# 3

# Comparability of Aggregated Group Scores on the "Same Test"

Leslie Keng and Scott Marion,
*National Center for the Improvement of Educational Assessment*[1]

**CONTENTS**



[1] We acknowledge the contributions of Susan Lyons to help us conceptualize this chapter and for producing Figure 3-2.

## INTRODUCTION

Chapter 2 outlined the challenges and opportunities associated with comparing scores among individual test takers on tests that are considered the same, such as the same end-of-year state achievement test. That discussion established key principles associated with comparing individual examinee scores. Some might think we must establish individual score comparability before establishing score comparability at various levels of aggregations such as school districts, states, and student groups, but there are many cases for which individual scores are not even generated (e.g., National Assessment of Educational Progress [NAEP], Programme for International Student Assessment [PISA]) and we still care deeply about aggregate comparability. Comparisons of aggregate scores, however, go beyond the typical units of analyses noted above and must include considerations of different test delivery platforms and modes of administration, the types of accommodations available to examinees in different settings, and many other factors. We limit this discussion to factors specific to group scores and do not rehash threats to individual score comparability discussed in the previous chapter.

Comparability is an essential requirement for establishing the validity of inferences of scores across individuals or other units, and validity is always evaluated in the context of specific purposes and uses. Therefore, after a brief introduction pointing out some differences between group and individual comparability is a discussion of the various uses and purposes associated with comparisons of aggregate performance for tests considered to be essentially the same. Following this initial framing, we describe the types of aggregate measures, or derived scores, used to compare group-level performance. We then present an analysis of some of the factors affecting the comparability of aggregate scores, drawing on examples from recent testing situations. This is followed by a discussion of comparability considerations unique to aggregate scores, such as when using matrix-sampling or computer-adaptive test designs. Because comparability exists on a continuum, there is rarely a dichotomous decision to indicate when comparability is either supported or violated. We conclude with a practical framework for evaluating and mitigating threats to the comparability of group scores.

### Student-Level Versus Group-Level Comparability

To illustrate how comparability of scores at the individual level does not guarantee that aggregates of the same scores are comparable, consider the example in Table 3-1 of a fictitious school (school A) across two academic years for the same test. Assuming all conditions for comparability of scores at the student level are met for this test (see Chapter 2), can we reasonably compare school A's performance across the 2 years and conclude that it has made significant improvements in eighth grade reading? Before the school starts celebrating its success in year 2, we should consider the notable drop in the number

**TABLE 3-1** School A's Performance on Grade 8 Reading Test Across Academic Years

| | Year 1 | Year 2 |
|---|---|---|
| Number of students | 120 | 50 |
| Average score | 425 | 512 |
| Proficient | 65% | 80% |

of students completing the test in year 2. This seems to indicate that there is something characteristically different between the composition of school A's test-taking populations in years 1 and 2. Why has there been such a precipitous drop in the number of test takers?

Suppose further that we discover the important information in Table 3-2 about school A's demographic composition for the eighth grade test takers in each year.

**TABLE 3-2** Demographics of School A's Grade 8 Reading Test Takers Across Academic Years

| | Year 1 | Year 2 |
|---|---|---|
| Free and reduced priced lunch | 61% | 30% |
| Special education | 24% | 12% |
| English learners | 12% | 4% |

How comparable are the school's average scores and percentage proficiency across the 2 years in light of this demographic shift? Is the improvement in year 2 due to the new academic initiative, or simply because of decreases in the participation of traditionally lower-performing student groups?

This example illustrates two factors that can affect the comparability of aggregated group scores even if we can assume the comparability of individual scores: group size and group composition. Before launching into a comprehensive discussion of these and other factors, we first discuss the importance of specifying purposes and uses associated with aggregated group scores.

## Purposes and Uses

It is axiomatic to say that validity is contingent upon intended purposes and uses and the claims users want to make based on the test scores. Given the close relationship between comparability and validity, it is fair to extend this axiom to comparability. Aggregated group scores from large-scale assessments are used for many purposes but generally fall into four major categories:

1. Monitoring population trends and patterns;
2. Comparing subgroup performance at specific time points and over time;
3. Evaluation of curriculum, instruction, interventions, and other programs; and
4. Accountability at various levels of the system (e.g., teacher, school, and district).

We expand on these broad purposes below and describe why comparability is essential to each of the categories, but to varying degrees.

### *Monitoring Full Population Trends and Patterns*

The National Research Council's Committee on Developing Assessments of Science Proficiency in K–12 described monitoring as the most important function of large-scale assessments, especially as it pertains to the role of large-scale assessments in systems of assessment (NRC, 2014). Being able to accurately and reliably document academic performance and progress is critical to understanding how educational programs are

working and whether investments in education, implementation of new standards or curricula, or other major policies are contributing to large-scale improvements in educational systems. The information provided from such monitoring assessments often supports useful descriptive purposes.

NAEP, in operation for more than 50 years, is the most well-known monitoring assessment in the United States. Monitoring full U.S. population trends over this time period has been of paramount importance to document the nation's academic progress. While the "state NAEP" has been receiving more attention since its inception, likely because it allows for state-to-state comparisons, the "long-term trend NAEP" is a critical function of the program because it allows policy makers and other education stakeholders to track trends over years and decades on how the nation's schools and students are performing at a point in time and compared to prior performance. This discussion of NAEP highlights an important consideration that might be lost in our discussions of how to evaluate and maintain comparability even when comparability is threatened. Difficult lessons have been learned over the years about maintaining comparability when tests and/or testing conditions have changed. The infamous 1986 NAEP reading anomaly occurred when changes introduced to the test led to unanticipated score drops. When describing the extensive analyses into the score drop, Beaton and Zwick (1990) emphasized, "When measuring change, do not change the measure" (p. 165). In other words, the first strategy for maintaining individual- and group-level comparability should be to avoid changing the assessment, population, conditions, and other factors.

### *Comparing Subgroup Performance at Specific Time Points and Over Time*

In addition to comparing full population trends, evaluating the performance of subgroups of students, particularly educationally disadvantaged subgroups, has been a key component of major equity initiatives in the United States since before the passage of the No Child Left Behind Act of 2001 (NCLB). Full population trends portray a particular picture of educational performance for a particular entity, but such a picture may be misleading if the performance of multiple subgroups differs from the full population results. Therefore, being able to accurately and consistently compare the performance of subgroups of students is critical for sustaining a meaningful equity agenda.

### *Evaluation of Curriculum, Instruction, Interventions, and Other Programs*

A key purpose of many large-scale assessments is to support program evaluation efforts of states, school districts, and other educational entities. School districts and states expend significant resources on a variety of educational materials and programs. Therefore, district and state leaders must exert their fiscal responsibility by evaluating the extent to which such programs are fulfilling the intended aims. Beyond the direct fiscal rationales for pursuing evaluations of programs and interventions, there is an opportunity cost associated with pursuing a less effective compared to a more effective educational program. For example, students cannot be taught using two different mathematics curriculum programs at once; if it turned out they were using the less effective curriculum, any loss of learning would be an opportunity lost. Therefore, states and districts must have the information necessary to evaluate educational programs and interventions. The quality and usefulness of evaluation studies are dependent on many factors, but high-quality data are critical. Test scores often serve as outcome data and

essentially all evaluation designs rest on assumptions of comparability of data across groups (e.g., control and treatment groups) and over time.

### *Accountability at Various Levels of the System*

Finally, school and more recently educator accountability systems designed to meet federal and many state mandates have been designed with the intended purpose of supporting an equity agenda. After all, the original Elementary and Secondary Education Act of 1965 was a key component of President Johnson's "War on Poverty." Being able to identify schools needing support to help students succeed and recognizing schools that can serve as models for others generally requires comparable data across units (e.g., teachers, schools, and districts) to support the intended uses. Furthermore, many accountability systems include goals and targets based on changes in performance over time. Without assurances of comparability at both the individual and aggregate levels, such performance goals and targets are meaningless. Being able to support assumptions of comparability within the accountability system is critical to the credibility of accountability determinations resulting from the system. Generally, assessment results represent an important part of school and educator accountability systems, and the comparability of the assessment results at the aggregate level is often a necessary condition for ensuring the comparability of the full accountability system. Comparability is so important to accountability systems that states have performed some impressive statistical gymnastics to attempt to maintain comparability of the accountability system, but it is always much easier to defend inferences of comparability when there is evidence the assessment results are comparable.

Addressing the four main purposes for aggregate score comparability described above does not mean the results can be used to support causal claims, even though many policy makers would like to do so. Establishing causality requires well-thought-out designs controlling for variables and factors that can influence the results such as context effects (e.g., community characteristics, available resources, and school and district size) and educational variables (e.g., teacher and leader expertise and experience, educator turnover, student turnover, class size, and curriculum choices). Making an inference about educational effectiveness or other quality attributes based on test scores alone—whether measures at a single point in time or growth measures—often ignores these other factors that can operate as intervening variables to affect and/or explain the observed score patterns.

## DERIVED SCORES

The focus of comparability claims for individuals usually involves either student scores (e.g., raw score or scale score) or performance-level classifications on the same test. Group-level comparability considerations often involve *derived scores*, that is, measures that are a summary or aggregation of individual scores or classifications within the group. Derived scores are helpful because they help reduce large quantities of individual scores into a singular value, or statistic, that represents an important quantitative characteristic of the group. This makes comparisons at the group level easier for score users to manage and interpret. In general, derived scores can be categorized in four

general ways: measures of central tendency, measures of variability, criterion-based measures, and growth/value-added scores.

## Measures of Central Tendency

Most individual student reports include measures of central tendency, such as mean or median scale scores for the school, district, and state, to help provide context for the student's performance. Aggregate-level reports usually include mean or median scale scores that facilitate the comparison of student groups or entities across a state. Most school accountability systems use measures such as mean scale scores or mean and median growth scores as the basis of their indicators. Finally, measures of central tendency for groups or entities across time are used to establish trends and compare longitudinal performance.

While it is tempting to simply compare average performance across time, we must attend to things that could influence our interpretations or inferences about a change in the average score. For example, if a school's mean scale score on a test changed from 262 in the previous year to 275 in the current year, we might infer that the school's performance improved. There is certainly a higher score associated with the school now compared to previously. However, what if the population of the school changed substantially due to a shift in attendance boundaries? Or, what if the state made a change in the test, such as the removal of a traditionally more difficult writing task, that led to an unexpected increase in scores? In both (and other) cases, we need to exercise caution when making inferences across time or contexts.

Population or sample size, often referred to as *n*-count or simply "*N*," is an important consideration when computing measures of central tendency because it affects the implicit weight associated with each test score. For example, entities (e.g., schools) could have different multiyear averages depending on if they used weighted or unweighted approaches to compute the multiyear average. Two approaches could yield different 3-year averages, especially if the *n*-counts fluctuate significantly across years. If the *n*-count in year 1 is much smaller than that in the other 2 years, then test scores in year 1 would carry higher implicit weights in the average scale score calculation under an unweighted than under a weighted approach. The same *n*-count/implicit weighting consideration applies when we compute measures of central tendency across groups or entities of differing sizes and, in fact, can lead to issues of Simpson's paradox (see discussion below in the section "Implications: Simpson's Paradox").

## Measures of Variability

Measures of variability indicate the degree of spread or dispersion in a set of test scores. Common measures of variability include the range, interquartile range, variance, and standard deviation. When summarizing and reporting aggregate-level scores for a test, measures of variability are often overlooked or even omitted. This is likely because variability is in general less understood than measures of central tendencies. Even those who know the definitions of measures of variability may not appreciate the utility of these measures when comparing groups. It is important, however, to include measures

of variability in reporting and use them to aid in the interpretation and comparison of test results, at both the individual and aggregate levels.

To illustrate this, suppose a student achieves a score of 85 on a test and the mean test score for the class is 80. Our view of the student's performance would be different if the standard deviation of test scores for the class is 15 compared to if it is 2. In the former case, the student's score is certainly above the mean, but in the latter case, the student's score is likely one of the top scores. Without an understanding of variability and distributions, users would not make this inference.

At the aggregate level, consider the example of a weighted composite score that is used to determine a school's annual summative rating for accountability purposes. The composite score is calculated by applying a weight to each indicator in the accountability system. Consider the following hypothetical equation for computing a composite score that includes four accountability indicators: academic achievement (ACH), academic progress (PROG), English language proficiency (ELP), and chronic absenteeism (CA):

$$\text{Composite Score} = \text{ACH} \times 40\% + \text{PROG} \times 30\% + \text{ELP} \times 20\% + \text{CA} \times 10\%$$

The weights in this equation are referred to as *nominal* or *policy weights* because they are usually set to reflect policy priorities for each indicator in the accountability system. The equation above, for example, would serve to communicate a high-priority emphasis on academic achievement (by weighing it at 40 percent in the composite score), followed by academic progress (with a 30 percent weight). Note that this prioritization plays out in the computation of the composite score for a *given school*. Academic achievement, for example, accounts for 40 percent of the school's composite score and directly influences the summative rating associated with the score. However, if the primary interest is to *compare* schools on their composite scores, then the indicator that is most consequential *may not* be the one that has the highest policy weight. Consider the simple scenario in Table 3-3 of six schools and their composite scores, computed using the formula above, along with the standard deviation of each indicator score across the six schools.

**TABLE 3-3** Composite Scores for Six Hypothetical Schools

| School | ACH (40%) | PROG (30%) | ELP (20%) | CA (10%) | Composite Score |
|---|---|---|---|---|---|
| A | 60 | 50 | 57 | 97 | 60 |
| B | 61 | 51 | 21 | 93 | 53 |
| C | 59 | 50 | 95 | 92 | 67 |
| D | 60 | 51 | 45 | 90 | 57 |
| E | 61 | 49 | 82 | 92 | 65 |
| F | 59 | 49 | 37 | 91 | 55 |
| Standard deviation[a] | 0.9 | 0.9 | 27.9 | 2.4 | 5.6 |

[a] This is the standard deviation of each indicator across the six schools.

NOTE: ACH = academic achievement; CA = chronic absenteeism; ELP = English language proficiency; PROG = academic progress.

In this scenario, the ELP indicator has significantly higher variability, as indicated by the standard deviation values in the final row, than the other indicators. As a result, ELP becomes more consequential in distinguishing schools on their composite scores than the other indicators, even though it has a lower policy weight than ACH and PROG. This illustrates the general idea of *effective weights*, which is directly related to the degree of dispersion in the set of scores (as well as the policy weights) for each indicator or element in a composite score. The concept of effective weights in multivariate indicator systems, such as current school accountability systems, is important to our discussion of comparability. When policy makers establish nominal or policy weights, they believe they are establishing the metrics by which schools or other entities will be compared. However, the effective weights change the means of comparison so it is important for users to understand how the differences between nominal and effective weights can influence aggregate comparisons.

Both of these examples show the importance of considering measures of variability, in conjunction with measures of central tendency and criterion-based measures, to aid in the interpretation and comparison of individual and aggregated group scores.

### Criterion-Based Measures

Criterion-based measures are calculated based on how a group of scores compares to a criterion, such as a benchmark or standard. These measures are often expressed as proportions or percentages and are referred to as *rates* or *percent above cut* (PAC) measures. Common examples include proficiency rates (the proportion of scores that meet the cut score for "proficient" on a test), graduation rates (the proportion of students who have met the requirements for graduation), and chronic absenteeism rates (the proportion of students who meet the definition of "chronically absent"). A collection of related criterion-based measures can be used to facilitate more in-depth comparisons of aggregated group performance. For example, the proportion of students that would be in each performance level for various demographic student groups based on a set of panel-recommended cut scores is typically used as part of a standard-setting workshop to evaluate the reasonableness of the recommendations.

Criterion-based measures are often transformed mathematically into whole-number values and referred to as *indices* or *scores*. For example, proficiency rates are transformed from a percentage (e.g., 72%) to a whole-number score (e.g., 72) so that they can be combined with other related measures. The example in Table 3-3 (in the previous section) includes several accountability indicator values that are rates transformed to the 0-to-100 scale so that they can be combined into a composite score for each school.

Criterion-based measures, such as PAC, have appeal over measures of central tendencies and variability because they are thought to be more easily understood. For example, it appears more intuitively understandable to learn that school A's "pass rate" on the test is 78 percent compared to reading that the school's average scale score is 725 with a standard deviation of 15. In fact, states are required by federal law (currently the Every Student Succeeds Act) to report the percentage of students scoring at the proficient level or higher. It is also often the metric by which trend or gap measures, such as the "achievement gap" between student groups, is quantified. However, Ho (2008)

noted that PAC measures offer very limited and potentially misleading representations of group-to-group or longitudinal comparisons. For example, the relationship between the location of the proficiency cut score and the distribution of test scores can have major effects on apparent changes in trends. Table 3-4 illustrates a seemingly paradoxical case in which a school observes a substantial 15 percent increase in proficiency rate on a test in year 2 and a lesser increase of 5 percent in year 3. However, the change in average scale scores appears to tell a contradicting story.[2] Thus, the substantial jump in proficiency rate for year 2 was due primarily to the movement of students who were just below the cut (742) to above the cut. The improvement of students well below or well above the cut scores is not captured by the proficiency rates but is accounted for in the average scale score and standard deviation.

**TABLE 3-4** Derived Scores for a Hypothetical School Across 3 Years

| Derived Score | Year 1 | Year 2 | Year 3 |
|---|---|---|---|
| Proficient | 45% | 60% | 65% |
| Average scale score | 740 | 745 | 760 |
| Standard deviation | 20 | 18 | 15 |

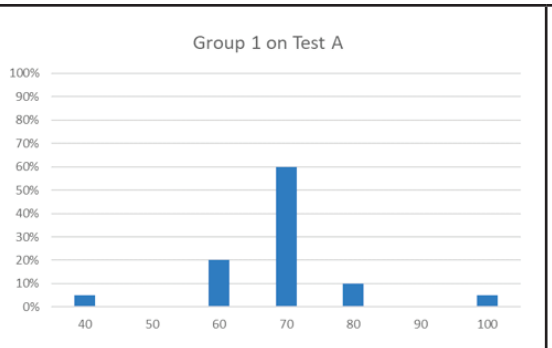
### Growth and/or Value-Added Scores

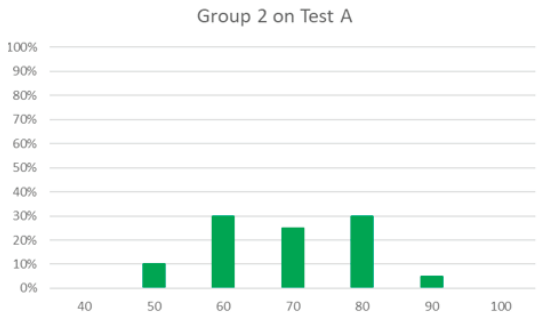
Every grade testing under NCLB changed much in the U.S. testing context, but it also opened the door to documenting students' longitudinal performance. Two prominent approaches have emerged as the main methods for evaluating changes in student test scores over time: value-added modeling (VAM) (NRC, 2010) and student growth percentile (SGP) (Betebenner, 2009). Both VAM and SGP can have substantial effects on individual and group-level comparability, but discussions of VAM and SGP are beyond the scope of this chapter.

### Using Multiple Derived Scores

As we stressed and illustrated with examples in this section, when comparing the performance of groups on a given test, it is important to not limit the comparison to only one type of derived scores. One recommended practice is to take an initial look at the distribution of test scores in the groups of interest, via visual representations such as histograms, before even calculating any of the derived scores. Figure 3-1 shows the merits of visually inspecting the distribution test scores. In this simple example, both groups have the same mean and median scores, the same standard deviation, and the same proficiency rates on the same test (Test A). However, the histograms indicate that there is something characteristically distinct about the performance of the two groups, which could lead to different conclusions or have varying implications in terms of support or interventions for the groups.

[2] For example, if an effect size is computed using the difference in average scale scores and the (pooled) standard deviation, then it would show that the improvement in year 3 (from year 2) is more significant than that in year 2 (from year 1).




| Derived Scores for Group 1 ($N$ = 20) | |
|---|---|
| Mean | 70 |
| Median | 70 |
| Standard Deviation | 11.2 |
| Proficient* | 5% |

* Based on a proficiency cut score of 85

| Derived Scores for Group 2 ($N$ = 20) | |
|---|---|
| Mean | 70 |
| Median | 70 |
| Standard Deviation | 11.2 |
| Proficient* | 5% |

* Based on a proficiency cut score of 85

**FIGURE 3-1** Visual comparisons of two groups of scores.

Even when we consider multiple derived scores and carefully examine the score distribution of two groups, there is still the essential question of whether the performance of two groups is comparable or, more generally, whether it is valid to compare two groups on their aggregated scores. To address this question, we need to take a close look at the contextual factors that can affect the comparability of derived scores.

## FACTORS AFFECTING COMPARABILITY OF AGGREGATE GROUP SCORES

In this section, we describe factors that affect the comparability of derived scores and the inferences that we can validly draw from comparing the aggregated group scores. We organize the factors into four broad categories: variations in group size and composition, variations across assessment conditions, variations in the composition of the assessment, and variations in administration and scoring procedures. Most of these factors also affect the comparability of individual student scores, as discussed in Chapter 2. In this section, we focus on the systematic issues influencing group-level score comparability.

### Variations in Group Size and Composition

The initial example in this chapter illustrated how group size and composition can influence the comparability of aggregated group scores. In that example, we illustrated how these two factors raised questions about the comparability of the overall performance of a single school at two time points on the same test. In most applied scenarios, the ways group size and group composition interact and influence the comparability of aggregate group scores tend to be more complicated. We outline below several group

size and composition factors that could influence the comparability of aggregate scores over time and/or over jurisdictions.

### *Definition of Subgroup Across Jurisdictions*

The focus on equity and closing achievement gaps in educational systems, in addition to comparing the overall performance of groups, requires us to compare the performances of subgroups across jurisdictions, such as schools, districts, and states. In some cases, the size and composition of a subgroup with the same label may differ across jurisdictions because of the geographical factors associated with population distribution. For example, the "English learner" (EL) subgroup in a southwestern U.S. state, such as Arizona, New Mexico, or Texas is generally large (in both absolute size and percentage within the state) and consists mainly of students whose first language is Spanish. The EL subgroup for a northeastern state, such as Maine, New Hampshire, or Rhode Island, tends to be significantly smaller and comprises fewer students whose first language is Spanish, but instead has more students whose first language is Somali. The EL subgroup from Hawaii might be large but includes more students who are Asians or Pacific Islanders with a variety of first languages. The definition of subgroups across jurisdictions may also differ because of policy. For example, rules for entering and exiting EL status, for identifying students with disabilities, and for determining racial and ethnicity groups can vary across districts and states, leading to different sizes and composition of subgroups. Thus, a seemingly newsworthy headline such as "ELs in State A Significantly Outperformed ELs in State B on the SAT Math Test This Year" could be misleading. Instead of drawing conclusions about the effectiveness of academic interventions or support programs for ELs in state A (or lack thereof in state B), it would be prudent to first carefully examine the size and composition of the EL subgroups in each state. For more on the comparability challenges associated with evaluating and maintaining such comparability, see Chapter 6, Comparability When Assessing English Learner Students.

This issue must be considered for other subgroups as well and not just for comparisons across states. For example, special education rates are notoriously variable across states and across districts within states. Even if the proportion of special education students in the population remains steady over several years, "special education" is an amalgamation of 14 specific disabilities and the constellation of the proportion of students with these specific disabilities can vary considerably even if the total proportion of special education students does not change. A shift in the makeup of the special education subgroup, such as a noticeable increase or decrease in students with intellectual disabilities compared with speech or language impairment, can lead to measurable changes in the performance of the special education subgroup. For more on the comparability of assessments concerning students with disabilities, see Chapter 7, Comparability When Assessing Individuals with Disabilities.

A casual reader might think these issues are unique to "educational" subgroups, such as ELs and students with disabilities, and not related to "natural" or "socially defined" subgroups, such as racial, ethnic, or poverty-related subgroups, but that is not true. The challenges faced by economically disadvantaged students have been well documented, but many acknowledge the differences between rural and urban poverty or the differences between those just below the poverty line and those far below.

Similarly, the Hispanic, African American, and Asian/Pacific Islander student groups could all vary considerably in the makeup of each subgroup in ways related to both performance and culture. The main point here is that comparing the performance of both the total population and specific subgroups over time involves understanding how the proportions of student groups have varied over time and how the constellations of the smaller subgroups vary within the larger student groups.

### *Group Size and Sampling Error*

State accountability and assessment leaders have learned a lot about the effects of group size on sampling error. Statistical purists might bristle at the term "sampling error" because many contend that a group of students tested in a particular year is a population. These debates played out early in the NCLB era when states first had to determine the minimum number of students needed to constitute a subgroup (i.e., minimum *n*). States were required by law to make valid and *reliable* determinations, but were also expected to include as many students and subgroups in the accountability determinations as possible. Researchers and state leaders witnessed the notable influence of group size on the variability of the estimates of indicator scores (e.g., percent of students scoring at the proficient level or graduation rates) and had to wrestle with the trade-offs between "reliability" and consequences associated with including as many subgroups as possible in accountability determinations (Kiplinger, 2008; Linn & Haug, 2002). While state leaders recognized the importance of reliable classifications, they quickly learned they would need minimum group sizes so large to meet reasonable reliability thresholds they would exclude many student groups from accountability. Even though many states used confidence intervals around score estimates for smaller groups (e.g., Marion et al., 2002), it became apparent that smaller groups had more volatile score trends than larger groups or schools. Therefore, group or school size is an important consideration for aggregate-level comparability because smaller schools (subgroups) bounce in and out of accountability determinations at higher rates than larger entities (e.g., Linn & Haug, 2002).

### *Changes Over Time Within Jurisdictions*

The size and composition of a group of students within a school, district, or state could change over time. Many schools and districts are in neighborhoods with highly transient populations. Natural disasters can have a significant impact on the constitution of jurisdictions at specific points in time. For example, Hurricane Katrina displaced millions of residents in the state of Louisiana in 2005, affecting the size and composition of school and districts not only in Louisiana, but also in its neighboring states in the Gulf Coast region of the United States. The recent gentrification within large U.S. cities has led to the movement of families with higher socioeconomic status (SES) to traditionally low-SES areas, changing the makeup of schools and districts in both urban and suburban neighborhoods.

Changing definitions or criteria for benchmarks or eligibility rules can also affect the group of test takers. For example, to better align with college and career benchmarks, the WIDA Consortium, which among other things develops assessments for ELP,

adjusted the cut scores of its ACCESS for ELLs 2.0 assessment starting in the 2016–2017 academic year. Many states use performance on ACCESS to determine whether an EL student has attained ELP. The number and composition of ELs who meet the ELP eligibility criteria are likely different in the years before and after the adjustment to the ACCESS cut scores.

Finally, politically motivated initiatives, such as the recent "opt-out" movement in several states where parents elect to excuse their children from taking standardized statewide assessments, can have a substantial effect on the size and composition of the test-taking population depending on the degree of opt-outs across the years in each jurisdiction. The opt-out movement likely had a bigger effect on comparability of statewide achievement test scores and on accountability results in several states with substantial opt-out rates, such as New York, Colorado, and Utah. However, even states with apparently minor opt-out issues can still face comparability challenges because students who opt out generally are a nonrandom portion of the tested population both within and across years.

### *Implications: Simpson's Paradox*

Simpson's paradox is a well-known statistical phenomenon manifest in the social sciences when the underlying population (or sample) is composed of subgroups and comparisons are being made across time or occasions (Blyth, 1972). The issues related to subgroup definitions and compositions described above may play out as a Simpson's paradox. This paradox gained notoriety with Wainer's (1986) explanation of the SAT score increases in the early 1980s.

The average total SAT score increased by 7 points from 1980 to 1984, yet the average score for whites increased by 8 points and 15 points for nonwhites during this time frame. Given the score increases for whites and nonwhites, many wondered why the overall increase was not somewhere between 8 and 15 points. Wainer explained that because the nonwhite scores started so much lower, their score increase of 15 points was not enough to bring them up to the performance of the white scores or even the overall average. Therefore, the weighted average score increase takes into account the size of each group, their starting point, and their score increase. This example demonstrates why it is important to pay attention to the potential of Simpson's paradox when making comparisons over time for an entity comprised of differentially performing subgroups.

## Variations Across Assessment Conditions

Chapter 2 discussed several threats to individual score comparability related to differences in assessment conditions. The threats include factors such as the mode of administration (e.g., paper versus desktop, or laptop versus tablet), test takers' familiarity with item formats, accessibility features and accommodation tools, availability of software and/or hardware for computer-based testing, and the general environment or context in which the assessment is administered. The salient point is that variations across assessment conditions could influence comparisons across groups because these potentially confounding factors are nonrandomly distributed across entities.

For example, a student's performance on a test could be affected by their level of comfort with responding to novel or "innovative" item types, navigating the computer-based testing interface, and/or taking a test on the specific digital device available in the student's school or testing location. In this case, differences in student scores on the test could indicate their experience or level of exposure to computers as much as their math achievement. When these differences are aggregated to the group level, it can manifest as performance differences that are exacerbated by access to technology or the level of computer literacy within a school or district. In other words, differences between groups that are referred to as "achievement gaps" may be due, in part, to gaps in technology access and/or technology literacy.

Several testing programs in recent years have wrestled with issues of mode comparability as school districts and states migrated from paper- to computer-based administrations. These differences were related to several of the issues discussed above (e.g., familiarity) and created challenges for state leaders. On the one hand, they were eager to shift their testing programs online, but on the other, they were reluctant to disadvantage any schools or subgroups that had not yet become used to the new testing system. Depending on the degree of novelty, such as with technology-enhanced items, these effects tended to be observed in the elementary grades' language arts performance—often writing—and the effects would dissipate after a few years. However, for schools that were concerned about accountability results, waiting a few years was not a satisfactory option. Therefore, several states conducted mode comparability studies and proposed making adjustments to the scores associated with the lower-performing mode. While this sounds straightforward, it was not. Rarely are the effects of mode uniform across the score distribution, so a single mean adjustment would not address the problem as fairly as intended. Therefore, many state leaders opted for a policy response by offering a "hold harmless" for schools experiencing score declines consistent with the shift to computer-based testing. In this case, state leaders offered schools the option of using the results based on computer-based testing or maintaining their accountability levels from the last year of paper-based testing, whichever was higher. The policy option was generally offered for a single transition year.

### Variations in the Composition of the Assessment

There are cases in which states or districts may make changes to the design of an assessment so what is touted on the surface as the "same test" may not in fact be the same in composition. Both Common Core–based consortia have examples of such modifications to existing operational tests. One state in the Smarter Balanced Assessment Consortium decided to remove some of the English language arts (ELA) performance tasks from its test forms. New Meridian Corporation, who manages what was formerly known as the Partnership for Assessment of Readiness for College and Careers (PARCC) consortium, made a notable change when it started offering a shorter blueprint in 2019 that several member states chose to administer. In both the Smarter Balanced and New Meridian cases, two group-level score comparability issues should have been considered: the comparability of scores over time within the same jurisdiction and the comparability of scores at the same time across jurisdictions. For example,

the Smarter Balanced state mentioned above should have evaluated whether proficiency rates and aggregated scores, such as average scale scores and growth measures, reported before and after the removal of the performance tasks were comparable within the state and therefore appropriate for use in its accountability system. If the state found larger-than-expected differences compared to normal year-to-year fluctuations, it could have decided to restart its accountability trends or it could have tried to link the two sets of scores using equipercentile linking or some similar method. Smarter Balanced should have considered whether it would be reasonable to compare the state results with other member states after the adjustment. For example, if the state's score change after removing the performance tasks was noticeably different than that of other states, especially in terms of subgroup performance, the consortium could have considered eliminating this state from consortium average performance.

### Variations in Administration and Scoring Procedures

Even if the test design is identical and assessment conditions are controlled for to the extent practicable, different entities or jurisdictions may vary in their approaches to implementation and rigor in enforcing the administration policies and scoring procedures for the "same test." While these are factors that can influence the comparability of individual scores, it is particularly noteworthy for aggregated group scores when different testing vendors or contractors are responsible for administering and scoring the test across time or across states. Many of these factors, such as different test security protocols, scorer qualifications, and psychometric procedures, were discussed in Chapter 2, but we emphasize that such variations may negatively affect the comparability of group scores from the same test within a state or jurisdiction across time, or across states or jurisdictions at a given point in time.

### Differential Item and Test Functioning

The four types of testing variations just discussed can affect state, district, school, and subgroup comparability within and across years. This lack of comparability could play out similarly among subgroups, but often the threats function nonrandomly across subgroups. Differential item functioning (DIF) and differential test functioning (DTF) encompass a substantial set of conceptualizations and analytic techniques used to evaluate these nonrandom outcomes across subgroups and can help shed light on the effects of noncomparability on aggregate-level performance.

DIF is said to occur when two or more sets of examinees, who are otherwise of equal ability (achievement), perform differently on specific items (AERA, APA, & NCME, 2014). In other words, when examinees have the same total test score, there would be no reason to expect systematic performance differences on any item on that test. When such differences occur, typically beyond prespecified thresholds, the item is said to function differentially for particular subgroups of students. Evidence of DIF is not necessarily evidence of test bias (Camilli & Shepard, 1994). Because investigations of item or test bias seek to determine whether scores for subgroups of students may be affected by attributes other than those the test is intended to measure, DIF procedures

may help shed light on the degree with which variations due to assessment compositions, conditions, and administration and scoring processes contribute to differential performance across subgroups that is unrelated to the measurement target. If DIF is detected, content, bias, and assessment experts are convened to try to ascertain whether evidence of item bias exists.

One could imagine a scenario where a large set of items on a test exhibits slight DIF, but none of the items are flagged for meeting prespecified criterion values; however, the direction of the DIF is consistent (i.e., favoring the same group). This could be due to a test that is functioning differently for various subgroups of students. DTF is like DIF, but based on the total test form (AERA et al., 2014). In DIF, however, the total test score is used to contextualize item performance. We need to use a different criterion, obviously, to evaluate DTF, and scores or performance on related measures or other external criteria are used in evaluations of DTF.

Again, observations of DIF or DTF do not mean the test is biased against specific subgroups of students. DIF also may be an indicator of multidimensionality when the test is being treated as a single dimension. DIF and DTF require reasonably sized samples (e.g., $n$ = 200) in order to conduct the analyses. Test makers are not off the hook with smaller samples because they can pursue qualitative approaches, such as cognitive laboratories, to investigate whether items are functioning as intended and similarly for various subgroups.

## PRACTICAL CONSIDERATIONS

Comparability is critically important for monitoring performance trends over time within and across groups; otherwise, educational leaders will not be able to accurately judge if their improvement efforts are working. Additionally, essentially all state accountability systems rely on strong assumptions of comparability to support normative comparisons (e.g., lowest-performing 5 percent of schools) and longitudinal comparisons (e.g., school progress toward long-term and interim goals). These assumptions require evidence documenting the threats to aggregate-level comparability and are not strong enough to invalidate the comparisons. In this section, we provide practical guidelines for practitioners and score users faced with the challenges of needing to make inferences and to act on conclusions drawn from imperfect group-level comparisons of assessment outcomes.

A popular adage in medicine is, "Prevention is better than cure." We suggest that the same idea applies to supporting the comparability of test scores at both the individual and aggregate levels. That is, the optimal approach to supporting claims of comparability is not a series of post hoc analyses, but rather it should begin with the design of the assessment system itself. It means *planning* for factors that may be threats to comparability during the development of the assessment system, *evaluating* the degree to which the threats are mitigated as the system is implemented, and then, if necessary, *adjusting* for any manifested threats or differences. Most importantly, much thought and planning should be put into *communicating* with the field and end users about appropriate score comparisons and interpretations. Figure 3-2 is a visual representation of this framework with a key guiding question for each step.

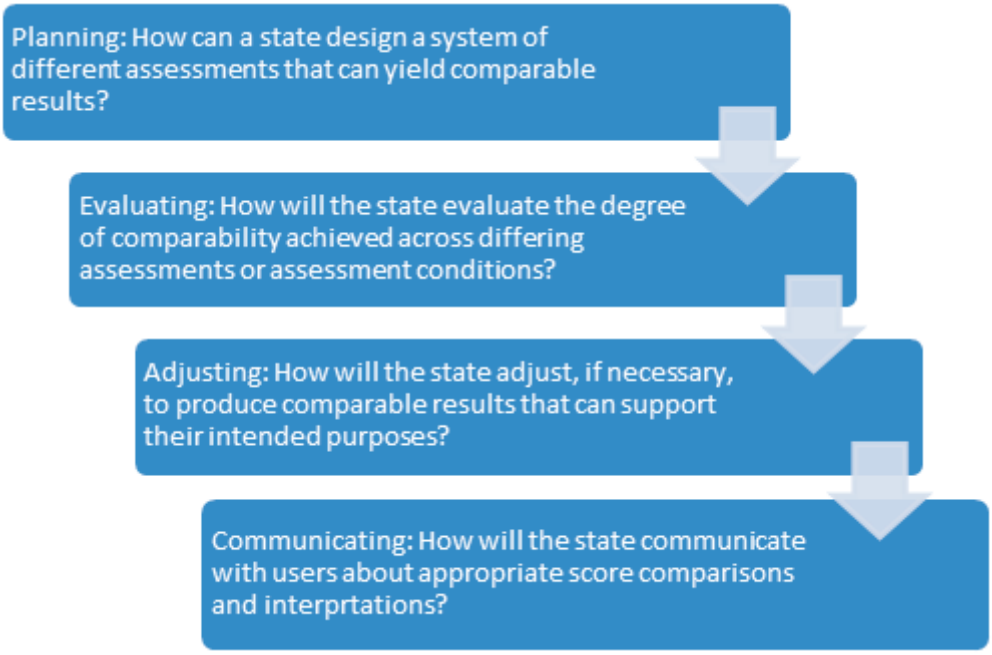



**FIGURE 3-2** Framework for supporting comparability claims in a state assessment system.

The order of these guiding questions is very important. It would not be possible to evaluate the factors influencing group-level comparability if comparability has not been carefully considered in spite of these factors. No amount of evaluation and adjustment can fix a system that has not been carefully designed to produce comparable scores. Thus, garnering evidence to support comparability of the test results requires thoughtful planning of the processes that promote comparability, and program monitoring mechanisms for evaluating comparability. Additionally, states must have a clear plan for effectively communicating with the field about the degree to which scores can be meaningfully compared among groups and entities, over time and across assessment conditions.

States should also consider the *people* who can provide support in each step of the framework. When identifying these supporting parties, the state should think not only about assessment and accountability professionals within the state education agency, but also those in local districts and schools; staff from its testing vendor or subcontractors; established practitioners and experts in the field, such as those on the state's various advisory committees; and educational stakeholders from the community, such as policy makers, teachers, and parents.

### Applying the Framework to Aggregated Score Comparability

To illustrate the use of this framework in the context of supporting the comparability of group-level scores, we provide key questions and considerations that a state can consider at each step to help mitigate or minimize the threats to comparability described earlier in the chapter. We recall the threats we focused on above:

- Variations in group size and composition,
- Variations across assessment conditions,
- Variations in the composition of the assessment, and
- Variations in administration and scoring procedures.

Table 3-5 below summarizes the key questions and/or considerations for each of the threats to group-level comparability for each of the steps in the framework.

**TABLE 3-5** Considerations for State Leaders to Help Mitigate or Minimize the Threats to Comparability

| Comparability Threat | Variations in Group Size and Composition |
|---|---|
| **Planning** | • What is the range of group sizes that the state observes within a given year? Across multiple years?<br>• How similar or different are the students that make up the groups in the state within and across years in terms of key demographic and educational characteristics? |
| **Evaluating** | • Is there a minimum group size at which derived scores are no longer reliable?<br>• What is the degree of uncertainty (e.g., standard error) of the aggregate scores for different group sizes?<br>• To what extent are the student characteristics that vary across groups correlated to the group's performance? |
| **Adjusting** | • For each purpose and use that the state is comparing groups, is it reasonable to combine or collapse certain groups (e.g., form a "super-subgroup") to increase group sizes or make the groups more similar in size?<br>• Are there statistical adjustments that that state can make to account for the larger degree of uncertainty associated with small group sizes? |
| **Communicating** | • If no adjustments are made, what explanations or disclaimers should the state include with the results of group comparisons to address the influence of uncertainty or precision resulting from differences in group sizes and/or composition?<br>• If adjustments are made to account for the variation in group sizes and/or composition, what information should the state include with the results of group comparisons to explain the adjustment procedures and rationale as well as the uncertainty associated with such adjustments? |

**TABLE 3-5** Continued

| Comparability Threat | Variations Across Assessment Conditions |
|---|---|
| **Planning** | • What protocols, instructions, and support can the state implement to minimize the impact of context effects due to variation in assessment conditions *across the general test-taking population*?<br>• What protocols, instructions, and support can the state implement to minimize the differential impact that assessment conditions can have on *specific subgroups*? |
| **Evaluating** | • Are there any group-level performance trends that are correlated with specific assessment conditions?<br>• Is there evidence of subgroups that are differentially affected by certain assessment conditions?<br>• Does the impact of any assessment conditions on group-level performance change (i.e., either weaken or grow stronger) over time? |
| **Adjusting** | • Should the state apply statistical adjustments to account for any of the following:<br>o An overall (main) effect for an assessment condition (e.g., a "motivation" or "opportunity to learn" adjustment for all students)?<br>o A differential (interaction) effect for an assessment condition and a subgroup of students (e.g., a "mode" adjustment for students who take the test online)?<br>o A change in the effect of an assessment condition over time (e.g., a "familiarity" effect applied to group-level scores in subsequent years of an assessment program)? |
| **Communicating** | • If no adjustments are made, what explanations or disclaimers should the state include with the results of group comparisons to address the potential impact of variations in assessment conditions?<br>• If adjustments are made to account for the variations in assessment conditions, what information should the state include with the results of group comparisons to explain the adjustment procedures and rationale? |

*continued*

**TABLE 3-5** Continued

| Comparability Threat | Variations in the Composition of the Assessment |
|---|---|
| **Planning** | • If changes in the composition of the assessment have been mandated, how can the state approach the changes to the test blueprints, design, content specifications, and/or performance-level descriptors, etc., to minimize the impact on comparability?<br>• Can the state propose alternatives to changing the assessment composition or request longer timelines to implement the change? |
| **Evaluating** | • What impact does the change in assessment composition have on the underlying scale, performance standards (i.e., cut scores), reliability, and validity of the assessment?<br>• Do the changes in assessment composition differentially affect certain groups of students in the state? |
| **Adjusting** | • Are the changes in assessment composition so substantial that the state cannot maintain the existing scale or cut scores?<br>  ○ If so, what processes does the state need to implement to generate a new scale and cut scores?<br>  ○ If not, what adjustments, if any, should be made to the existing scale or cut scores? |
| **Communicating** | • What information should the state provide to the field about the changes to the assessment composition and any potential implications to group-level performance comparisons?<br>• If a new reporting scale and cut scores are introduced, or the existing scale and cut scores are modified, what guidelines can the state provide to help the field interpret the assessment outcomes before and after the change? What cautions or disclaimers should the state provide in terms of interpreting group-level trends over time? |

**TABLE 3-5** Continued

| Comparability Threat | Variations in Administration and Scoring Procedures |
|---|---|
| **Planning** | • What training, documentation, and real-time support can the state provide to local testing personnel to ensure that the administration procedures are implemented with fidelity?<br>• What test security protocols and procedures does the state need to enforce to minimize testing irregularities or improprieties during administration?<br>• What qualification criteria, scoring protocols, and monitoring procedures should the state put in place to support reliable scoring processes, including both machine and human scoring?<br>• How can the state minimize the impact of transitioning to innovative scoring approaches on score comparability? |
| **Evaluating** | • What evidence does the state need to collect to confirm that administration procedures have been implemented with fidelity?<br>• What data forensics analyses should the state conduct to detect potential testing irregularities or improprieties?<br>• What metrics should the state calculate and monitor regularly to confirm that the scoring processes are reliable and implemented with fidelity?<br>• Does the state have evidence of differential scorer effects on responses from different subgroups?<br>• What research studies does the state need to conduct to support the validity of innovative scoring approaches? |
| **Adjusting** | • If there is evidence that administration or scoring procedures have not been implemented with fidelity, what adjustments, if any, does the state need to make to affected student scores? Does the state need to apply any adjustments to group-level scores?<br>• If there is evidence of testing irregularities or improprieties, how should the state handle the student scores in question? Should the state apply any adjustments to group-level scores?<br>• If there is evidence of differential scorer effects on specific subgroups, what adjustments should be made to student scores in the impacted groups? Should the state apply any adjustments to aggregated scores for the impacted groups? |
| **Communicating** | • If there are issues related to test administration, scoring, or incidents of testing irregularities or improprieties, how can the state communicate the issues, potential impacts, and mitigation strategies to the field in a clear and transparent manner? |

### An Example

Consider a scenario in which a state is legislatively required to remove writing from its ELA assessments. To minimize the potential impact on longitudinal trends, the state would like to maintain comparability of the ELA scale score and performance levels. How should the state approach this change? Table 3-6 outlines potential approaches the state may employ to evaluate and perhaps maintain comparability. Note that, based on our categorization of threats to group-level score comparability, the removal of writing from ELA is a variation in the composition of the assessment, and perhaps a significant variation.

**TABLE 3-6** Example Application of Comparability Support Framework

| Framework Step | Potential Courses of Action |
|---|---|
| **Planning** | • The state conducts comparative analyses of the old and new test blueprints, design, content specifications, and achievement-level descriptors to determine whether the underlying ELA construct is substantively affected by the removal of writing prompts.<br>• The state convenes meetings with ELA content specialists and educators from across the state to provide input and feedback on the proposed changes to the test blueprints, design, content specifications, and performance-level descriptors.<br>• The state examines its school accountability system and identifies aggregate measures, indicators, classifications, and/or identification business rules that are potentially affected by the removal of writing from the ELA assessments.<br>*Supporting Parties*<br>• ELA content specialists and educators from the state education agency, testing vendor, and representatives from across the state<br>• Accountability specialists at the state education agency |

**TABLE 3-6** Continued

| Framework Step | Potential Courses of Action |
|---|---|
| **Evaluating** | • The state conducts empirical studies, based on data from the most recent operational administration, to evaluate the impact of removing writing tasks on item calibration, scaling, test reliability, predictive validity, and classification accuracy and consistency for the ELA assessments. The studies are conducted at each grade level for all students and by subgroups.<br>• The state replicates the empirical studies during the upcoming operational administrations.<br>• The state continues to monitor for unexpected shifts in ELA performance, especially for the female student group (which traditionally scores higher on writing) and schools that previously showed notable improvement in writing.<br>• The state performs impact analyses with its accountability system to evaluate whether there are any unexpected changes in school ratings or identifications because of the removal of writing. If the analyses reveal such changes, the state examines the affected schools to see if there are any discernable trends in terms of the characteristics of the schools. If the state judges the trends to be substantial, the state may choose to reset accountability goals and establish a new baseline.<br>*Supporting Parties*<br>• Psychometric and research experts<br>• Technical advisory committee (TAC)<br>• Accountability implementation specialists and programmers |
| **Adjusting** | • Based on the findings from the empirical analyses, the state makes adjustments to the underlying scale or establishes a new scale for its item bank.<br>• The state convenes an ELA standards validation meeting to recommend potential adjustments to the cut scores on the ELA assessments. Depending on the committee's recommendations, the state adjusts its reporting scale.<br>• If unexpected shifts in ELA performance are detected in subsequent years, the state considers additional adjustments to the scale and cut scores.<br>• Based on changes made to the assessment system and the impact analysis on the accountability system, the state makes decisions such as whether to adjust the affected accountability components (i.e., aggregate measures, indicators, ratings, etc.), to introduce new accountability components, and/or to suspend reporting of accountability outcomes during the transition year.<br>*Supporting Parties*<br>• Psychometric and research experts and TAC<br>• ELA content specialists and educators from across the state (to participate in the standards validation meeting)<br>• Accountability implementation specialists and programmers<br>• Accountability leadership and advisory committee |

*continued*

**TABLE 3-6** Continued

| Framework Step | Potential Courses of Action |
|---|---|
| **Communicating** | • The state convenes focus and/or advisory groups to review and provide input on the updated individual student reports (ISRs).<br>• The state makes changes to the ISRs based on feedback from the focus and/or advisory groups.<br>• The state organizes community outreach meetings to explain the changes to the assessments, especially in terms of whether schools and districts can maintain trends or if new baselines must be established. In addition to clearly communicating the decisions, a key communications goal is getting buy-in from key stakeholder groups.<br>• The state publishes communication resources that highlight findings from the empirical studies, outcomes from the standards validation process, changes to the reporting scales, and impacts to the accountability system components and outcomes.<br>• The state updates its annual assessment and accountability technical manuals with details about the empirical studies, changes to the scale and cut scores, and changes to the accountability system components.<br>*Supporting Parties*<br>• Assessment and accountability reporting specialists<br>• District and school administrators<br>• Community leaders and educational stakeholders<br>• Psychometric and research experts and TACs |

## CONCLUSION

The focus of this chapter has been on the comparability of aggregated group scores. In our experience, states often attend to the comparability of scores at the individual student level because they are perceived as having a more direct impact on students. Less attention is often afforded to score comparability at the aggregate level. We have attempted to highlight the importance of considering the comparability of group-level scores by describing the purposes and uses of comparing scores at the aggregate level, citing common threats to group-level score comparability, and proposing a framework that states can use to evaluate and build its case for comparability. Chapter 2 and this chapter should provide assessment and accountability professionals with broad knowledge and practical guidance on how to establish the validity of test scores and their inferences through comparability at all levels of reporting.

## REFERENCES

AERA (American Educational Research Association), APA (American Psychological Association), & NCME (National Council on Measurement in Education). (2014). *Standards for educational and psychological measurement*. Washington, DC: AERA.

Beaton, A. E., & Zwick, R. (1990). *The effect of changes in the National Assessment: Disentangling the NAEP 1985-86 reading anomaly* (Rep. No. ETS-17-TR-21). Washington, DC: National Center for Education Statistics. Retrieved August 1, 2018, from https://files.eric.ed.gov/fulltext/ED322216.pdf.

Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*(4), 42–51.

Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association, 67*(338), 364–366. http://doi.org/10.2307/2284382.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items* (Vol. 4). Thousand Oaks, CA: Sage Publications.

Ho, A. D. (2008). The Problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher, 37*(6), 351–360.

Kiplinger, V. L. (2008). Reliability of large-scale assessment and accountability systems. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test based accountability* (pp. 93–114). New York: Routledge.

Linn, R. L., & Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis, 24*(1), 29–36.

Marion, S. F., White, C., Carlson, D., Erpenbach, W. J., Rabinowitz, S., & Sheinker, J. (2002). Making valid and reliable decisions in the determination of adequate yearly progress. In *Implementing the state accountability system requirements under the No Child Left Behind Act of 2001*. Washington, DC: Council of Chief State Schools Officers.

NRC (National Research Council). (2010). *Getting value out of value-added: Report of a workshop.* Washington, DC: The National Academies Press.

NRC. (2014). *Developing assessments for the Next Generation Science Standards.* Washington, DC: The National Academies Press.

Wainer, H. (1986). Minority contributions to the SAT score turnaround: An example of Simpson's paradox. *Journal of Educational Statistics*, *11*(4), 239–244.