5

Comparability Across Different Assessment Systems

Marianne Perie, *Measurement in Practice, LLC*

CONTENTS

INTRODUCTION

With the production of the Common Core State Standards (CCSS), many state and national policy makers indicated a desire to compare performance across states and jurisdictions. For example, could the performance in Florida be compared to that in Illinois or Texas, and could performance in Miami be compared to that in Chicago or Dallas? The National Assessment of Educational Progress (NAEP) provides comparable scores at the state level and for some urban districts; however, state leaders have indicated the desire for comparability that goes beyond the level that is currently available. On the one hand, they wanted to compare their school or district to top achieving countries in the world, and on the other, they wanted to be able to understand equivalent scores from students coming into their school from another state. In an early meeting,

prior to the development of the multistate testing consortia funded by the federal government under the Race to the Top funding statute, state leaders verbally indicated the following reasons for wanting comparable scores:

- To understand the performance of a student transferring from another state;
- To compare schools across states;
- To know that the term "college ready" has the same meaning from one state to another;
- To compare proficiency across schools, districts, and states; and
- To compare performance in a district to performance in another country.

As policy has moved the focus to college readiness, there is also a desire to compare state assessments to the tests traditionally used for college admissions: the ACT and the SAT. The consortia leaders wanted to replace those tests with their own high school equivalents, and states wanted to be able to predict ACT and SAT scores using their state assessments.

Finally, to achieve all of these outcomes, instruction needed to be aligned with the summative goals. Districts often purchase interim assessments that are purported to align with their state standards to predict success on the summative assessment and improve outcomes.

All of these goals require a degree of comparability among each of these assessments depending on how they are used to meet each goal. Many researchers have struggled with the question of how to compare scores across tests. For example, Braun and Qian (2007); Mislevy (1992); and NRC (1999a, 1999b) have produced important publications documenting the challenges to and potential approaches for comparing scores across states using NAEP, state tests, and other measures. The central challenges that Mislevy outlined in *Linking Educational Assessments* remain:

- "discerning the relationships among the evidence the assessments provide about conjectures of interest and
- determining how to interpret this evidence correctly" (p. 21).

In this chapter, we examine how different assessment systems can be designed to answer questions about comparability of students, schools, districts, and states. Specifically, the focus is on required elements of the assessments for comparability, understanding the comparability of scores at different levels of aggregation, and psychometric constraints in making the desired inferences about students and schools across states and countries.

REQUIREMENTS FOR COMPARABILITY

Returning to the definition of comparability used in this volume, scores are comparable when they can be validly related even when they come from measurements taken at different times, in different places, or using variations in assessments and assessment procedures. Ideally, users could be assured that students with the same score on the same scale possessed the same level of proficiency with respect to the domain of knowledge and skills that a test was intended to measure.

Although the requirements for comparability differ depending on the comparisons made, there are some basic principles that should be followed. Variations in the degree of similarity of these principles affect the degree of comparability that can be assumed: first is the stated purpose of the tests, which affects the test takers' motivations and often differs across tests even within the same system; second, the similarity of the content of the assessments influences what can be said about the scores; third, administration conditions can affect the degree of comparability; and fourth, the psychometric properties of the assessments may change the interpretability of any comparison.

Purpose

Depending on the intended use of the test, a student may spend more or less time preparing for it and may give varying levels of effort during the test-taking session. There is a large body of research documenting the correlation between motivation and achievement (e.g., Finn, 2015; Pintrich, 1989; Pintrich & DeGroot, 1990). Motivation can affect the score because unmotivated students may not give their full effort, resulting in a score that underestimates their actual knowledge and skills (Mislevy, 1992). Even when there are no stakes for the students, such as state assessments under the No Child Left Behind (NCLB) era, the results may affect teachers, which could lead to increased test preparation activities.

Purpose could also affect performance, for example, when the difference is between a benchmark test used to inform the teacher on where to start teaching the student and a summative assessment that "grades" students or their teachers. Putting motivation aside, the interim assessment is likely to be a part of classroom instruction and to have less fanfare associated with it across the school. Because the results can be shared immediately with the student, the test could be perceived differently, and the student may put forth a greater effort than with many state assessments.

Comparing results from a test with no stakes, and a likely proportion of unmotivated test takers, to one with high stakes and highly motivated test takers can result in a large degree of error. For example, predicting a college-readiness score on a test like the ACT from a grade 10 state assessment with no student-level consequences could result in underpredicting performance by several points.

Content

Typically, content similarities are shown through alignment studies that compare the specific content and depth of knowledge assessed by the items on one assessment to that of another (see, e.g., Webb, 1997, 2007). However, traditional alignment studies simply document the percentage of items that can be mapped directly to a standard, the percentage of standards that have items measuring them, the range of depth of knowledge of the items across the standards, and the balance of representation of items across the standard. Two tests can receive similar alignment ratings to the same set of standards and still measure the subject area quite differently. Conversely, two tests could be very similar in what they measure and how they measure it and not receive the same alignment score. Alignment is too superficial as currently defined to be a sole requirement for content comparability.

Although strict content alignment is not required for comparability, the tests should measure the same construct at the same level. Tests that differ in terms of what is assessed or even the distribution of emphasis on the knowledge and skills assessed can affect the comparability of the two scores. When the two consortia met to discuss comparing scores across the two sets of tests, some argued that building the test to the same content standards was a necessary but insufficient condition for comparability. Additionally, the tests would need to use the same blueprints that emphasized the same content areas as well as the same performance-level descriptions (e.g., Marion & Perie, 2011). Even if all that was desired was to claim that the percentage of students who reached proficiency could be compared across all states using one of the consortia assessments, the same definition of proficiency would need to be adopted for each. Using the same method to set the cut scores for proficiency and incorporate the same external benchmarks would also strengthen comparability claims. Ultimately, none of this was done, and scores from the two consortia are not considered interchangeable.

Going further than test specifications or blueprints and performance expectations, item types can also influence the score interpretation. Two tests purporting to measure the same construct but using different item types may be measuring similar content at different levels of rigor. Consider a math test that is all multiple choice versus one that contains additional items asking students to show their work and explain their reasoning. Although both of those constructs could be measured with selected-response items, asking a student to generate a response taps into a different psychological construct that may affect the comparability of the results while providing different insights into the student's learning.

Going even further, if two tests both include open-ended items but are scored with rubrics that emphasize different aspects of the construct, those results could also be limited in comparability. Consider, for example, a rubric that focuses on the quality of evidence used to defend an argument compared to one that focuses on the organization of the paragraph. And, even if the rubrics are similar, differences in scoring processes could also affect comparability, as discussed in the next section.

Finally, even when tests are designed based on the same specifications and using the same performance-level descriptions, differences in alignment of instruction to those standards can influence statistical linking of assessments. For example, Reardon, Kalogrides, and Ho (2018) showed that the linkage between state tests and NAEP showed higher-than-expected scores on the state tests than indicated by linked NAEP scores for districts participating in the Trial Urban District Assessment. One explanation they provide is that the district instruction may be more closely aligned with the state test than that of other districts. When the linkage between the state tests and NAEP is done at the state level, district means predicted from one test to the other have differing degrees of accuracy because curriculum and instruction vary by district.

One alignment report by Forte (2017) discusses the level of alignment needed for comparability, as she describes an approach that starts with the claims made about the resulting scores and traces them through the content standards, blueprints, specifications, and performance-level descriptors. Her theory is that all pieces of an assessment must be aligned to claims made about the student, classroom, school, or state in the final reports.

Administration Conditions

The degree of comparability between two scores also depends on the conditions under which the tests were administered. An issue that states struggled with within each consortium was the degree of flexibility that should be allowed for administration. Take, for instance, the testing window. Some states traditionally set aside one week in the spring for testing, and all assessments are completed that week under strict schedules. Other states have much longer windows, and districts and schools have the flexibility to schedule the assessments within that window. Even with the same length of window, some testing programs assess earlier in the spring than others. Allowing for different assessment windows could have the effect of providing fewer or more instructional hours prior to the assessment, which would negatively affect the comparability of the assessment scores.

A second type of administration condition related to time is the speededness of the test. Although the research is mixed about the direction and significance of this effect, whether the test is timed can affect performance (Haniff, 2012). Therefore, equating error is introduced if we try to link scores taken in a timed test condition to those taken under untimed conditions.

Additionally, the same assessment could be given at the same time through multiple platforms. Some districts may provide the assessment using paper and pencil, others on a personal computer, and others on a tablet; still others may use a combination of platforms. Multiple studies over the past few years have focused on the comparability of paper and pencil to computer-based assessment and on device comparability within computer-based assessment (see, e.g., DePascale, Dadey, & Lyons, 2016; Kingston, 2009; Way, Davis, Keng, & Strain-Seymour, 2016). Differences have been found between paper and pencil and computer assessments as test designers take advantage of technology in ways that can be difficult to translate to paper and pencil. Among technology devices, however, few differences have been found. As long as the students are familiar with the device on which they take the test, the results are assumed to be comparable across devices.

A fourth difference in administration conditions that can affect comparability is the accommodations allowed. Although there is now general consensus on most accommodations, there continue to be different policies on when a read-aloud accommodation is used and when calculators may be used by students with disabilities.[1] There is almost universal agreement that students may have instructions read aloud, regardless of whether they have a disability. Likewise, there is general agreement that reading aloud a math item to those who need that form of communication does not alter the construct being assessed. However, there exist greater differences in opinion about when and how the read-aloud accommodation should be permitted in an English language arts (ELA) assessment (Rogers, Lazarus, & Thurlow, 2014). Some policy makers do not allow it until students have reached a certain grade level so that decoding can be measured in the lower grades. Others argue that reading aloud any part of a reading test at any grade changes the construct assessed and should not be allowed. Still others

[1] For further discussion on accommodations, please refer to Chapter 6, Comparability When Assessing English Learner Students, and Chapter 7, Comparability When Assessing Individuals with Disabilities, in this volume.

believe that we need to assess a student in any way that elicits information from them and thus they allow the read-aloud accommodation at all grades. These differences of opinion manifest themselves in different accommodations policies found across states and other jurisdictions. These differences mean that the scores across jurisdictions with different policies would not be comparable for students needing the accommodation.

Psychometric Characteristics

Other factors of the assessment can also affect the degree of comparability of the scores. Both tests being compared should have similar, high reliabilities in order to make interpretable comparisons. A low reliability on either test would increase the error in linking them and increase the confidence interval around the linked score.

Likewise, the model used to scale the assessments should be the same. If one test uses a one-parameter model to scale the tests and the other uses a three-parameter model, the linkage will have a greater degree of error. When equating two forms of a test, items are calibrated together using the same model. But, often, in trying to project the score from one test onto another, recalibration is not an option. For example, when trying to link a state end-of-course math assessment to the SAT, one of the factors resulting in a high degree of error was that the state test used an item response theory model to create its scale while the College Board uses a classical approach that norms the results (Roeber et al., 2018).

Finally, as discussed in the section on content, different item types can affect comparability. Moreover, within constructed-response item types, different scoring rules can also affect the results, usually by increasing or decreasing the reliability of the overall assessment. Clearly, the rubrics themselves can affect comparability. If different content expectations are emphasized in the rubric, it can reduce the clarity of score interpretation when the two scores are compared. Furthermore, if one program uses a rigorous scorer training protocol with frequent validity checks and read behinds, while the other relies more on remote training with few checks, the resulting score discrepancies can affect the comparability of the results. Likewise, the density of items around specific points in the scale affects reliability, so tests designed to spread items across the scale will have different precision at specific points than will tests designed to maximize information at a specific cut score of the scale.

Ultimately, the comparison between tests could be made at the score level or at some benchmark. For example, under NCLB, states wanted to compare the percentage of students reaching "proficient." Under the revised act, called the Every Student Succeeds Act (ESSA), the focus is more on comparing the percentage of students who are "college ready." And while one could argue that "proficiency" means "ready for college," the same terms may not have the same definitions across different states. And without the same definitions, the percentages are not comparable.

COMMON PRACTICE

Regardless of the principles of comparability, statistically, most tests can be linked. Multiple researchers have linked scores from state assessments to NAEP (Bandiera de Mello, Blankenship, & McLaughlin, 2009; Braun & Qian, 2007), from NAEP to the

Programme for International Student Assessment (PISA) (Stephens & Coleman, 2007) and Trends in International Mathematics and Science Study (TIMSS) (Jia et al., 2014), from one state to another (Bandeira de Mello, Rahman, & Park, 2018), from interim assessments to summative assessments (Reardon, Kalogrides, & Ho, 2018), and from state or local assessments to the ACT and the SAT (Roeber et al., 2018). Test score equating or linking, the more general term, is the most common way we address comparability goals in our current testing context.

The goal of equating is to disentangle differences (across different forms or tests) in item or form difficulty from changes in actual student achievement. A common current example is ensuring that the scores on the state's fifth-grade mathematics test in 2019 can validly be placed on the same scale as the 2018 scores. In this example different students (the fifth graders in 2018 and 2019) have completed tests containing different sets of items, except for a subset of items that were administered in both 2018 and 2019. It is this subset of items—assuming many conditions are met—that allows us to disentangle the changes in student achievement from the changes in the difficulty of the other (nonlinking) items on the test. The challenge is to ensure that the assumptions are actually met.

Although equating is the strongest form of linking, it can only be conducted when the two tests were designed from the same test blueprint to measure the same construct(s). Holland (2007) describes the purpose of equating as making it possible to use scores interchangeably, which can result when the tests measure the same construct with the same intended difficulty and reliability. The most common example is to use two or more forms of the same test. This is not the level of comparability of interest in this chapter, so equating is not discussed here.

The two most common forms of linking when comparing one test score to another different score are calibration and projection. Calibration is used when the tests were not designed from the same test blueprint, but both have been constructed to provide evidence about the same type of achievement (e.g., the same construct). "Unlike equating, which matches tests to one another directly, calibration relates the results of different assessments to a common frame of reference, and thus to one another only indirectly" (Mislevy, 1992, p. 24). Calibration is described as a type of scale aligning with the purpose of "transforming scores from two different tests onto a common scale" (Holland, 2007, p. 12). Projection is used to make statements like "a student who scores X on Test A would have a 75% probability of scoring between Y and Z on Test B." It has a looser set of requirements for the comparability of the two tests, but, as described throughout this chapter, when assessments are constructed around different types of tasks, administered under different conditions, or used for purposes that bear different implications for test takers' affect and motivation, then mechanically applying linking or aligning formulas can prove seriously misleading (Mislevy, 1992).

More recently, the term "concordance" has been used to refer to linking scores on assessments that measure similar (but not identical) constructs and in which scores on any of the linked measures are used to make a particular decision (Kolen, 2004). Subsumed in this definition is the assumption that scores are highly correlated and test takers are similar. For example, ACT and the College Board design studies that result in concordance tables for the ACT and the SAT, which allow one to determine the equivalent score on the test not taken based on the score of the test taken (College

Board & ACT, 2018). Dorans (2004) recommends using regression methods to link scores on measures that cannot be related using concordance procedures. Additional detail on linking can be found in Chapter 2, Comparability of Individual Students' Scores on the "Same Test."

LEVELS OF COMPARABILITY

Thus, the question is no longer how to link one test to another but how to interpret the results and determine if they are truly comparable. Examining the error associated with the linkage will tell us the precision with which we can estimate what score a student would have likely received if they had taken the other test. However, policy makers may be more interested in how scores from one group of students compare to scores from another group. For instance, can we compare Algebra I performance in Los Angeles to that in Chicago when the two districts take two different state end-of-course assessments? Moreover, the comparison might not be made in terms of average scale score but in terms of the percentage of students who "pass" or reach a specific standard.

Figure 5-1 graphically presents a selection of statements that one might want to make about linked scores (Marion & Perie, 2011). As can be seen from the figure, the strictest student-level comparability requires the same test to be administered under the same conditions. The authors of the figure acknowledge that it likely oversimplifies the ordering of the statements and assessment conditions and that the order could change slightly as they attempt to display several factors on a single line. The figure provides a good general overview of the trade-offs between comparability statements and design and administration conditions but should not be viewed as a menu. In the next section, specific examples of different types of linkage are given along with levels of comparability attained.

EXAMPLES

Next, we walk through examples of the ways two different tests are linked to create comparable scores and examine how comparable they truly are. The important piece is the claim being made after scores are linked. The claim that scores are interchangeable requires much more rigorous conditions of comparability than the claim that the rank order of schools would be the same on two assessments or that the percentage of students reaching a benchmark would be comparable.

Interim to Summative Assessments

The creation of consortia also led to the ability of states to purchase balanced assessment systems. Smarter Balanced continues to provide summative assessments, interim assessments, and formative resources. Chapter 4, Comparability Within a Single Assessment System, discusses comparability within such systems. However, only 14 states and 2 territories remain in either consortia at the time of this writing. Most states develop their own summative assessments through a contractual process with a test development contractor. The development or purchase of interim assessments and formative tools are then left to the districts. This practice can lead to a large amount of local

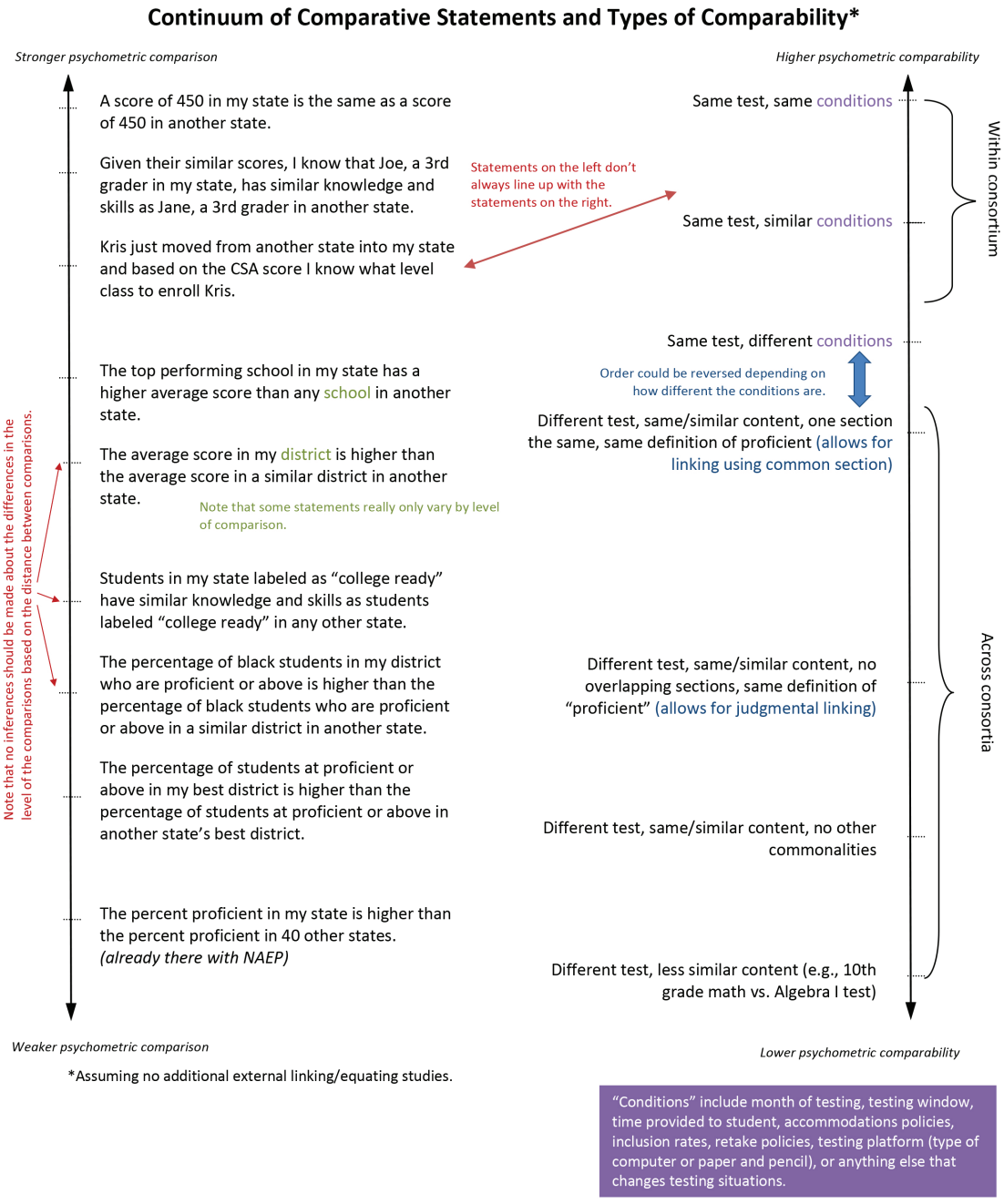

FIGURE 5-1 Continuum of comparative statements and level of comparability.
SOURCE: Marion & Perie (2011).

variation in the comparability of scores between interim and summative assessments. Some states (e.g., Arizona, Florida, and Louisiana) review and approve interim assessments as being sufficiently aligned with the state standards and of sufficient technical rigor to produce results that should be in line with information from the summative assessment. Other states leave to their districts the task of reviewing the technical quality and comparability of the interim assessments.

Many interim assessment companies tout their products as useful in improving scores on summative assessments (Perie, Marion, & Gong, 2009). In order to make claims about how growth on the interim assessment leads to higher summative assessments, they link the two tests. Typically, there are students taking both tests, and a regression equation is set up to predict the score on the summative assessment from the interim assessment. Depending on how close in time the interim assessment is administered to the summative assessment, the prediction can be fairly accurate (see, e.g., Immekus & Atitya, 2016). However, according to Li, Marion, Perie, and Gong (2010), there are several other conditions that should be met to enhance the comparability of the assessments.

Purpose

The interim assessment typically serves a much different purpose than a state summative assessment. An interim assessment is intended to provide instructional feedback to teachers that can lead to corrective action and/or be used to measure growth in understanding over smaller bits of time (Perie et al., 2009). Students are more likely to be engaged because they receive immediate feedback, and the results will affect their learning opportunities. The state summative assessment typically takes months to produce student scores and they may affect school and teacher accountability. However, interim assessments could also be used by policy makers to ensure teachers follow a similar scope and sequence of instruction or to predict performance on the summative assessment, which could yield less desirable results.

Content

Chapter 4, Comparability Within a Single Assessment System, discussed the situation where an interim assessment is built as part of a balanced assessment system. In that case, the standards and item specifications should be the same. However, in many districts, curriculum directors buy an off-the-shelf product that appears to align with their standards and curricular goals.

Many interim assessment companies have spent the past several years aligning their content to the CCSS. During that same time, many states have revised their standards and thus they can no longer be called Common Core State Standards. Depending on the degree of changes made by the state, interim assessments will be more or less aligned to the state standards. Some companies allow districts to select items (or standards) from an item bank to build assessments, a practice that should lead to better alignment. As discussed by Perie et al. (2009), interim assessments can be built as miniature summative assessments, covering all standards, or as a section of the content that reflects a small number of standards that were expected to be taught by that point in the year. The

alignment between the expected scope and sequence of the interim assessment and the actual scope and sequence of the instruction will affect the interpretability of the results.

Finally, differences in item type can affect comparability. Interim assessments that serve the primary purpose of diagnosing a student's understanding may have more open-ended and probing questions. Interim assessments can consist solely of performance tasks, but those are rarely assessed on a large-scale test because of the cost of scoring them. Moreover, to serve a strong predictive function, Perie et al. (2009) argue that the item types on an interim assessment should match the item types on the summative assessment. So, those tests that provide rich instructional feedback are less comparable to the summative assessments and provide less reliable predictive information.

Administration Conditions

If an interim assessment is built or purchased as part of a balanced assessment system, it is likely to have the same administration conditions. Typically, in those cases, the assessments are given on the same computer platform and students receive the same accessibility tools and accommodations every time they access the system. However, when an interim assessment is purchased separately, it is typically given on a different platform and may not include all of the same tools or accommodations. If the summative assessment has more rigorous administration conditions, the interim assessment may overpredict performance on the summative assessment. Conversely, if the interim assessment system does not include all of the necessary accommodations for a student, it may underpredict performance on the summative assessment for that student.

Psychometric Characteristics

Typically, interim assessments have lower reliability than summative assessments because they tend to be shorter. However, the reliability still tends to be well in the acceptable range. Scaling models will often differ because the tests are developed and analyzed by different vendors. However, as long as the standard error associated with the prediction is reported clearly, the results should still be interpretable. As described by Immekus and Atitya (2016), the total score of an interim assessment is typically the best predictor of performance on the summative assessment. Subscores provide little additional value to the prediction equation or comparability.

State High School Assessment to College-Readiness Exam

Currently, several states are replacing their high school assessments with either the ACT or the SAT and using that test for the dual purpose of high school accountability and a measure of preparedness for college coursework. Students can use the scores they receive on the state-administered test for admission into college. Other states are funding one administration of either the ACT or the SAT for every high school student but not using it for school accountability. But, for the states that are allowing the ACT or the SAT to be used as an alternative to their high school assessment, claims they make about the comparability of the two should be examined.

As with interim assessments, some states link their state assessment through a common student approach to make claims about expected performance on the ACT or

the SAT. For example, Kansas provides information on where the equivalent of an ACT benchmark falls on its statewide summative assessments in ELA and math at grades 8 and 10. It also created the linkage shown in Table 5-1 displaying likely ACT scores for student scores at each performance level of the Kansas Assessment Program.

TABLE 5-1 Projected ACT Scores for Grade 10 KAP Performance Levels

English Language Arts		
KAP	ACT Reading	ACT English
Level 1: 220–268	1–17	1–16
Level 2: 269–299	18–23	16–22
Level 3: 300–333	23–29	22–28
Level 4: 334–380	29–36	28–36
Mathematics		
KAP	ACT	
Level 1: 220–274	1–19	
Level 2: 275–299	19–22	
Level 3: 300–332	23–27	
Level 4: 333–380	28–35	

NOTE: KAP = Kansas Assessment Program.
SOURCE: https://ksassessments.org/act.

As with interim assessments, the college-readiness assessments were written to different standards and specifications. However, with a common-population linking design, regression equations can be used to predict performance on one test given performance on the other. More information about the error involved should be provided to help interpret the scores, but the test scores are not intended to be used interchangeably.

A different situation arose in Florida, which had legislation that required the state to analyze the possibility of replacing the Florida State Assessment (FSA) high school ELA test and the Algebra end-of-course assessment with either the ACT or the SAT. The decision would be made at the school level, but then Florida would compare schools based on scores placed back on the FSA scale. In this case, the scores from the FSA, ACT, and SAT would have been considered interchangeable, which requires a high level of comparability. A report commissioned by a group of researchers showed that the criteria needed for this level of comparability could not be met (Roeber et al., 2018). Although the SAT was fairly well aligned to the state content standards, the ACT would need to be supplemented to measure the same detail as measured by the FSA. More importantly, statistical linking showed a large degree of error in trying to predict scores from the FSA to the ACT and the SAT or vice versa. Because the results would be used for school accountability purposes, decision accuracy was calculated using the ACT and the SAT and assuming the FSA decision was "correct." Results of this classification consistency analysis indicate that many students would be placed at different performance levels on the three tests, some by as much as four out of the five performance levels. Particularly concerning was that the direction of the error varied depending on the ability level of the students. Larger schools with a greater number

of lower-performing students have an advantage in using the alternate tests (the ACT and the SAT). Schools with a higher-performing population fared better when graded using the FSA. Florida ultimately withdrew legislation to allow the three assessment programs to operate as if the scores were interchangeable.

The most common approach currently for states is to allow for the "local option" clause in the ESSA. That is, a district can choose to give a "nationally-recognized college-ready assessment" in place of the state high school assessment if it meets peer-review requirements, including content alignment. At least a dozen states are using either the ACT or the SAT as the high school assessment used for accountability (Gewertz, 2019). A couple of states are allowing districts to choose the other assessment as the local alternative, if they desire. For instance, Oklahoma elected to have the SAT become its new state assessment in high school, replacing its end-of-course exams. But it also allows districts to choose to use the ACT instead and rely on the College Board/ACT concordance tables to place all the scores on the same scale. At the time of this writing, no state had both assessments pass peer review. A big issue is the comparability claim and conditions that must be met for it to be true.

Purpose

Both the ACT and the SAT serve the same purpose of informing college admissions offices of a student's level of knowledge, skill, and reasoning ability. Interestingly, though, they both now serve the additional purpose of school accountability at the high school level. A state summative assessment serves the latter purpose, but because there are no other stakes attached, students may be less motivated to do their best on the state assessment. There should be no difference in motivation or effort on the SAT compared to the ACT.

Content

Beginning in March 2016, the College Board administered a new version of the SAT that was revised to better align with the CCSS. The ACT does not align to any particular standards, and its focus is on a framework rather than alignment to standards. Alignment studies have shown differences in the content covered by the two assessments, particularly in mathematics. The SAT includes items on linear equations, systems, problem solving, data analysis, complex equations, geometry, and some trigonometry. The ACT assesses pre-algebra, elementary algebra, intermediate algebra, plane geometry, and coordinate geometry.

The Florida study described earlier (Roeber et al., 2018) showed stronger alignment between the SAT and Florida's state standards than between the ACT and the state standards, indicating a mismatch in content coverage between the SAT and the ACT. Achieve (2018) conducted an alignment study of the ACT to the CCSS and found a weak match between the two. Earlier, the Delaware and Maine State Departments of Education commissioned a study by the Human Resources Research Organization on the alignment of the SAT with the CCSS (Nemeth, Michaels, Wiley, & Chen, 2016). They found fairly strong alignment for ELA and slightly lower alignment for math, although it was concluded that the SAT met the minimum requirements for an aligned high-quality assessment.

These and other alignment studies often conclude with a recommendation that a state augment the college-readiness assessment with a set of items that cover the standards not assessed by the ACT or the SAT. While this approach can close gaps in content alignment, it becomes trickier to ensure these supplemental items are scaled appropriately and lead to valid score interpretations.

Administration Conditions

The ACT and the SAT are administered on different platforms. Although they are now both offered as computer-based tests, different software is used to administer them. The majority of accessibility and accommodation options are the same. The primary difference is in reading directions aloud for English learner students. However, policies are continually reviewed and updated by both companies, so information here may no longer be accurate. The bigger difference is often between state administration policies and the ACT and the SAT policies. Both the ACT and the SAT are timed tests while the majority of state tests are not, and some states allow for different accommodations than others.

Psychometric Characteristics

The comparability report commissioned by the Florida State Department of Education included a table comparing the psychometric properties of an administered form of the ACT, the SAT, and the equivalent FSA. The table is reproduced as Table 5-2 (Roeber et al., 2018, p. 78).

As seen in Table 5-2, there are more differences with the FSA than between the ACT and the SAT. They have similar reliabilities and mean item difficulties. The SAT has greater variability in item difficulty and includes grid-in items in the mathematics test. The ACT is a slightly longer assessment.

Even though the College Board and the ACT work diligently, using a common student approach, to link the scores of the two assessments, the level of comparability is not rigorous enough to assume the scores are as interchangeable as they appear in the concordance tables. The two tests do measure different content, particularly in mathematics, using different item types. And, certainly, they are quite different from typical state summative assessments.

Cross-State Comparisons

A desire of many state commissioners is to compare performance in their state to others. Some describe practical reasons, such as being able to place the student's score from a different state assessment onto their scale to help with placement. Others simply want to raise their relative ranking. The consortia were born of the desire to have scores that can be transported as well as compared across states. Although one consortium, the Partnership for Assessment of Readiness for College and Careers (PARCC), has lost the vast majority of its state participants as of this writing, the other, Smarter Balanced, maintains sufficient numbers of states that comparisons can be made.

TABLE 5-2 A Comparison of Form Characteristics of Florida State Assessment, ACT, and SAT

	ELA			Math		
Criterion	FSA: ELA 10	ACT	SAT	FL EOC: ALG I	ACT	SAT
Form reliability	0.91	0.89	0.89	0.92	0.91	0.90
Form length	53 items + writing prompt	115 items + writing prompt	96 items + writing prompt	58 items	60 items	58 items
Distribution of item types	58% MC; 23% Editing text choice; remaining is multiselect, hot text, and evidence-based Selected Response	MC + essay	MC + essay	Vast majority of item types are MC and SCR (SCR = grid-in or equation editor). Other = table and matching	MC	MC + grid-in
Item difficulty[a]						
Mean	0.65	0.58	0.58	0.21	0.58	0.58
Min	0.12	0.20	0.03	0.00	0.20	0.03
Max	0.92	0.89	0.98	0.75	0.89	0.98

NOTE: FL EOC = Florida end-of-course exams; MC = multiple-choice item, with four options and one correct answer; SCR = short constructed-response item.

[a] Item difficulty is shown as the percentage of students answering an item correctly. The minimum and maximum show the percentage of students answering the hardest and easiest item on a form correctly. The mean gives an indication of the overall difficulty of the form by summarizing the percentage of items answered correctly.

State to State on Consortia Assessments

On the surface, it appears that a consortium of states giving the same assessment should have full comparability of the scores. That was certainly the intent of forming the consortia. Indeed, the consortia tests generally had the same purpose, content, and psychometric characteristics. However, administration conditions were not always the same.

Working within either consortium during development raised many issues of comparability. In order to have interchangeable scores across states, many administration decisions needed to be made and adhered to in every state. It was not sufficient to simply give the same test. It needed to be given at the same time. However, districts choose the starting day for schools, so giving the test to all schools on the same day means there will be a different number of learning days prior to the assessment, and giving the test after a specific number of school days had occurred could lead to security concerns with tests being given on different days across districts. Broadening the assessment window and recommending the number of learning days prior to assessing

helped alleviate that concern. Agreeing on an accommodations policy was also necessary, albeit one of the more difficult discussions among states.

PARCC struggled with comparability of administration platform as more states and districts in this consortium took a paper version of the test. In 2015, nearly 5 million students took the PARCC assessments, 81 percent on computer. Analyses by several states, including Illinois, Maryland, and Ohio, indicated that students did better on the paper version than on the computer version. A subsequent research study (Backes & Cowan, 2018) found mode effects of about 0.10 standard deviations in math and 0.25 standard deviations in ELA, which amounts to up to 5.4 months of learning in math and 11 months of learning in ELA in a single year. Interestingly, this mode effect was cut in half the second year of testing and was almost nonexistent by year 3. Possible reasons for these effects include unfamiliarity with devices, scrolling through reading passages versus flipping back and forth between pages in a booklet, and technology-enhanced items that cannot be fully replicated on paper. All of these factors affect the comparability of results, not just across states but within states, when students take the test on different modes.

State Versus National Assessments

Since 2003, the National Center for Education Statistics (NCES) has released reports that map state proficiency levels onto the NAEP scale. Using an equipercentile linking approach, researchers match the percentage of students reported in the state assessment to be meeting the standard in each NAEP grade and subject to the point on the NAEP achievement scale corresponding to that percentage. They can thereby determine the NAEP equivalent of the state proficiency cut score. Next, a determination is made as to which NAEP performance level best matches the proficient level in each state. In more recent years, the match has been done at the school level. For example, if a state reports that 70 percent of the students in fourth grade in a given school are meeting their math achievement standard and 70 percent of the students in the NAEP achievement distribution in that same school are at or above 229 on the NAEP scale, then the best estimate from that school's results is that the state's standard is equivalent to 229 on the NAEP scale (see Figure 5-2). Results are then aggregated over all schools participating in NAEP in the state to provide an estimate of the NAEP scale equivalent of the state's threshold for its standard. Although not every school is assessed by NAEP, the sampling is done such that generalizations can be made to the entire state and standard errors are reported.[2]

The comparability results are reported and interpreted only at the aggregate state level. Additionally, only the proficient score is mapped, although, theoretically, additional cut points could be mapped. On the surface, such a broad comparison appears valid. However, a deeper dive into the warrants made by such a linking is needed.

A second type of state versus national assessment is conducted at the Stanford Education Data Archive. They use the state accountability data as well as NAEP data to conduct finer-grained analyses of issues like gender gap down to the district level. Their data set currently runs from 2008–2009 through 2014–2015. Some of their findings

[2] Additional details can be found at http://nces.ed.gov/nationsreportcard/studies/statemapping.

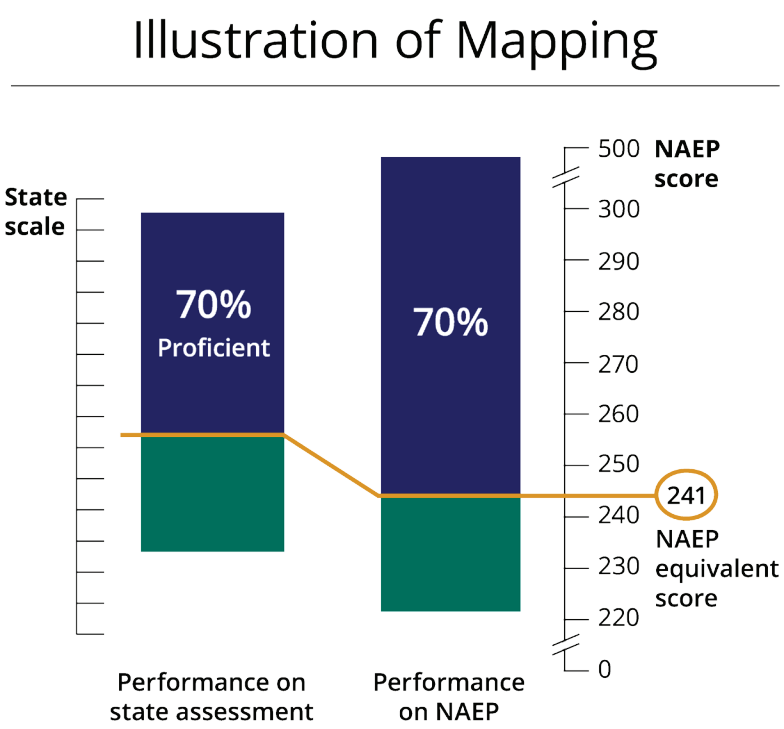

FIGURE 5-2 Illustration of mapping a state cut score onto the NAEP scale.

include that the average school district has no gender achievement gap in math but has a gap of roughly 0.23 standard deviation in ELA that favors girls. Both math and ELA gender achievement gaps vary among school districts but math gaps tend to favor males more in socioeconomically advantaged school districts and in districts with larger gender disparities in adult socioeconomic status. These two variables explain about one-fifth of the variation in the math gaps. However, they found little or no association between the ELA gender gap and either socioeconomic variable and can explain virtually none of the geographic variation in ELA gaps (Reardon, Fahle, Kalogrides, Podolsky, & Zárate, 2018).

Purpose The goals of state assessment systems are very different from those of NAEP. State assessments are specifically designed to be used in school accountability programs while NAEP is intended to be a snapshot of the performance of each state and the nation as a whole. Even though few state assessments are high stakes for students, the students do receive individual report cards that are sent home to their parents. Their teachers also understand that state assessments can affect them either directly, through a connection with performance reviews, or indirectly, through school ratings. This difference could affect the preparation teachers give to students prior to the testing window. Going into the assessments, students know that they will not receive scores on NAEP. Teachers also know that their school will not receive any feedback on student performance. Theoretically, then, students could approach the tests with differing levels of motivation to persevere on the more difficult items.

Content NAEP is not built to any particular set of content standards but rather to a framework determined by subject-matter experts and practitioners working for the National Assessment Governing Board (NAGB). NAEP frameworks provide the blueprint for the content and design of each NAEP assessment. In order to measure trends in student performance, NAEP frameworks are designed to remain stable for as long as possible; however, the frameworks are revisited approximately every 10–15 years to be responsive to changes in national standards and curricula. The current math and reading frameworks were published in 2009.[3]

In 2015, the NAEP Validity Studies Panel released a report of an alignment study conducted between the NAEP frameworks in mathematics and the CCSS at grades 4 and 8. The researchers found "reasonable agreement" overall but also some areas of fourth and eighth grade math where there was less of a match (Daro, Hughes, & Stancavage, 2015).

Specifically, the study found that 79 percent of NAEP items in fourth grade math assessed content included in the CCSS at grade 4 or below. However, the match rate was lower in some areas: 47 percent for data analysis, statistics, and probability; 62 percent for algebra; and 68 percent for geometry. In grade 8 the link was stronger, with 87 percent of NAEP items assessing math included in the CCSS at grade 8 or below. However, the authors noted that 42 percent of the CCSS for grades 6, 7, and 8 were not being tested by any items in the 2015 NAEP item pool. Therefore, there are definite content differences between the states using CCSS in 2015 and NAEP. For those states that were not teaching to the CCSS, the link is unknown but presumably no better.

Administration conditions Administration conditions can vary significantly across the two types of assessments. NAEP reading and math will be administered digitally for the first time in 2019. Many state assessments moved online years ago. As discussed in the previous section, there can be differences between online and paper versions of an assessment.

Students are given 60 minutes to take the NAEP items that have been selected for them. It is, therefore, a much shorter test than many state assessments, and each student only takes one subject. Because only a handful of students are selected to take NAEP in each sampled school, the assessment is given in a small-group setting, which is different from the classroom setting of most state assessments.

In the past, there have been concerns about students being opted out of NAEP because they either had a disability or were in an English learner program. In March 2010, NAGB adopted a new policy to maximize the participation of students with disabilities and English learners. Matching instructions under NCLB, NAGB recommended that exclusion rates should not exceed 5 percent of all sampled students. In 2017, approximately 90 percent of students with disabilities were included in the assessment. The only English learners that should be excluded are those who have been in U.S. schools for less than 1 year and for which a translated form of the assessment is not available.

Accommodation policies differ in some respects between some state assessments and NAEP. NAEP only allows a translated form for students who have been in U.S.

[3] Additional details can be found at https://www.nagb.gov/naep-frameworks/frameworks-overview.html.

schools for less than 1 year. After that, a word-to-word bilingual dictionary is provided. And these accommodations are not permitted on the reading or writing tests. Other accommodations match state accommodations such as extended time, directions read aloud, and test items read aloud for all but the reading test. The reading test also may not be presented to students in American Sign Language. Students are not permitted to have a calculator as an accommodation for the math or science tests. These accommodation policies are similar to those in some states but not all.

Psychometric characteristics Figure 5-3 shows the result of the state mapping done in 2015 for grade 4 mathematics. The claims made are that the proficient cut score set for PARCC in grade 4 math was slightly higher than the NAEP proficient cut score in the same grade and subject. Conversely, states such as Iowa, Louisiana, Nebraska, and Virginia set their proficient cut score at a level that was roughly equivalent to the NAEP basic level.

Although the psychometric characteristics of NAEP differ substantially from the state assessments, the claims made about NAEP seem reasonable. NAEP uses a matrix sampling approach to assessments, meaning different students receive different blocks of items, but all blocks are paired with one another to allow for an estimation of a full covariance matrix. That is, by giving a few students a few items, but by systematically spiraling those items in blocks and randomly sampling the students, inferences can be made about the full population. Because the performance on all items must be imputed based on multiple students taking a few items, background variables of those students are included in the estimation calculations. Multiple conditioned scores are produced and a sample of them is drawn to derive an ability estimate.

These sampled values are described as "plausible values" and only published in sufficiently sized aggregates, typically at the state and large-district levels. In this case, the only claims made are about the rigor of the benchmarks on the two assessments. Similarly, it is becoming common state practice for psychometricians to bring information about the percentage of students scoring at or above proficient on the NAEP grades 4 and 8 reading and math tests to the standard-setting workshops where cut

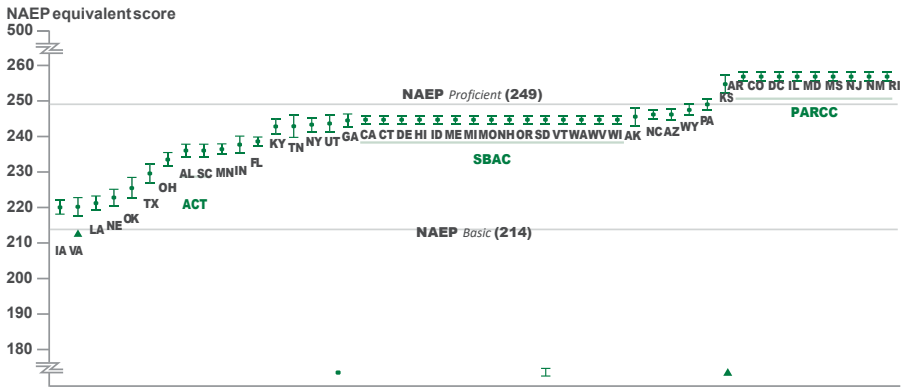

FIGURE 5-3 NAEP scale equivalents of state grade 4 mathematics standards for proficient performance by state, 2015.

scores are placed on state assessments. Comparability is only discussed in terms of the expected level of rigor.

National Assessments Compared to International Benchmarks

NCES ran a special study to link the NAEP scale to the TIMSS scale so that states could compare the performance of their students with that of students in other countries. First, it modified the NAEP assessment schedule so that eighth graders in all 50 states, the District of Columbia, and the U.S. Department of Defense schools could be assessed in mathematics and science in 2011. They were administered NAEP item booklets with some TIMSS items woven throughout. Then, nine states participated in the 2011 administration with a large enough sample to produce state-level results on TIMSS. They took TIMSS booklets that had NAEP items woven in. The NAEP results were used to link the two tests, and the TIMSS results from the nine participating states were used to validate the results. The design of the 2011 study allowed for the use of several different linking methods: statistical moderation, statistical projection, and calibration to predict TIMSS results for the U.S. states that participated in NAEP. All three methods produced similar results, so NCES chose to publish results from the statistical moderation analysis. "Statistical moderation aligns score distributions such that scores on one assessment are adjusted to match certain characteristics of the score distribution on the other assessment. In this study, moderation linking was accomplished by adjusting NAEP scores so that the adjusted score distribution for the public-school students who participated in 2011 NAEP had the same mean and variance as the score distribution for public school students in the TIMSS U.S. national sample. This allowed NAEP results to be reported on the TIMSS scale."[4] The analysis resulted in statements such as "Massachusetts and Vermont scored higher in science than 43 of the 47 participating education systems, while the District of Columbia scored higher than 14 education systems."

Purpose

NAEP, TIMSS, and PISA all purport to have different purposes and certainly serve different audiences. NAEP is a congressionally funded assessment that measures what U.S. students know and can do in various subjects across the nation, across states, and in some urban districts. It has multiple components dating back to 1969. TIMSS has measured trends in mathematics and science achievement at the fourth and eighth grades every 4 years since 1995. The goal is to get a snapshot of performance across multiple countries and to gauge progress over time. Finally, PISA is an international assessment that measures 15-year-old students' reading, mathematics, and science literacy every 3 years, emphasizing functional skills that students have acquired as they near the end of compulsory schooling. The age of 15 was chosen as it is the last age in which education is compulsory for most countries.

[4] Taken from https://nces.ed.gov/nationsreportcard/studies/naep_timss/about_timss.aspx.

Content

All three assessments are written to different test blueprints. NAEP and TIMSS are both curricula based while PISA is skills based, meaning it takes a broader approach to assessing student conceptual understanding. PISA assesses a different age of students than does NAEP or TIMSS, so the content is predictably different.

Administration Conditions

The NAEP-TIMSS linkage was designed with embedded items, meaning the administration conditions were exactly the same. However, that is not true for comparisons made between NAEP and PISA. As the next section describes in detail, NAEP tends to be more inclusive and offer more accommodations than PISA.

Psychometric Characteristics

From the student and teacher perspective, both national and international assessments have limited consequences. Therefore, the likelihood that teachers would engage in test-prep activities for either assessment is low, and students are likely to be equally (un)motivated for each. The linking study between NAEP and TIMSS used sound methodology, and the types of comparisons made appear reasonable. However, digging below the surface, TIMSS is based on specific science and math curricula that may be taught to students in the United States at different times. NAEP and TIMSS test specifications are not the same, so, even though the scores may be transferrable, assumptions about the level of knowledge and skills of a particular jurisdiction may not be.

Different International Assessments Compared

There are three well-known international assessments used to rank-order countries based on student achievement. However, each of the assessments has differences in what and whom they assess. PISA focuses on reading, mathematical, and science literacy at age 15, rotating the emphasis each year; TIMSS assesses mathematics and science in grades 4 and 8; and the Progress in International Reading Literacy Study (PIRLS) focuses on reading at age 10. Few comparisons are made among the tests, with one exception. When PISA and TIMSS are given in the same year (e.g., 2003 and 2015), comparisons are made between the overlapping portions of each assessment. In 2003, comparisons were made regarding math achievement and in 2015 on science achievement.

Table 5-3 shows how each assessment differs on key features. Between PISA and TIMSS, it is important to note that they are given at two different ages (15 and 14, respectively) and are based on two different content frameworks. PISA is intended to be more general and is built around key concepts while TIMSS is curriculum driven, tests more specific knowledge and skills, and may show more differences in scores based on alignment with instruction.

TABLE 5-3 Comparison of Key Features of Three International Assessments

	PISA	TIMSS	PIRLS
Full name	Programme for International Student Assessment	Trends in International Mathematics and Science Study	Progress in International Reading Literacy Study
Assesses	Reading, mathematics, science, problem solving	Mathematics and science	Reading
Age	15	10 and 14	10
Grade	9/10	4 and 8	4
Frequency	Every 3 years, since 2000	Every 4 years, since 1995	Every 5 years, since 2001
Last assessment	2018	2019	2016
When	Autumn	March–June	March–June
Purpose	Evaluates education systems by assessing to what extent students at the end of compulsory education can apply their knowledge to real-life situations and be equipped for society	Measures trends in math and science achievement	Measures trends in reading comprehension
Focus	Skills based	Curriculum based	Curriculum based
Parent organization	Organisation for Economic Co-operation and Development (OECD)	International Association for the Evaluation of Educational Achievement (IEA)	IEA
Countries	72 countries and economies in 2015	57 countries and 7 benchmarking entities in 2015	50 countries and 11 benchmarking entities in 2016
Test length	120 minutes, plus 35-minute background questionnaire	72 minutes at grade 4; 90 minutes at grade 8 plus 15-minute background questionnaire	80 minutes, plus 15-minute background questionnaire
Testing format in most recent year	Computer based	Computer based	Paper and pencil with an ePIRLS extension assessed online
Number of students assessed per country	More than 5,000	At least 4,000	About 3,500–4,000

Purpose

Although the official purposes are listed differently, they appear to be used in similar manners. And student engagement, particularly in the United States, should not vary among the tests. Students are told that they are representing their country but will not receive their score.

Content

There are clear differences in the content among the three assessments. First, they assess different, but overlapping, subject areas. Second, they assess different grades and ages and presumably align the content to be age and/or grade appropriate. Third, the tests developed by IEA are curriculum based while the tests developed by the Organisation for Economic Co-operation and Development are skills based, meaning that TIMSS and PIRLS test more specific knowledge and skills related to curriculum. Opportunity to learn should, therefore, have more of an impact on TIMSS and PIRLS scores than on PISA, which tests more general understandings.

Administration Conditions

PISA is longer than TIMSS or PIRLS and has been assessed via computer. TIMSS was administered through paper booklets until 2019 when it moved to an electronic delivery of the assessment. PIRLS was also administered through paper booklets through 2016, although an optional technology literacy component was administered online. In 2021, PIRLS will transition to a fully online assessment. An additional distinction is that PISA is administered in the fall, while TIMSS and PIRLS are spring assessments.

TIMSS allows read-aloud accommodations only for the directions, as well as a calculator in grade 8 for all students. For students with disabilities who need a read-aloud accommodation, they may request words, phrases, or sentences be read aloud. For students requiring a calculator as an accommodation at grade 4, a school-supplied, four-function calculator is permitted. English learners may use a word-for-word dictionary for translation on TIMSS. Standard setting and presentation accommodations are provided. As of the 2015 assessment, PISA offers only limited accommodations for students with special needs, such as small-group settings, and their exclusion rate is higher. They do not permit extended time or allow large print, Braille, or even magnification. No assistance is provided for English learners. These differences could severely impact the comparability of TIMSS and PISA scores. PIRLS does not offer accommodations for English learners nor does it offer special forms for students with disabilities. Setting accommodations are allowed if typically used for other U.S. assessments.

All three of these policies are much less inclusive than typical U.S. state policies and result in more students being excluded from testing, a decision made at the school level. Indeed, exclusion rates across countries ranged from 0.0 percent in several smaller countries to 8.2 percent in the United Kingdom on the 2015 administration of PISA; the exclusion rate in the United States was 3.3 percent. For TIMSS, in that same year, student exclusion rates varied between 0.0 percent in eight countries to 6.8 percent in the United States. Again, assessing different populations could have a significant impact on comparisons of scores between the two assessments.

Psychometric Characteristics

Typically, comparisons focus on how a country compares to others. One study from Germany found a strong correlation in mean scores by country across TIMSS and PISA (Kleime, 2016). In math, the coefficient of correlation is .923, indicating that 85 percent of the between-country variance in PISA mathematics literacy can be explained by TIMSS, and vice versa. Likewise, in science, the coefficient of correlation is 0.926, accounting for 86 percent of between-country variance. This indicates that the relative rankings could be compared, although the test takers and level of specificity of the content differ. It should be noted, however, that although individual-level correlations are unavailable due to the data-collection design, they would be expected to be much lower than these correlations between national averages.

Examination of claims made about country scores in analyses comparing TIMSS to PISA show that the statements are primarily about the rank ordering. For example, Wu (2009) writes

> It is found that Western countries generally performed better in PISA than in TIMSS, and Eastern European and Asian countries generally performed better in TIMSS than in PISA. Furthermore, differences between the tests on two factors, content balance and years of schooling, can account for 93% of the variation between the differential performance of countries in PISA and TIMSS. Consequently, the rankings of countries in the two studies can be reconciled to a reasonable degree of accuracy.

These claims seem reasonable as they focus on overall trends and not specific comparisons of growth or absolute amount of knowledge demonstrated in the two assessments.

CONCLUSION

There are many necessary conditions for full comparability of test scores. However, those conditions are not necessary to link scores; they are only necessary to validly interpret the results. That is, even though test scores can be linked, it does not mean that the interpretation is the same.

Determining the statement one wants to make about performance on the two tests determines the degree of comparability needed. For example, because the national-international comparisons focus on rank order of countries, the comparability rules are less strict. Conversely, colleges often assume that ACT and SAT scores are interchangeable simply because they can be concorded, even though the content differs, and they currently have different accommodation policies. Assuming interchangeable scores at the student level requires a higher degree of comparability than currently exists between the ACT and SAT.

Even when giving the same assessment on the same platform, states grappled with comparable administration times and conditions in the state consortia, where they wanted transportable scores. Full comparability is difficult to achieve, so it is important to understand the different characteristics and conditions of the assessments and to determine appropriate statements of comparison that can be made.

REFERENCES

Achieve. (2018). *Independent analysis of the alignment of the ACT to the Common Core State Standards*. Washington, DC: Author. Retrieved February, 25, 2019, from https://www.achieve.org/files/ACTReport.pdf.

Backes, B., & Cowan, J. (2018). *Is the pen mightier than the keyboard?: The effect of online testing on measured student achievement.* Washington, DC: American Institutes for Research. Retrieved February 27, 2019, from https://caldercenter.org/sites/default/files/WP%20190.pdf?platform=hootsuite.

Bandeira de Mello, V., Blankenship, C., & McLaughlin, D. (2009). *Mapping state proficiency standards onto NAEP scales: 2005–2007: Research and development report*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Bandeira de Mello, V., Rahman, T., & Park, B. J. (2018). *Mapping state proficiency standards onto NAEP scales: Results from the 2015 NAEP Reading and Mathematics Assessments* (NCES 2018-159). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Retrieved February 10, 2019, from https://nces.ed.gov/nationsreportcard/subject/publications/studies/pdf/2018159.pdf.

Braun, H. I., & Qian, J. (2007). An enhanced method for mapping state standards onto the NAEP scale. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales*. New York: Springer.

College Board & ACT. (2018). Guide to the 2018 ACT®/SAT® Concordance. Retrieved February 12, 2019, from https://collegereadiness.collegeboard.org/pdf/guide-2018-act-sat-concordance.pdf.

Daro, P., Hughes, G., & Stancavage, F. (2015). *Study of the alignment of the 2015 NAEP mathematics items at grades 4 and 8 to the Common Core State Standards (CCSS) for mathematics*. Retrieved February 20, 2019, from https://www.air.org/sites/default/files/downloads/report/Study-of-Alignment-NAEP-Mathematics-Items-common-core-Nov-2015.pdf.

DePascale, C., Dadey, N., & Lyons, S. G. (2016). *Score comparability across computerized assessment delivery devices*. Washington, DC: Council of Chief State School Officers.

Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement, 28*(4), 227–246.

Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series RR-15-19*. Retrieved February 13, 2019, from https://doi.org/10.1002/ets2.12067.

Forte, E. (2017). *Evaluating alignment in large-scale standards-based assessment systems*. Washington, DC: Council of Chief State School Officers.

Gewertz, C. (2019, April 19). Which states were using PARCC or Smarter Balanced in 2016-17? An interactive breakdown of states' 2016-17 testing plans. *Education Week*. Retrieved July 15, 2019, from https://www.edweek.org/ew/section/multimedia/which-states-were-using-parcc-or-smarter.html.

Haniff, R. E. (2012). *The impact of timed versus untimed standardized tests on reading scores of third grade students in Title I schools*. Electronic Theses and Dissertations 2202. Retrieved from http://stars.library.ucf.edu/etd/2202.

Holland, P. W. (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales*. New York: Springer.

Immekus, J., & Atitya, B. (2016). The predictive validity of interim assessment scores based on the full-information bifactor model for the prediction of end-of-grade test performance. *Educational Assessment, 21*(3), 176–195.

Jia, Y., Phillips, G., Wise, L. L., Rahman, T., Xu, X., Wiley, C., & Diaz, T. E. (2014). *2011 NAEP-TIMSS linking study: Technical report on the linking methodologies and their evaluations* (NCES 2014-461). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Kingston, N. M. (2009). Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education, 22*(1), 22–37.

Klieme, E. (2016). *TIMSS 2015 and PISA 2015: How are they related on the country level?* (DIPF Working Paper). Retrieved February 24, 2019, from https://www.dipf.de/de/forschung/publikationen/pdf-publikationen/Klieme_TIMSS2015andPISA2015.pdf.

Kolen, M. (2004). Linking assessments: Concept and history. *Applied Psychological Measurement, 28*(4), 219–226.

Li, Y., Marion, S., Perie, M., & Gong, B. (2010). An approach for evaluating the technical quality of interim assessments. *Peabody Journal of Education, 85*(2), 163–185.

Marion, S., & Perie, M. (2011). Some thoughts about comparability issues with "common" and uncommon assessments. Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved February 2, 2019, from www.nciea.org/sites/default/files/publications/Marion_Perie_Comparability%20paper_NCME_032511.pdf.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: Educational Testing Service.

Nemeth, Y., Michaels, H., Wiley, C., & Chen, J. (2016). *Delaware system of student assessment and Maine comprehensive assessment system: SAT alignment to the Common Core State Standards*. Alexandria, VA: HumRRO. Retrieved February 25, 2019, from https://www.doe.k12.de.us/cms/lib/DE01922744/Centricity/Domain/414/SATalignment.pdf.

NRC (National Research Council). (1999a). *Uncommon measures: Equivalence and linkage among educational tests* (M. J. Feuer, P. W. Holland, B. F. Green, M. W. Bertenthal, & C. Hemphill, Eds.). Washington, DC: National Academy Press.

NRC. (1999b). *Embedding questions: The pursuit of a common measure in uncommon tests* (D. M. Koretz, M. W. Bertenthal, & B. F. Green, Eds.). Washington, DC: National Academy Press.

Perie, M., Marion, S., & Gong, B. (2009). Moving towards a comprehensive assessment system: A framework for considering interim assessment. *Educational Measurement: Issues and Practice, 28*(3), 5–13.

Pintrich, P. R. (1989). The dynamic interplay of student motivation and cognition in the college classroom. In C. Ames & M. Maehr (Eds.), *Advances in motivation and achievement: Vol. 6. Motivation enhancing environments* (pp. 117–160). Greenwich, CT: JAI Press

Pintrich, P. R., & DeGroot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*(1), 33–40.

Reardon, S. F., Fahle, E. M., Kalogrides, D., Podolsky, A., & Zárate, R. C. (2018). *Gender achievement gaps in U.S. school districts.* Palo Alto, CA: Standford Education Data Archive. Retrieved August 20, 2019, from https://cepa.stanford.edu/content/gender-achievement-gaps-us-school-districts.

Reardon, S. F., Kalogrides, D., & Ho, A. (2018). *Linking U.S. school district test score distributions to a common scale* (CEPA Working Paper No. 16-09). Retrieved February 13, 2019, from http://cepa.stanford.edu/wp16-09.

Roeber, E., Olson, J., Topol, B., Webb, N., Christophersen, S., Perie, M., Pace, J., Lazarus, S., & Thurlow, M. (2018). *Feasibility of the use of the ACT and SAT in lieu of Florida Statewide Assessments*. White paper commissioned by the Florida State Department of Education. Retrieved June 25, 2019, from http://www.fldoe.org/core/fileparse.php/5663/urlt/ACTSATFSA.pdf.

Rogers, C. M., Lazarus, S. S., & Thurlow, M. L. (2014). *A summary of the research on the effects of test accommodations, 2011–2012* (Synthesis Report 94). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved June 25, 2019, from https://nceo.umn.edu/docs/OnlinePubs/Synthesis94/Synthesis94.pdf.

Stephens, M., & Coleman, M. (2007). *Comparing PIRLS and PISA with NAEP in reading, mathematics, and science* (Working Paper). Washington, DC: National Center for Education Statistics, U.S. Department of Education. Retrieved February 22, 2019, from http://nces.ed.gov/Surveys/PISA/pdf/comppaper12082004.pdf.

Way, W. D., Davis, L. L., Keng, L., & Strain-Seymour, E. (2016). From standardization to personalization: The comparability of scores based on different testing conditions, modes, and devices. In F. Drasgow (Ed.), *Technology in testing: Improving educational and psychological measurement* (Vol. 2). Abingdon, UK: Routledge.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6). Madison: Wisconsin Center for Education Research, University of Wisconsin.

Webb, N. L. (2007). Issues Related to Judging the Alignment of Curriculum Standards and Assessments. *Applied Measurement in Education, 20*(1), 7–25.

Wu, M. (2019). A comparison of PISA and TIMSS 2003 achievement results in mathematics. *Quarterly Review of Comparative Education, 39*(1), 33–46.