1

Introduction: Framing the Issues

Amy I. Berman, *National Academy of Education*
Edward H. Haertel, *Stanford University*
James W. Pellegrino, *University of Illinois at Chicago*

"How is my child doing?" "What are my child's strongest and weakest subjects?" "Have my child's test scores improved from last year?" "How does my child's test scores compare to others looking to go to college?" "Should I move to this school zone?"

—Parent questions

"How do the assessment scores of schools within our district compare?" "How are our English learner students doing compared with our native English speakers?" "Are we closing the achievement gap?" "How do our assessment scores compare to others within the state?"

—District administrator questions

"How do our kids measure up to kids in other states?" "Within districts?" "How are the scores of various student subgroups changing over time?"

—State administrator/policy maker questions

While such questions are common, finding accurate and satisfactory answers is not an easy task given the countless factors influencing assessment scores. Stakeholders often simply assume that scores obtained from different students in different times and places, using different test forms, are directly comparable. Moreover, the questions come from a range of stakeholders each with a separate vested interest in educational assessments, ranging from parents worried about individual student test scores, to local district leaders interested in a specific population, to state policy makers looking at the broad aggregate data. They are often asking different questions about the same

assessments, but answers to these questions do not always coincide with the interpretive uses for which the assessments were originally designed and validated. While their interests and questions may differ, these stakeholders all have one thing in common: they are asking questions that assume scores can be validly compared—that a lower score means less proficiency, similar scores mean similar proficiency, a higher score means greater proficiency, and a positive change in scores from one year to the next means improvement, regardless of the specific details of how each student was tested. In other words, they *assume* the *comparability* of scores from educational assessments.[1]

And while much of educational news reporting relies on testing data, it is often reported at a high aggregate level without descriptions of the assessments, their purposes, and possible explanatory variables. Recent headlines such as these have the potential to influence people, even if they do not tell the full story: "Minnesota Report Card: Small Schools Score Higher" (Sethrie, 2020); "Maryland's PARCC Results Show Dip in Math, Improvements in English" (Ryan, 2019); "Oregon Dips in Standardized Test Scores, Mixed Bag for Mid-Valley" (Rimel, 2019); "New Statewide Test Results Show Achievement Gap Throughout Cedar Rapids Community School District" (Kalk, 2019); "Survey: 45% of Test-Takers Boycott ELA Exam [Long Island, NY]" (Tyrrell, 2019); and "Majority of South Bend Schools Do Not Meet Federal Expectations, New Report States" (Kirkman, 2020). Such articles can influence individual decisions concerning where to live or whether to apply to a nontraditional public school as well as state and federal policy makers' decisions about investments and policies related to educational reform.

This National Academy of Education (NAEd) volume provides guidance to key stakeholders on how to accurately report and interpret comparability assertions as well as how to ensure greater comparability by paying close attention to key aspects of assessment design, content, and procedures. The goal of the volume is to provide guidance to relevant state-level educational assessment and accountability decision makers, leaders, and coordinators; consortia members; technical advisors; vendors; and the educational measurement community regarding *how much* and *what types* of variation in assessment content and procedures can be allowed, while still maintaining comparability across jurisdictions and student populations. At the same time, the larger takeaways from this volume will hopefully provide guidance to policy makers using assessment data to enact legislation and regulations and to district- and school-level leadership to determine resource allocations, and also to provide greater contextual understanding for those in the media using test scores to make comparability determinations.

To accomplish these ambitious goals, the NAEd organized a steering committee comprised of Edward Haertel (Co-Chair), James Pellegrino (Co-Chair), Louis Gomez, Larry Hedges, Joan Herman, Diana Pullin, Marshall S. Smith, and Guadalupe Valdes. The topical foci of the eight chapters following this introduction are the result of the committee's extensive efforts to determine the most pressing comparability issues currently affecting educational assessment while also ensuring that particular subgroups for which comparability issues often arise are included in the discussion instead of shelved with an asterisk for later discussion. The committee organized these issues into

[1] The words *assessment* and *test* are used throughout this volume, and though to some extent they are interchangeable, they do have different meanings. *Assessment* is the more general of the words, conveying the idea of a process providing evidence of quality. *Assessment* covers a broad range of procedures to measure teaching and learning. A *test* is one product that measures a particular set of objectives or behavior.

the following chapters: (1) comparability of individual students' scores on the "same test," (2) comparability of aggregated group scores on the "same test," (3) comparability within a single assessment system, (4) comparability across different assessment systems, (5) comparability when assessing English learner (EL) students, (6) comparability when assessing students with disabilities, (7) comparability in multilingual and multicultural assessment contexts, and (8) interpreting test-score comparisons. The first four chapters progress from narrower to broader interpretive contexts, with comparability claims in each chapter building on those preceding. Chapters 6 through 8 address specific populations meriting additional attention. The final chapter offers a synthesis of best practices for interpreting test-score comparisons. After identifying the chapter themes, the steering committee outlined the chapter goals and identified experts to develop and author the individual chapters. The steering committee, as well as other chapter authors, provided critical feedback on draft chapters, including at a 2-day workshop of authors and the steering committee in June 2019.[2] The results of these efforts comprise this volume.

BACKGROUND TO THIS VOLUME

Student testing has played an important role in the American education system since its creation. Each day students take tests, most of which are devised by teachers, to monitor student learning and guide instruction. Testing students for the purposes of classroom feedback, system monitoring, and selection and placement decisions have existed for more than 180 years. Standardized written exams began in the mid-19th century (OTA, 1992).

The mid-19th to the mid-20th century served as a time of great expansion for educational testing. Entire books and articles have been written about the history of educational testing (see, e.g., Kaestle, 1983, 2012; OTA, 1992; Resnick, 1982; Vinovskis, 2019). While we cannot do justice to such a history in so short an introduction, we point out that, with both population growth and urbanization, public school enrollment more than doubled from 1870 to 1900 and with it the desire to use educational testing for accountability and classification purposes (OTA, 1992). By 1900, intelligence testing had begun and, following the extensive use of intelligence tests in the Army during World War I, these tests proliferated into American schools (Kaestle, 2012). During the 1920s and 1930s, cost-effective, multiple-choice standardized tests became entrenched in schools (OTA, 1992). And, in 1950, the automatic scoring machine was invented by the Iowa Testing Program and large-scale state and national testing became feasible (OTA, 1992).

Of course we would be remiss in not acknowledging the equity concerns that have abounded in standardized testing. Issues have been raised about the equity (bias) of tests, as well as disparate educational outcomes that result from the use of results from educational tests. Moreover, there is increased diversity of test takers with our ever-changing population as well as expansion of test taking, including greater racial and ethnic diversity, language and cultural diversity, and the inclusion of students with disabilities. While we again cannot do justice to this history in this introduction, others

[2] The steering committee also called on the expertise of Christian Faltis to serve as both a discussant at the June 2019 workshop and a reviewer of several chapters. The committee is grateful for his contributions to this volume.

have described these issues (e.g., Moss, Pullin, Gee, Haertel, & Young, 2008; Symposium, 1994) and some of these concerns are raised throughout this volume, including in our chapters addressing English learner students, students with disabilities, and nondominant language and cultural groups.

The Elementary and Secondary Education Act (ESEA) of 1965 included test-based evaluation measures—albeit weak and weakly enforced (Kaestle, 2012; Vinovskis, 2019)—as part of an effort to raise educational achievement and make education more equitable. Then, in 1969, the first national assessments of academic achievement, now known as the National Assessment of Educational Progress (NAEP), were administered.

In 1983, President Ronald Reagan's National Commission on Excellence in Education released *A Nation at Risk*, which asserted that "the educational foundations of our society are presently being eroded by a rising tide of mediocrity that threatens our very future as a Nation and a people," and "[i]f an unfriendly foreign power had attempted to impose on America the mediocre educational performance that exists today, we might well have viewed it as an act of war." This rallying cry led to reform efforts to set high standards and increase accountability measures, often described in the form of testing. In 1991, President Bush proposed the "America 2000" program (and implemented portions through executive order), which called for challenging national standards and voluntary national tests (Vinovskis, 2019). And in 1994, President Clinton's Goals 2000 Act and the Improving America's Schools Act (the latter being the reauthorization of the ESEA) both passed, calling for high educational standards and systems of testing accountability (NAEd, 2009). Finally, in 2002, the federal government mandated annual educational testing in grades 3 through 8 and once in high school for accountability purposes with the passage of the No Child Left Behind Act (NCLB). While NCLB set the impossible goal of all students reaching proficiency on state reading and math tests by 2014, the states' response to NCLB also highlighted the lack of comparability of state standards and assessments.

In 2009, nearly all states, along with the District of Columbia, came together to develop common academic standards in mathematics and English: the Common Core State Standards (CCSS) Initiative. Common standards led to the call for common assessments and, in 2009, through the Race to the Top (RTT) program, the Obama administration announced a competition for grant funding of $350 million for the development of tests aligned with the CCSS (Jochim & McGuinn, 2016). In 2010, the U.S. Department of Education awarded grants to two state consortia, the Partnership for Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (Smarter Balanced), which represented 44 states and the District of Columbia, to develop assessment tools aligned with the common standards adopted by states (DOEd, 2010, n.d.; Robelen, 2010).[3] As noted by the U.S. Department of Education in its award letters to PARCC and Smarter Balanced, for public schools to succeed we need "a first-rate assessment system to measure progress, guide instruction, and prepare students for college and careers."[4] Moreover, through RTT, the Obama

[3] There were also several smaller awards to consortia addressing assessments for students with severe disabilities and for English learner students.

[4] The U.S. Department of Education award letters can be found here: https://www2.ed.gov/programs/racetothetop-assessment/parcc-award-letter.pdf and https://www2.ed.gov/programs/racetothetop-assessment/sbac-award-letter.pdf.

administration offered states a competitive grant program to enact preferred education reform policies, which included adoption of high-quality common standards (which could be demonstrated by participating in a consortium of states) and new assessments aligned to those standards (DOEd, 2009).

Common standards and common assessments, among other things, would address the variation in the stringency of state standards. And, common standards aligned with common assessments were expected to greatly enhance the interpretability of achievement results within and across states. The hope was that if states adopted the CCSS and signed on to one of the state assessment consortia (Smarter Balanced or PARCC), then policy makers, schools, and parents could finally gauge how their students were performing relative to their peers in others parts of the country. The same goals were behind the decisions of the consortia of states that developed common assessments for use with students with severe disabilities that would be assessed against alternative achievement standards derived from the Common Core standards (e.g., Dynamic Learning Maps and the National Centers State Collaborative).

Perfect comparability in testing, however, is not achievable (NRC, 1999a). From the inception of the Common Core assessments in 2010, questions arose about whether results within each consortium, let alone across consortia, were comparable when students were taking the tests via paper and pencil or computer, were using different electronic devices, were tested on different dates stretching over a multiweek administration window, or were subject to different accommodation policies (Hess, 2014). Moreover, well before the advent of Common Core assessments, NAEP was facing issues of comparability because states use different procedures for inclusion, accommodations, and so forth (NRC, 1999b). But such issues were not generally considered reason enough to abandon the idea of having common assessments that could provide comparable results across states.

The U.S. Department of Education incentivized states to adopt the Common Core standards and assessments but soon there was backlash: some felt that federal involvement in education had gone too far. In an effort to take back some local control over assessment, states started backing away from Common Core assessments. Relative to its predecessor version of the law, the Every Student Succeeds Act of 2015 allows states more flexibility in designing their accountability systems. Some of the possibilities include using the Common Core standards and assessments (Smarter Balanced, PARCC, or ACT Aspire), having their own customized state standards and tests, using one of the nationally recognized college entrance exams (ACT or SAT) as their high school assessment, or even giving districts within a state a menu of assessments to choose from.[5]

Over time, fewer states have been administering Common Core consortia assessments in their entirety as intended, and more states are moving toward creating their own unique assessment systems that include a blend of shared and customized elements (Marion, 2017). At their inception in 2010, 44 states and the District of Columbia joined either Smarter Balanced or PARCC; in spring 2019 only 15 states and the District of Columbia administered PARCC or Smarter Balanced and many of them did not do

[5] State Responsibilities for Assessments & Locally Selected, Nationally Recognized High School Academic Assessments. 34 C.F.R. § 200.2-3 (2016).

so for high school (Gewertz, 2019; Robelen, 2010). Some states are creating state tests using a combination of Common Core assessment items and their own state-customized items, some are partnering with vendors such as Pearson and Cambium Assessment (formerly American Institutes of Research Assessment) to develop their own tests, and some are using consortia tests in grades 3 through 8 and the ACT or the SAT at high school. In addition to choosing their assessment and vendor, states also define achievement levels differently.

There is a trade-off, however, between variability and comparability. At what point are comparisons between state test results no longer defensible? To what extent can states that modify assessment content and/or procedures continue to use the consortia's validity studies to support claims about the validity of their own state uses of the assessment? At what point in the modification of content and/or procedures does a state's use of the consortia's score scale become no longer meaningful?

As observed by Haertel and Linn in 1996, when examining issues of comparability in the context of performance assessment,

> Different aspects of comparability will be more or less relevant in a given situation. As with any psychometric desiderata, the stringency of comparability requirements will depend on the kind of decision being made (e.g., "absolute" decisions about status with respect to a cutting score versus "relative" decisions about the rank ordering of students or schools); the importance of the consequences attached to those decisions; the level of aggregation at which scores will be reported and used (individuals versus aggregates like classrooms, schools, or states); the relative costs of mistakenly passing versus mistakenly failing an individual; the quality of other relevant, available information and how it is combined … and the ease with which faulty decisions can be detected and revised. (p. 60)

The same principles apply to the current assessment context. This volume seeks to inform the design and use of large-scale assessments to help support intended inferences and actions. Chapter authors, who are all experts in educational assessment, examine the most pressing comparability issues in the current assessment system context and provide suggestions for moving forward. However, before turning to the comparability issues discussed in this volume, we first offer two critical definitions: (1) comparability and (2) assessment system.

DEFINITION OF COMPARABILITY

Users of educational assessments assume that students' scores can be validly compared—they assume *score comparability*—even if those scores come from measurements taken at different times, in different places, or using variations in assessment content and procedures. Ideally, users could be assured that students with the same score possessed the same level of proficiency with respect to the domain of knowledge and skills a test was intended to measure (AERA, APA, & NCME, 2014).

Broadly speaking, there are at least three ways actual test scores necessarily fall short of this ideal. *First*, scores are imprecise—various sources of measurement error affect scores, introducing random error that limits score interpretations. *Second*, with few exceptions, the knowledge and skills a test actually measures do not perfectly

match the range of knowledge and skills that test users wish or intend to measure. *Third*, a range of influences can give rise to systematic differences in scores (i.e., differences in "expected scores") among students who in fact possess equal proficiency with respect to the qualities the test actually measures. This third kind of imperfection, systematic influences that differentially affect scores of different examinees, comprises threats to score comparability and can arise from many sources. These three kinds of limitations can interact in complex ways, but, by and large, the first two—random errors and imperfections in the scope of knowledge and skills measured—on average affect all students' scores in the same way. The third kind of limitation—factors affecting comparability—introduces systematic distortions that may affect score patterns across individuals or groups. This kind of limitation is the primary focus of the present volume. The following brief discussion is by no means exhaustive but is intended to clarify the scope of these comparability concerns addressed by the papers in this volume, and the importance of doing so.

Most obviously, if test performance requires proficiencies irrelevant to the knowledge and skills the test is intended to measure, and if some students' performance suffers due to lack of those irrelevant proficiencies, then the scores of those students are not comparable to the scores of other students. This is a comparability concern because it systematically affects the scores of some students differently from others. On tests intended to measure knowledge and skills other than language proficiency per se (e.g., mathematical computational skills), scores of students hampered by limited language proficiency may be depressed for reasons unrelated to the construct the test is intended to measure. For assessments administered on digital platforms, if some students are unfamiliar with the technology employed, a similar issue may arise. Closely related to issues of irrelevant skill demands are issues of test bias. If item content is more interesting or more familiar to one or another identifiable group of students, score comparability may be compromised.

Threats to comparability may also arise due to differences in test administration or scoring conditions. The scores being compared may have been obtained using different test forms or may be based on different scorers' judgments of students' responses. Students may take digitally administered items on different kinds of devices or use test forms administered at substantially different times during the academic year. Score comparability may also be compromised if students in one jurisdiction perceive a test as "high stakes" and those in another jurisdiction do not, giving rise to differing levels of effort and engagement. In some cases, test administration conditions are deliberately altered to enable more valid measurements of target constructs for students requiring testing accommodations. Although appropriate accommodations can undoubtedly improve score comparability, sound and defensible use of testing accommodations can be challenging. Many of these threats to comparability are amplified when comparisons are made across different assessments and assessment systems.

When scores are compared for groups of students, comparability also demands that the groups be defined consistently, with proper attention to sampling, rules for exclusions and exemptions, and retesting practices.

Additionally, the issue of score comparability requires attention to the inferences drawn from test scores. Consider this scenario: A new, high-stakes test is introduced. Students are retested annually, and, over the first 2 or 3 years the test is in place, average

scores rise dramatically. If the intended inference as to the meaning of the test scores was limited to proficiency with respect to the content sampled on the test, demonstrated in just the ways the test called for, then one might validly infer that the rising scores showed proficiency increasing from year to year. If, however, the test scores are interpreted as indicators of a proficiency with respect to the broader domain of content the test was designed to represent, including both sampled and unsampled content, then the same pattern of rising scores might be attributed, at least in part, to realignment of curriculum and instruction to tested content elements at the expense of untested content elements. From the perspective of that broader intended inference, first-year and subsequent-year scores might not be entirely comparable. As these examples show, comparability is contingent on arguments and evidence about the intended purposes and uses of the test scores being compared. Comparability may be adequate for one interpretive purpose but not another.

DEFINITION OF ASSESSMENT SYSTEM

Throughout this volume, the term *assessment system* is used and we need to be clear about the meaning and scope of this term as used in this volume, especially with respect to other discussions in the broader educational assessment literature (e.g., Herman, 2016; NRC, 2001, 2006). In general, an assessment system implies the existence of multiple assessments designed to function together to fulfill specific interpretive goals and purposes. The assessment system may be composed of assessments that range in form and content from teachers' classroom quizzes and midterm or final exams, to district, national, or international standardized tests. Whatever the specific tests included, the overarching purpose of the collective set of assessments making up the system should be to provide information that serves to promote student learning (e.g., Herman, 2016; Wiggins, 1998). The focus in this volume is on comparability concerns involving assessments that are primarily distal to the classroom—district, state, national, and international assessments.[6]

As noted by Coladarci (2002), "a collection of assessments does not entail a system any more than a pile of bricks constitutes a house." Rather, an assessment system is an assemblage adhering to principles that ensure that the elements are complementary and work together. In the National Research Council (NRC) report *Knowing What Students Know: The Science and Design of Educational Assessment* (NRC, 2001), three major system properties were described: *comprehensive, coherent*, and *continuous*.

Comprehensive means that a range of approaches is used to provide a variety of evidence to support educational decision making. Using multiple types of assessments and indicators that span the ways that a subject is expressed in the curriculum, and in typical instructional practices, enhances the validity and fairness of the inferences drawn by giving students various ways and opportunities to demonstrate their competence.

For the system to support learning, it must also have the property of *coherence*. One dimension of coherence is that there is consistency in the conceptualization of student learning underlying the various assessments within the system. While a state-level

[6] We are not suggesting that this restriction to the definition of assessment system should be broadly employed, just that we are focusing on more summative assessments in this volume.

assessment might be based on a model of learning that is broader and thus less fine grained than the model underlying the assessments used in classrooms, the conceptual base for a state assessment should be the same as that guiding assessment at the classroom level. In this way, results from assessments external to the classroom will be consistent with the more detailed understanding of learning underlying classroom instruction and assessment. The NRC (2006) also discusses the important property of vertical coherence, whereby the different levels of assessments conceptually align with curriculum and instruction at the given grade or academic level.

Finally, an ideal assessment system would be designed to be *continuous*. That is, assessments would measure student progress over time. To provide such pictures of progress, multiple sets of observations over time must be linked conceptually so that change can be observed and interpreted. Models of student progress in learning should underlie the assessment system, and individual assessments should be designed to provide information that maps back to the progression. Thus, continuity calls for alignment along the dimension of time.

Much of what concerns the chapters in this volume are assessments that have been designed for use at levels that are relatively distal in time and space from ongoing classroom instructional and assessment practice. The inferences made about student learning based on such distal assessments require levels and forms of comparability that are typically less critical for the highly contextualized interpretive uses associated with formative and summative classroom purposes.

Unless otherwise indicated, "assessment system" throughout this volume is therefore meant to apply to the types of systems designed to operate outside the classroom interpretive context.[7] It refers to a collection of assessments designed and used to measure student achievement with respect to some common content framework. In addition to the assessments themselves, an assessment system also refers to (1) the rules and policies governing uses of those assessments, (2) the infrastructure required to administer the assessments and to acquire and score students' responses, and (3) the associated reporting structures and associated professional development designed to help users (i.e., students, teachers, parents, educational administrators, and policy makers) interpret the results. An assessment system may serve as the foundation for an accountability system that employs test scores, usually in conjunction with other kinds of information, to quantify the performance of students, schools, or districts and possibly to determine rewards or sanctions. As used here, however, "assessment system" is limited to the mechanisms for measuring and reporting student achievement to promote student learning and does not include the additional data sources and decision rules incorporated in an accountability system. However, at points we do make reference to the links and tensions between an assessment system and its accompanying accountability system.

This volume's working definition of an assessment system is motivated by the high-stakes accountability context of K–12 education and testing. As used here, in addition to academic achievement tests, "assessment system" also encompasses tests of English language proficiency (including initial screening tests) used to classify students

[7] Such assessment systems may be within an individual school district or state, may span multiple states, or may span countries.

as English learners or as fully English proficient. It excludes, however, assessments of classroom climate and measures of socioemotional learning, important as these may be. Indicators of various student demographic variables may be used to report student achievement according to racial/ethnic group, gender, socioeconomic status, student language background, or other categories, but in this volume, these demographic variables are not treated as part of the assessment system itself. Also excluded are indicators of opportunity to learn (OTL), although consideration of OTL and related contextual factors may be essential if certain test score interpretations are to be fair and useful.

Content framework. The content framework undergirding an assessment system describes, in greater or lesser detail, what is to be taught and learned through formal schooling. At some level of abstraction, all of the assessments within a single assessment system can be linked back to a single, common framework, such as the CCSS. On closer examination, however, there may be multiple content definitions at various levels of specificity. The foundational document may set forth broad instructional goals but is unlikely to provide sufficient detail to guide either classroom instruction or the design of assessments. The CCSS, for example, is explicitly *not* a curriculum framework or test specification. Various intermediate documents may elaborate on the overarching framework. Some may prescribe the scope and sequence of instruction, and others may include "test blueprints" prescribing the mix of item types and content elements in particular assessments. The same assessment system may serve classrooms in which various textbooks are used for a given subject at a given grade level. These different textbooks may differ somewhat in the content and organization of instruction they prescribe, and, of course, individual teachers may adapt curriculum materials in different ways.

Types of assessments. The assessments within an assessment system may span multiple grade levels and subject areas. They may include specific assemblies of items used together (fixed-form tests), item assemblies created dynamically from calibrated item pools (computer adaptive testing), or both. Typically, there will be multiple forms of any given test for use over time (e.g., annual testing), as well as special forms for students requiring accommodations. Assessments may include multiple-choice items, other forms of selected-response items, constructed-response exercises, performance tasks, or various mixtures of these or other item formats.

Comparability and context. If an assessment system is to provide accurate, fair, and useful information to meet the needs of various audiences, it must be carefully designed to work within a given context. Alignment with content frameworks is fundamental to meeting virtually all such information needs. Users of information from an assessment system will appropriately assume that test scores reflect students' mastery of significant content, going beyond the answers to specific questions actually administered. Alignment is essential if content frameworks are to provide trustworthy guidance as to the meaning of test scores.

In addition to alignment with content frameworks, many uses and interpretations will depend on the *comparability* of scores across students, across student groups, across

schools, across years, and, in some cases, across different kinds of assessments included within the assessment system. Clearly, not all assessments within an assessment system can or need to be directly comparable. There is not a requirement for a common scale for scores from all of the constituent assessments. It should also be noted that comparability is a matter of degree. At one extreme, scores from alternate forms of the same test might meet stringent psychometric requirements for *equating*, a fine tuning of score scales from different test forms that renders their scores entirely comparable. At the other extreme, there may be no common scale connecting a teacher's informal classroom assessment, used formatively to guide instruction, and the end-of-year, external summative assessment covering the same content, even though scores on those two very different tests would probably be positively correlated.

Forms and degrees of comparability for different purposes are complex and resist easy categorization. To give just a few examples, absent some compelling rationale, achievement standards defining (for example) "proficient" should be established in such a way that aggregate proportions designated as "proficient" do not change erratically from one grade level to the next, nor should they be grossly disproportionate across subject areas. If an assessment system offers the choice, for some assessment, of paper-and-pencil versus computer-based testing, or, more generally, a choice among digital platforms for computer-based assessments, then in order for the obtained scores to be reportable on a common scale, they should meet stringent standards for comparability. If scores for a certain demographic subgroup are to be compared across jurisdictions, those subgroups should be defined everywhere in the same way. To the degree possible, scores from students tested with accommodations should be reportable on the same scale, and interpretable in the same way, as for students tested without accommodations. These and other comparability issues are discussed throughout this volume.

COMPARABILITY ISSUES ADDRESSED IN THIS VOLUME

As noted above, this volume is an attempt to provide guidance to key stakeholders, including state-level educational assessment and accountability decision makers, leaders, and coordinators; consortia members; technical advisors; vendors; and the educational measurement community regarding *how much* and *what types* of variation in assessment content and procedures can be allowed, while still maintaining comparability across jurisdictions and student populations. The volume also provides guidance and caveats to policy makers using assessment data to enact legislation, regulations, and district- and school-level guidance and also provides greater context for media using test scores to make comparability determinations. Here we briefly summarize the comparability issues addressed in this volume.

Comparability of Individual Students' Scores on the "Same Test" (Chapter 2). While comparability is often thought of as comparability across states or different tests, the first chapter in this volume begins by grounding the reader in comparability issues in the interpretation of a single test score of a single student. Charles DePascale and Brian Gong explain that while on large-scale assessments, individual student test scores on the same test are expected to be interchangeable (i.e., the student would be expected to receive the same test score if they took a different form of the test or took

the test under different conditions), meeting this goal is challenging. The term "same test" refers to various cases in which students may take different sets of items under different conditions. This chapter addresses how to evaluate whether comparability across forms and/or conditions is sufficient to support a particular inference or test use. Intended comparability may be supported through careful design decisions and psychometric procedures. There are also external threats that might affect the accuracy and/or interpretation of students' scores. Students' opportunity to learn the content assessed and familiarity with the item formats and tools used on the assessment are two types of comparability threats related primarily to their prior experiences. Threats to comparability that may arise from differences in the intended uses of the assessment and from different assessment contractors' processing of the "same test" are also discussed. The process of establishing the comparability of individual student scores on the same test involves compiling sufficient evidence to support inferences and actions related to student performance based on those test scores.

Comparability of Aggregated Group Scores on the "Same Test" (Chapter 3). After examining individual students' scores in Chapter 2, Leslie Keng and Scott Marion address the considerations and challenges associated with comparing scores from the same test at the aggregate level, such as between student groups, schools, districts, and states. While many principles and methodological approaches are similar to those addressed in Chapter 2, comparisons of aggregated group scores also must include, among other things, differences across jurisdictions in test delivery platforms, modes of administration, and testing accommodation policies. Since comparability is essential for establishing the validity of inferences, and validity is evaluated in the context of specific purposes and uses, this chapter explores the various uses and purposes associated with comparisons of aggregate performance for tests considered essentially the same; the categories of aggregate measures, or derived scores, used to compare group-level performance; and factors that can affect aggregate-score comparability. Because comparability exists on a continuum, the authors propose criteria that can be used to determine whether the preponderance of evidence supports comparability claims for an intended aggregate-score use or purpose and conclude with a practical framework for evaluating and mitigating threats to the comparability of group scores in current policy and practical contexts.

Comparability Within a Single Assessment System (Chapter 4). Mark Wilson and Richard Wolfe address comparability issues that arise within a single assessment system, focusing on summative results for individuals and aggregates (classrooms, schools, districts, and states). This chapter examines the validity of comparisons across grades, subjects, and years, and in interim results where they are strongly aligned to summative tests. The authors address the question of whether the different parts of the system measure the same or similar variables. As the authors note, test-to-test concordances only are useful or valid if there is confidence that the tests are addressing essentially the same underlying variables; as such, the chapter examines the alignment of subject-matter content, the design of the measurement constructs within the system, and the stringency of the different tests within the system. In essence, are the tests aligned and designed to attend to their intended uses? The chapter also addresses the reliability of the tests with respect to different uses and different levels of aggregation,

as well as the need for transparency in the system (i.e., what information should consumers have available to make decisions, and what level of technical documentation is needed to ensure that a system can be fully reviewed by expert evaluators).

Comparability Across Different Assessment Systems (Chapter 5). In this chapter, Marianne Perie expands the discussion beyond one assessment system and examines comparability issues when interpreting scores across more than one large-scale assessment. Policy makers want to compare performance across states and districts, using measures that go beyond NAEP. For instance, as policy has moved to focus on college readiness, there is also a desire not only to compare tests and state assessments across consortia but also to compare the results of such tests with traditional college admissions tests such as the ACT and the SAT. And, there is interest in international comparisons of state assessments to multinational tests such as the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA). This chapter examines how different assessment systems might address score comparability of students, schools, districts, and states. Specifically, the focus is on elements of assessments required for comparability, understanding score comparability at different levels of aggregation, and psychometric constraints on desired inferences about students and schools across states and countries.

While issues pertaining to EL students, students with disabilities, and students from nondominant linguistic and cultural backgrounds permeate the volume, the steering committee determined that in addition to attention within chapters, comparability issues for these groups should also be the foci for individual chapters. As such, Chapters 6 through 8 address EL students, students with disabilities, and students from varying linguistic and cultural backgrounds as described below.

Comparability When Assessing English Learner Students (Chapter 6). In this chapter, Molly Faulkner-Bond and James Soland identify several decisions and test-score uses specific to EL students in the United States and introduce potential comparability issues concerning generalizations or comparisons about this population of students. These issues begin at the level of defining the population itself; because ELs are identified on the basis of test-based processes, decisions about who belongs in this subgroup, as well as reclassification criteria, may lack comparability across settings. Within the EL subgroup, comparing and interpreting English language proficiency scores is challenging due to differences in how tests are developed and scored, how states weight various subscores, and even how the construct of language proficiency is operationalized across measures. Achievement test score comparisons between ELs and non-ELs may be distorted by potential confounds between language and academic ability. Furthermore, many ELs take achievement tests using accommodations that complicate comparisons if not properly addressed, and ELs can also be part of other subgroups like students with disabilities that necessitate additional accommodations. Finally, using scales to estimate and compare growth for ELs (including comparisons to growth for non-ELs) is complicated by the shifting nature of the EL subgroup. In the chapter, the authors present several considerations for minimizing threats and supporting valid score use, both within and across populations and systems.

Comparability When Assessing Individuals with Disabilities (Chapter 7). Standardized testing procedures are meant to provide a level playing field for all examinees with respect to content tested, test administration procedures, and scoring processes. However, in some cases, aspects of standardized procedures may prevent examinees with disabilities from fully demonstrating their proficiencies. In such cases, accommodations may enable individuals with disabilities to better demonstrate what they know and can do. In this chapter, Stephen Sireci and Maura O'Riordan describe the various types of accommodations provided on statewide and college admissions tests, the resulting issues in score comparability, and how to evaluate the effects of test accommodations. The authors also examine test development procedures that may help make educational tests more accessible to individuals with disabilities, thereby reducing the need for accommodations.

Comparability in Multilingual and Multicultural Assessment Contexts (Chapter 8). Kadriye Ercikan and Han-Hui Por examine the impact of score comparability for students from different linguistic and cultural backgrounds on the validity of inferences from assessments. In addition to comparability issues arising in the context of international assessments given in multiple languages, the issue of consistent score meaning is also a concern for countries with populations from diverse language and sociocultural backgrounds, including countries with large immigrant populations. Recognition of the diversity within the United States led states to develop assessments in multiple languages and provide language tools and accommodations. This chapter highlights the complexity of comparability issues when tests are administered in multiple languages to students from diverse backgrounds and provides recommendations for optimizing comparability of adapted versions of tests.

Interpreting Test-Score Comparisons (Chapter 9). The concluding chapter of the volume, authored by Randy Bennett, is a cross-cutting chapter that examines—with all of the caveats and warnings described in prior chapters—how to best interpret test scores. And, as is likely evident by now, getting meaning from test results requires some type of comparison, be it to other test takers, oneself, or some absolute standard. Comparisons are strongest when the same measure is given under substantively the same conditions to comparable student samples at the same point in time. Comparisons become weaker as the measure, the assessment conditions, student samples, and the time of administration diverge. This chapter addresses when conditions are substantially the same as well as when divergence can occur. With respect to good practice, it is well to note that comparative claim statements can appear (or be implied) in score reports, press releases, websites, and other communications. When making such statements, it is best to determine first whether the same test is being used and if it is administered under the same conditions to comparable student samples at the same point in time. If not, the divergence(s) should be identified and a logical rationale for making the comparison should be articulated. The strength of the comparative claim should be adjusted as a function of (1) the extent to which the instruments, assessment conditions, student samples, and time between administrations diverge, and (2) the extent of the logical and empirical support available to back the claim and technical assistance committee review of this support. This chapter explores comparative claims

across this spectrum and suggests adjustments in terms of level of confidence based on either of these two factors.

As the chapters in this volume show, issues of comparability of assessment results are numerous and challenging but they are not insurmountable. It is our hope that, by surfacing these issues across a range of contexts where comparisons are inevitable, and often critical for informing policy and decision making, such comparisons can be approached in ways that are appropriate and useful. Each of the chapters offers cautions with respect to the types of comparisons of assessment results that are typically desired while also offering recommendations that can lead to more valid and useful inferences for those contexts of use that in turn can support equity, fairness, and enhancement of educational opportunities and outcomes.

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.

Coladarci, T. (2002). Is it a house…or a pile of bricks? Important features of a local assessment System. *Phi Delta Kappan 83*(10), 772–774.

DOEd (U.S. Department of Education). (2009). *Race to the Top program executive summary*. Washington, DC: U.S. Department of Education. Retrieved from https://www2.ed.gov/programs/racetothetop/executive-summary.pdf.

DOEd. (2010). *U.S. Secretary of Education Duncan announces winners of competition to improve student assessments.* Retrieved from https://www.ed.gov/news/press-releases/us-secretary-education-duncan-announces-winners-competition-improve-student-asse.

DOEd. (n.d.). *Awards: Race to the Top assessment program.* Retrieved from https://www2.ed.gov/programs/racetothetop-assessment/awards.html.

Gewertz, C. (2019, April 9). Which states are using PARCC or Smarter Balanced? *Education Week*. Retrieved from http://www.edweek.org/ew/section/multimedia/states-using-parcc-or-smarter-balanced.html.

Haertel, E. H., & Linn, R. L. (1996). Comparability. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 59–78). Washington, DC: National Center for Education Statistics. Retrieved from https://nces.ed.gov/pubs/96802.pdf.

Herman, J. (2016). *Comprehensive standards-based assessment systems supporting learning*. The Center on Standards & Assessment Implementation, WestEd. Retrieved from https://www.csai-online.org/sites/default/files/resources/4666/CAS_SupportingLearning.pdf.

Hess, R. (2014, April 30). SBAC responds to my queries about the Common Core tests. *Education Week*. Retrieved from http://blogs.edweek.org/edweek/rick_hess_straight_up/2014/04/sbac_offers_answers_to_my_queries_about_the_common_core_tests.html.

Jochim, A., & McGuinn, P. (2016). The politics of the Common Core assessments. *Education Next, 16*(4).

Kaestle, C. F. (1983). *The pillars of the republic: Common schools and American society, 1780–1860*. New York: Hill and Wang.

Kaestle, C. F. (2012). *The testing policy in the United States: A historical perspective.* The Gordon Commission on the Future of Assessment in Education. Retrieved from https://www.ets.org/Media/Research/pdf/kaestle_testing_policy_us_historical_perspective.pdf.

Kalk, J. (2019, December 11). *New statewide test results show achievement gap throughout Cedar Rapids Community School District*. Retrieved from https://www.kcrg.com/content/news/Cedar-Rapids-schools--566098691.html.

Kirkman, A. (2020, January 3). Majority of South Bend schools do not meet federal expectations, new report states. *South Bend Tribune*. Retrieved from https://www.southbendtribune.com/news/education/majority-of-south-bend-schools-do-not-meet-federal-expectations/article_44cfffd2-2da0-11ea-a537-a7e687e54948.html.

Marion, S. (2017, January 3). What's next for the Common Core and its assessments? *Future Ed.* Retrieved from https://www.future-ed.org/whats-next-for-the-common-core-and-its-assessments.

Moss, P. A., Pullin, D. C., Gee, J. P., Haertel, E. H., & Young, L. J. (Eds.) (2008). *Assessment, equity, and opportunity to learn*. New York: Cambridge University Press.

NAEd (National Academy of Education). (2009). *Education policy white paper on standards, assessments, and accountability* (L. Shepard, J. Hannaway, & E. Baker, Eds.). Washington, DC: Author.

NRC (National Research Council). (1999a). *Uncommon measures: Equivalence and linkage among educational tests* (M. J. Feuer et al., Eds.). Washington, DC: National Academy Press. Retrieved from https://www.nap.edu/read/6332/chapter/1.

NRC. (1999b). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress* (J. W. Pellegrino, L. R. Jones, & K. J. Mitchell, Eds.). Washington, DC: National Academy Press. Retrieved from https://www.nap.edu/read/6296/chapter/1#ii.

NRC. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

NRC. (2006). *Systems for state science assessment* (M. Wilson & M. Bertenthal, Eds.). Washington, DC: The National Academies Press.

OTA (Office of Technology Assessment). (1992, February). *Testing in American schools: Asking the right questions* (OTA-SET-519). Washington, DC: U.S. Government Printing Office.

Resnick, D. P. (1982). History of educational testing. In *Ability testing: Uses, consequences, and controversies* (Vol. 2, pp. 173–194). Washington, DC: National Research Council.

Rimel, A. (2019, September 21). Oregon dips in standardized test scores, mixed bag for mid-valley. *Corvallis Gazette-Times*. Retrieved from https://www.gazettetimes.com/news/local/oregon-dips-in-standardized-tests-mixed-bag-for-mid-valley/article_cc900868-a925-59ce-923e-14af4ce01da3.html.

Robelen, E. W. (2010, September 2). Two state groups win federal grants for Common tests. *Education Week*. Retrieved from https://www.edweek.org/ew/articles/2010/09/02/03assess.h30.html.

Ryan, K. (2019, August 27). *Maryland's PARCC results show dip in math, improvements in English*. Retrieved from https://wtop.com/maryland/2019/08/marylands-parcc-results-show-dip-in-math-improvements-in-english.

Sethrie, J. (2020, January 6). Minnesota report card: Small schools score higher. *Fillmore County Journal*. Retrieved from https://fillmorecountyjournal.com/minnesota-report-card-small-schools-score-higher.

Symposium: Equity in educational assessment. (1994). *Harvard Educational Review, 64*(1).

Tyrrell, J. (2019, April 4). *Survey: 45% of test-takers boycott ELA exam*. Retrieved from https://www.newsday.com/long-island/education/schools-ela-opt-outs-test-boycott-1.29381145.

Vinovskis, M. A. (2019). What use is educational assessment? In A. I. Berman, M. J. Feuer, & J. W. Pellegrino (Eds.), *The Annals of the American Academy of Political and Social Science, 683*, 22–37.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco, CA: Jossey-Bass.