, one must report reliability/precision information for each of the scores and for the observed differences between the scores.

In addition, anyone making claims about the meaning of score comparisons should establish and make available all of the evidence we have discussed in this chapter. Simply comparing the scores for two groups without evaluating the comparability of the test forms, formats, testing conditions, and students in the tested groups is irresponsible and can lead to inappropriate and unsupportable statements about differences in the groups.

Appropriate Uses of Score Comparisons



7. What evidence supports the appropriate use of the scores involving comparisons across students, sites, forms, formats, and time?



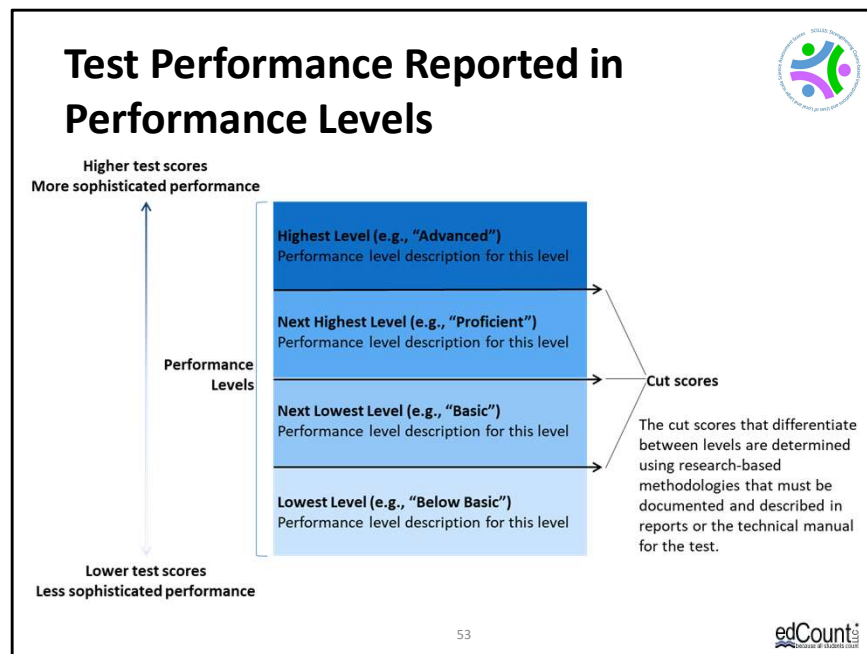
52

edCount³
Measuring up to the future

Our last question in this chapter is:

7. What evidence supports the appropriate use of the scores involving comparisons across students, sites, forms, formats, and time?

Evidence for this question comes from the reporting and use phases of the testing life cycle.



Reports for large scale assessments, such as those a state administers annually, should provide teachers, parents, and students with background information about how the tests were designed and built and how the scores should – and should not – be interpreted and used. While some information about these tests must remain confidential to protect both the meaning of the scores and students’ rights to privacy, the vast majority of information about the test development, administration, scoring, and analysis phases should be available to the public.

To this point, we’ve discussed many types of evidence necessary for testing data to be comparable across several types of variations. Here, we’ll focus on another common type of comparison and then on the evidence necessary when one makes claims about what a student should do based on his or her test performance.

We say test performance here because the results of tests are not always reported as numbers. As we described in chapter 2 of this workbook, performance on some tests may be reported as pass or fail or, as is the case for all large-scale, statewide assessments in the United States, in terms of the performance levels in a set of performance standards. These levels might be numbered or have labels such as proficient or basic or needs improvement or advanced. Even when no particular label

is applied, anytime a score is associated with a decision, such as when a student must achieve a score of at least, say, “20” to be admitted into a program, that’s an implied performance level.

In all cases, the performance levels correspond to ranges of the score scale and the term “cut score” refers to the scores that differentiate one level from the level below it. It’s the score one needs to achieve to pass in pass/fail results. These cut scores should be determined through a rigorous process using research-based methodologies; this process should be described in detail in the technical documentation for any large-scale test.


Reports that present performance in levels must include information to help test users interpret the meaning of students’ performance at each level. This typically includes text associated with each level that describes the kinds of skills that students whose test score falls into that level may have. These performance level descriptors, or PLDs, should also be developed using a rigorous process and, ideally, the technical documentation for a test would include evidence from studies that support claims that student performance on a test reflects the skills described in the PLDs.

Test Performance Reported in Performance Levels



Reports that include a student's performance level should include:

- The student's scale score;
- The name of the performance level where the student's scale score fell;
- A description of what performance in that level generally means;
- Descriptions of what performance in other levels general means; and
- Information about the error around the student's score and around the cut scores that differentiate between the performance levels.



Advanced Performance level description for this level
Proficient Performance level description for this level
Basic Performance level description for this level
Below Basic Performance level description for this level

54

edCount
California's Data Center

Reports that include performance level information for individual students indicate which level the student's scores fell into and should also describe that level as well as the other levels in the performance standards. Reports that present aggregated performance level data do so by indicating the number and percent of students who scored in each level.

As is the case for all other types of scores, performance level scores should be accompanied by information about their reliability/precision. For performance levels, that should include the conditional standard error of measurement at the cut scores, which is specific to the cut scores and different from the standard error of measurement for the test as a whole. In addition, test publishers should provide the results of classification analyses, which indicate the accuracy and reliability of students' classification into the performance levels. All of these statistics should be presented in the technical manuals along with interpretive guidance that helps test users understand the significance of these reliability estimates.

Our Standards: Reporting to Support Appropriate Score Uses



- **Standard 1.5:** When it is clearly stated or implied that a recommended test score interpretation for a given use will result in a specific outcome, the basis for expecting that outcome should be presented, together with relevant evidence.

(AERA, APA, NCME, 2014, p. 24)

This means that if a test report or its accompanying materials indicates that, based on his/her test score, a student should engage in some specific task, activity, or instructional unit, the vendor must provide evidence to support the reasonable expectation that the task, activity, or instructional unit will be beneficial to the student.

55



Performance level scores are sometimes used to assign students, groups of students, or even schools and districts to some form of intervention. This may be, for example, a lesson for a student, a unit or program for a student or group of students, or resources to support improvement efforts in a school. This is a type of comparison because the test scores are used to distinguish one student's skills from another or to categorize schools.

Our professional standards clearly state that those using scores for these purposes are obligated to establish evidence to support such uses.

Standard 1.5: “When it is clearly stated or implied that a recommended test score interpretation for a given use will result in a specific outcome, the basis for expecting that outcome should be presented, together with relevant evidence.”

It's common for tests adopted and used by schools and school districts to offer recommendations for next steps in instruction in relation to students' scores. In these cases, the vendor is obligated to provide evidence to support their claims about score meaning as well as the efficacy of the recommended next steps. That is, if a test report or its accompanying materials indicates that, based on her test score, a

student should engage in some specific task, activity, or instructional unit, the vendor must provide evidence to support the reasonable expectation that the task, activity, or instructional unit will be beneficial to the student.

Our Standards: Reporting to Support Appropriate Score Uses



- **Standard 7.1:** When particular misuses of a test can be reasonably anticipated, cautions against such misuse should be specified.
(AERA, APA, NCME, 2014, p. 125)
- **Standard 9.0:** Test users are responsible for knowing the validity evidence in support of the intended interpretations of scores on tests that they use, from test selection through the use of scores, as well as common positive and negative consequences of test use.
(AERA, APA, NCME, 2014, p. 142)
- **Standard 7.12:** When test scores are used to make predictions about future behavior, the evidence supporting those predictions should be provided to the user.
(AERA, APA, NCME, 2014, p. 129)

56



Our professional standards also obligate test vendors and those who use test scores in making decisions to identify and take steps to avoid the misuse of test scores, which includes inappropriate categorization into levels and subsequent, inappropriate interventions.

Standard 7.1: “When particular misuses of a test can be reasonably anticipated, cautions against such misuse should be specified.”

Standard 9.0: “Test users are responsible for knowing the validity evidence in support of the intended interpretations of scores on tests that they use, from test selection through the use of scores, as well as common positive and negative consequences of test use.”

Similarly, if a testing vendor makes claims that scores can be used to predict later performance, that vendor must provide evidence to support those claims.

Standard 7.12: “When test scores are used to make predictions about future behavior, the evidence supporting those predictions should be provided to the user.”

This obligation applies to tests such as the ACT and SAT, where the scores are interpreted as indicators of likelihood of success in subsequent college settings, and to scores from tests used for selection purposes or to predict performance on subsequent end-of-year tests.

Comparability Questions

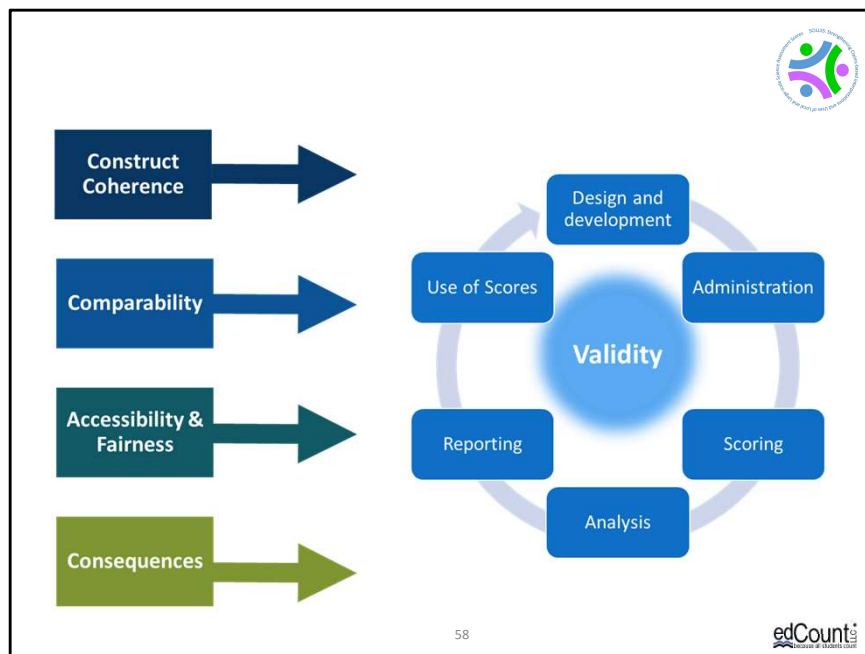


1. How is the assessment designed to support comparability of scores across forms and formats?
2. How is the assessment designed and administered to support comparable score interpretations across students, sites (classrooms, schools, districts, states), and time?
3. How are student responses scored such that scores accurately reflect students' knowledge and skills across variations in test forms, formats, sites, scorers, and time?
4. How are score scales created and test forms equated to support appropriate comparisons of scores across forms, formats, and time?
5. To what extent are different groups of students who take a test in different sites or at different times comparable?
6. How are scores reported in ways that support appropriate interpretations about comparability and disrupt inappropriate comparability interpretations?
7. What evidence supports the appropriate use of the scores in making comparisons across students, sites, forms, formats, and time?

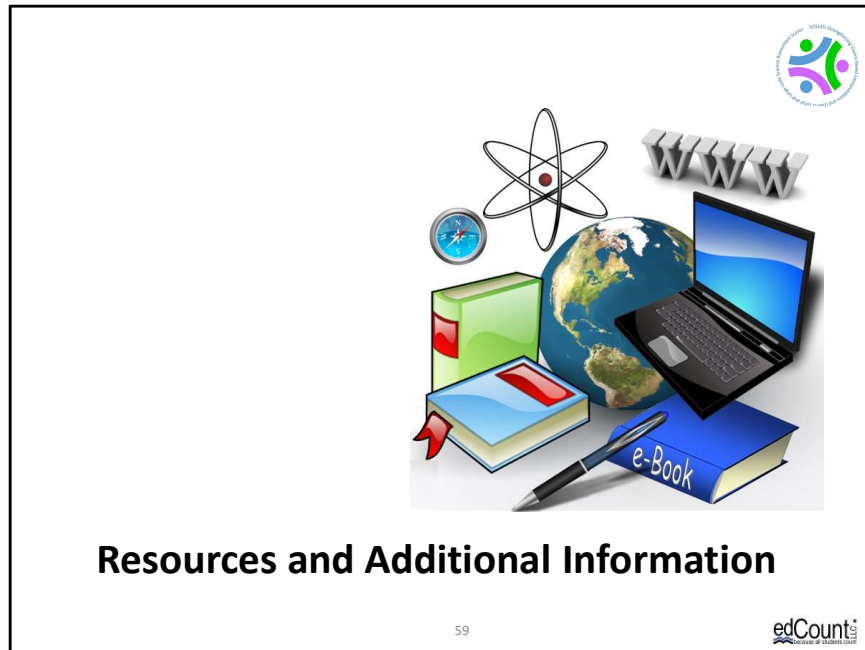
57



We have reached the end of our seven questions in this chapter. As we've seen in each chapter so far, when it comes to reliability/precision and comparability, test developers, or those using test scores for any purpose, are obligated to establish a great deal of evidence to support the ways they interpret and use test scores. Keep in mind that reliability/precision evidence is necessary for all tests and comparability is a concern for every test that is administered to more than one student or on more than one occasion. Any time we want to combine or compare scores we need evidence of comparability.



Thus far in this series, we have addressed questions related to construct coherence and comparability. In the chapters that follow, we will address questions related to accessibility and fairness in chapter 4 and consequences of test use in chapter 5.



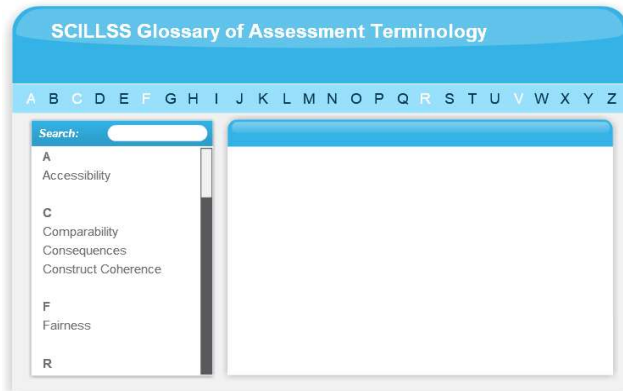
Finally, we offer additional resources that may be helpful to anyone interested in learning more about the concepts presented in this chapter. A glossary of terms and our reference list follow.

Thank you for your engagement in this third chapter of the SCILLSS digital workbook on educational assessment design and evaluation.

SCILLSS Glossary



Please refer to the SCILLSS Glossary for operational definitions of terms used.





Web links

In the web links pod, you can find the following resources.

- American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.
- National Research Council. 2014. *Developing Assessments for the Next Generation Science Standards*. Washington, DC: The National Academies Press.
- SCILLSS Website



References

American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) Joint Committee on Standards for Educational and Psychological Testing. (2014). Standards for educational and psychological testing. Washington DC: American Educational Research Association.