Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Fall 2020, Vol. 39, No. 3, pp. 100–105

Standardization and *UNDERSTAND*ardization in Educational Assessment

Stephen G. Sireci, *University of Massachusetts Amherst*

Abstract: *Educational tests are standardized so that all examinees are tested on the same material, under the same testing conditions, and with the same scoring protocols. This uniformity is designed to provide a level "playing field" for all examinees so that the test is "the same" for everyone. Thus, standardization is designed to promote fairness in testing. In practice, the material tested, the conditions under which a test is administered, and the scoring processes, are often too rigid to provide the intended level playing field. For example, standardized testing conditions may interact with personal characteristics of examinees that affect test performance, but are not construct-relevant. Thus, more flexibility in standardization is needed to account for the diversity of experiences, talents, and handicaps of the incredibly heterogeneous populations of examinees we currently assess. Traditional standardization procedures grew out of experimental psychology and psychophysics laboratories where keeping all conditions constant was crucial. Today, accounting for and measuring what is not constant across examinees is crucial to valid construct interpretations. To meet this need I introduce the concept of* understandardization, *which refers to ensuring sufficient flexibility in standardized testing conditions to yield the most accurate measurement of proficiency for* each *examinee.*

Keywords: culturally responsive assessment, educational testing, scaling, standardization, test accommodations, validity

The history of modern educational testing is often traced to Alfred Binet, who in 1904 developed the first test of educational proficiencies to ensure Parisian children would not be denied the education they deserve (Kaestle, 2012; Sireci & Randall, in press). However, the process of measuring unobservable "constructs" within people actually began about 40 years earlier with the work of Ernst Weber and Gustav Fechner, who were physicists working at the University of Leipzig in Germany. These physicists, who eventually became known as "psychophysicists," were exploring the lack of a direct link between changes in a physical object, such as the weight of a block, and people's perceptions of those changes. They found that people did not directly perceive the magnitude of physical differences until the difference hit some threshold—a "just noticeable difference," as Weber termed it. Fechner later used these just noticeable differences to develop the first psychological scale of "sensation" (Sireci, Wainer, & Braun, 1998).

The conditions under which Weber and Fechner conducted their experiments were meticulously controlled so that the most precise values of the just noticeable differences could be calculated. Participants in these experiments became the instruments used to develop the sensation scale. It was the measurement, not the people themselves, that were of primary interest in the emergence and evolution of psychophysics. The carefully controlled procedures of Weber and Fechner led to the birth of experimental psychology, via Wilhem Wundt, the father of experimental psychology, who was also at Leipzig around this time. It was from this orientation that modern psychometrics was born. As Geisinger (2000) described,

> Wundt's influence was so strong that perhaps the most common theme among the early leaders in testing was that the administration of measures needed to be strictly controlled so that they were interchangeable across individuals. With such strict controls, all differences in performance were the result of individual differences rather than differences in test administrations or "error" as had been believed previously (p. 118).

As educational testing became widespread in the early 20th century (e.g., the large-scale administrations of the army alpha and beta tests around 1917, the Scholastic Aptitude Test [SAT] around 1928), standardization became one of the most critically important requirements to defend fair and accurate measurement. The uniform conditions achieved in the laboratories of psychophysicists and experimental psychologists were adhered to as closely as possible. As large-scale tests began to permeate American society "Keep everything the same for everyone" was the mantra for valid measurement.

Although we can sympathize with this perspective today, we realize that in educational testing, *students* are the most important part of the measurement process, not the measure itself, or the measurement scale. Contemporary psychometrics and educational research have clearly determined that overly rigid testing procedures can impede accurate measurement of students' proficiencies, and distort test score interpretations (e.g., Arbuthnot, 2020; Berman, Haertel, & Pellegrino, 2020; Gordon Commission, 2013; Pellegrino,

Chudowsky, & Glaser, 2001; Winter, 2010). Universal test design and accommodations to standardized testing conditions are examples of efforts to account for overly prescriptive standardization procedures in testing. However, these examples leave the impression of being exceptions to the rule, rather than redefining the rule.

As the other articles in this special issue emphasize, it is time to redefine the rule. It is time to move away from the ideals of 19th-century psychological measurement to arrive at better measurement of, and better outcomes for, the students we measure. In this article, I describe a new way of thinking about standardization in educational testing. The reconceptualization I propose aims to retain the main goals of standardization—providing a level playing field for all examinees—but accomplishes it by building flexibility into the standardization process, rather than rigidity. Rigidity leads to exclusion, and the goal of educational measurement is not to measure the students who are easiest to measure and who conform to the most dominant culture associated with the measurement enterprise, but rather to obtain the best measure of each and every student's proficiencies. The main point of this reconceptualization is we must *understand* the numerous dimensions of heterogeneity that exist within the populations of people we test, and embed that understanding in our standardization processes. Hence, the term *UNDERSTANDardization*.

Understanding Understandardization

It is important to note from the outset that the key change in moving from standardization to understandardization is not the prefix "under," but rather the prefix "understand." That is, incorporating understanding of examinee heterogeneity into the formal standardization process does not give us something less than standardization; it gives us a better understanding of the testing conditions that are best for each examinee. Such understanding facilitates valid interpretation of test performance for each examinee. The goal in understandardization is to understand (a) what each student brings to the testing situation in addition to the proficiency measured, (b) how these personal characteristics may interact with testing conditions, and (c) how the testing conditions can be flexible enough to accommodate and account for these potential interactions. Thus, understandardization first requires better understanding of the student population, which helps anticipate potential test administration problems. Anticipating these problems allows for the development of strategies to mitigate them through more flexible standardized testing procedures that ultimately lead to more accurate interpretations of students' true proficiencies. In this section, we describe these three aspects of understandardization.

Understanding Student Diversity

Like many countries, the United States has a richness of diversity with respect to language, history, and culture. However, the large-scale educational measurement community did not emerge from this rich diversity of culture; rather, it emerged from the dominant culture—from those who were in power in the early 20th century. Those who were not in power were easily marginalized. Thus, the culture of educational testing in the United States today, grew out of the dominant culture of the times from the early-to-mid 20th century.

As described earlier, it was the work of the U.S. armed services during World War I and World War II that led to the emergence of large-scale testing in the United States (Bunch & Clauser, in press; Kaestle, 2012; Lehman, 1999; Sireci & Randall, in press). Large-scale testing focused first on military recruits and then on college applicants (Lehman, 1999). However, the most common types of educational assessment in use in the United States today are extremely different from those that emerged in the early-to-mid 20th century.

Although testing in the military and for college admissions continues to this day, most educational tests in the United States are administered in public schools and are used for purposes such as reporting on students' mastery of curriculum standards, accountability (for teachers, schools, and districts), and high school graduation. The No Child Left Behind Act of 2001, and its reauthorization—the Every Student Succeeds Act of 2015—mandated annual testing of public school students in the United States in grades 3 through 8, and one grade in high school, in reading and mathematics. States are also required to assess students' science proficiency in three grade levels, and English learners' English proficiency every year until they are reclassified as former English learners. Formative assessments, which are used "by teachers and students during instruction…to adjust ongoing teaching and learning to improve students' achievements…" (Council of Chief State School Officers, 2012, p. 4) are also commonplace in U.S. schools.

Unlike the tests developed for men in the military or predominantly white men applying to colleges in the 1940s, contemporary educational tests involve assessing the full range of student diversity in the United States. Thus today, valid interpretation of students' test scores requires understanding the heterogeneity of the student population with respect to community resources, home resources, family structures, culture, language, communication norms, religious beliefs, educational experiences, and other factors. By understanding the different "funds of knowledge" (González, Moll, & Amanti, 2005) students bring to the testing situation, we can better standardize that situation to support, rather than prohibit, diversity. Standardization should not "wash out" student heterogeneity, it should embrace it.

To use my current home state of Massachusetts as an example, in some school districts, Spanish is the most common language spoken at home. In other districts, it is Polish, Vietnamese, or Portuguese. Although English is the native language for about 90% of the students in Massachusetts, there are entire schools where English is far from the predominant language spoken at home. Can a test designed for the 90% majority of students who are native English speakers work the same way for the 10% who are not? This rhetorical question is not only a research question; it is a test development question. Without considering the special characteristics of these students from the earliest stages of test development, a test will only appropriately measure the students who are easy to measure, and will result in poor measurement for many other types of students. Sireci, Wells, and Hu (2014), for example, found that about 90% of English learners on a statewide biology test either left a constructed-response item blank or received a score of zero on the item. Think about that statistic—the item provided no information for 90% of the students from an important, and populous, subgroup of students in the state. Given that the constructed-response items are worth more points than the multiple-choice items on the test, the impact of this nonresponse on estimation of

English learners' science proficiency is substantial. However, given the relatively low proportion of English learners in the state, this problem would not even be flagged by a traditional item analysis.

Before thinking about ways standardized testing procedures could prevent such nonresponse, remember first that English learners are only one subgroup of students from the dozens that are equally important to consider. Many discussions of fairness in testing focus on English learners and students with disabilities (e.g., Faulkner-Bond & Soland, 2020; Sireci & O'Riordan, 2020). Attention in the measurement literature on these two groups is probably due to Federal laws that have been established to take into account the needs of English learners and students with disabilities (e.g., IDEA, 2004; NCLB, 2002). These laws have allowed accommodations to standard testing conditions to promote fairness in testing (e.g., Faulkner-Bond & Sireci, 2015; Faulkner-Bond & Soland, 2020; Sireci, Banda, & Wells, 2018). However, there are no laws that explicitly address accommodations for other important subgroups of interest such as African Americans or other cultural groups. African Americans represent about 15% of the population of public school students in the United States (National Center for Education Statistics, 2020). Although test development guidelines stress the importance of sensitivity reviews and differential item functioning analyses (i.e., item bias analyses) for identifying and screening test material that may disadvantage or offend cultural groups such as African Americans (e.g., American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014), there are no guidelines for adjusting test administration conditions to promote more valid assessment of this, or other, historically marginalized groups (e.g., indigenous students; refugees, transgender students, etc.).

In a subsequent section of this article I discuss ideas for increasing flexibility in standardization testing conditions to address student diversity head on. The point for the present section is that test developers and testing agencies need to conduct research to understand the diversity within the student population when developing tests and establishing the standardized testing conditions. Cultural diversity should be a key focus in developing this understanding; and as this exploration begins, culture should be broadly defined. For example, Montenegro and Jankowski (2017) suggested viewing culture as,

> (1) the explicit elements that makes people identifiable to a specific group(s) including behaviors, practices, customs, roles, attitudes, appearance, expressions of identity, language, housing region, heritage, race/ethnicity, rituals, religion; (2) the implicit elements that combine a group of people which include their beliefs, values, ethics, gender identity, sexual orientation, common experiences (e.g. military veterans and foster children), social identity; and (3) cognitive elements or the ways that the lived experiences of a group of people affect their acquisition of knowledge, behavior, cognition, communication, expression of knowledge, perceptions of self and others, work ethic, collaboration, and so on. (pp. 8–9)

As a better understanding of student diversity increases, so too will the standardized procedures used to measure these students. Gaining an understanding of the diversity of students within the target population will require reviewing census data, demographic data from school systems, and actually talking to teachers and school leaders about the different types of students in their classrooms. Yes, understanding student diversity requires test developers and testing agencies to break away from their computers and data files, and have conversations with educators.[1]

Identifying Potential Interactions between Student Characteristics and Testing Conditions

Once the heterogeneity and diversity within the student population is understood, potential ways in which different student characteristics can interact with anticipated testing conditions can be identified. As a very simple example, if the test is to be administered on a desktop computer, and the student population contains very tall and very short students, having a standard chair size and computer terminal height will not be appropriate for all students. For student populations that contain English learners, test instructions and other test material in the English language may be similarly problematic. In some testing situations, African American students may find themselves in a space that is very culturally unfamiliar to them, which can cause undue stress. Choices of reading passages, graphs, and tables may be differentially familiar across student groups, and differential familiarity with colloquialisms used on a test could also cause confusion for students from nonmajority cultures. Item formats, such as multiple-choice items may be completely unfamiliar to some cultural groups, particularly those still acculturating to American society. Understanding the importance of doing well on an assessment is also likely to differ across groups of students who differ with respect to culture, poverty, and other sociodemographic variables. Clearly, understanding the importance of the test directly impacts test performance.

The examples in the previous paragraph are but a short list of the types of potential problems that could exist if traditional standardized test development and administration procedures are applied to a contemporary educational assessments in a given context. Once the work to understand the student population is done, brainstorming the potential negative interactions that could occur with traditional standardized testing conditions will lead to improving those conditions. Thus, the initial process in understandardization is to first, understand diversity within the student population; then, critically evaluate the ways in which traditional standardized testing procedures may lead to biased estimates for some students within that population. The third step, adjusting those standardized procedures to eliminate those potential biases is described next.

Making Standardized Testing Conditions More Flexible

The idea of making standardized testing conditions more flexible is not new. It has been proposed and initiated for students with disabilities and English learners for some time (e.g., Koenig & Bachman, 2004). Two important areas of progress for these groups of students are *test accommodations* and *universal test design*. These concepts are relatively well known, and so only brief descriptions are provided here. They are important aspects of understandardization, but they are not the only ones.

Test accommodations refer to deviations from standardized test administration procedures with respect to the presentation of test content, testing time limits, setting in which the test is administered, or way in which students present their responses (Abedi & Ewers, 2013; Sireci & O'Riordan, in press; Sireci, Scarpati, & Li, 2005). Examples of these

accommodations include administering the tests in a separate setting, providing oral accommodations such as reading the test directions aloud to the student, or providing a translated version of the exam in an alternate language (e.g., Spanish). These accommodations are provided to remove any barrier from standard test administration conditions that might hinder students from demonstrating their full potential on an exam.

Universal test design seeks to make test accommodations for students with disabilities and English learners unnecessary. It is "an approach to test design that seeks to maximize accessibility for all intended examinees" (AERA et al., 2014, p. 50). The idea behind universal test design is that if a test and its administration conditions are designed with students with disabilities and linguistic minorities in mind, there will be no need to provide accommodations to these groups. Thus, the goal of universal test design is to make the test and testing situation flexible enough so that accommodations are not necessary (Thompson, Blount, & Thurlow, 2002; Thurlow, Lazarus, Christensen, & Shyyan, 2016).

A popular example of universal test design is removing time limits on a test for all students. What was formally thought of as an accommodation becomes part of the standardized testing conditions. The playing field is still the same for all, but now all students have all the time they need to complete the test. Other examples of universal design are allowing all students to be tested in a separate room, or all students to screen-reading software to read the text associated with a test aloud. Universal design makes the standardized testing conditions more flexible for all students, which is the same idea motivating understandardization. That is, provide flexible testing conditions so that the needs of specific students are taken into account before testing begins.

Although the ideas underlying universal design are the same as those motivating understandardization, with understandardization, we extend the concern for more accurate measurement from students with disabilities and English learners to *all* subgroups of students. For example, in understandardization we explicitly consider historically marginalized groups such as African Americans and impoverished youth, as well as any other groups of interest discovered when investigating the diversity of the student population.

How can we make standardized testing conditions more flexible for these other important subgroups of students via understandardization? Although research in this area is just beginning, the concept of *culturally responsive assessment* represents a promising approach for developing assessment conditions that are more appropriate for African American and other potentially marginalized groups of students. Relatively recent and complex accommodations such as *translanguaging* offer another example. These and other ideas are described next.

Culturally Responsive Assessment as Understandardization

The idea of culturally responsive assessment (Hood, 1998; Koelsch, Estrin, & Farr, 1995; Lee, 1998; Montenegro & Jankowski, 2017) grew out of the call for more culturally responsive pedagogy (Ladson-Billings, 1995). A key idea of culturally responsive pedagogy is to allow students to bring aspects of their culture into the classroom as assets to learning, rather than viewing nonmajority culture as a deficit. Similarly, in culturally responsive assessment, the assessment situation invites students to draw from their cultural experiences to demonstrate their knowledge and skills with respect to the educational constructs measured. Culturally responsive assessment argues that if we accept the fact that students can learn in multiple ways, why should we require them to demonstrate that understanding in a single, specific way? Lee (1998) proposed that culturally responsive assessments be integrated directly with curriculum and instruction and involve tasks that draw on culturally based funds of knowledge from students' families and communities, and from their youth culture.

The term *funds of knowledge* refers to the accumulated knowledge, experiences, and ways of interacting and communicating that are associated with specific cultures, broadly defined. González et al.(2005) defined funds of knowledge as "historically accumulated and culturally developed bodies of knowledge and skills essential for household or individual functioning and well-being" (p. 72). The important point for education and educational assessment is that students can draw from these funds of knowledge to develop strategies for solving problems. Thus, culturally responsive assessment, and hence understandardization, seeks to allow students to draw from these funds of knowledge while taking a test.

One way to allow for funds of knowledge to be accessed during an assessment is to allow students to use *translanguaging* when taking a test. Translanguaging refers to "flexible use of linguistic resources that characterizes bilinguals in their attempt to make sense of their bilingual worlds" (Gándara & Randall, 2019, p. 63). For example bilinguals or English learners can be permitted to switch back and forth between languages when reviewing test instructions, reading test content, or providing their responses to test items. The idea is to allow students' to use their knowledge of more than one language to demonstrate their proficiency in a way that is not restricted via the monolingual restrictions that are typical in standardized testing. As Gándara (2017) explained, "Solutions based on monolingual expectations will never tap the critical characteristic that emerging bilinguals possess: a flexible, fluid and strategic use of multilingual resources" (p. 30). Although test development, administration, and scoring becomes more complicated when translanguaging is permitted as part of standard test administration conditions, it will result in better measurement of the proficiencies of bilingual and emerging bilingual students. Thus, more flexible test administration conditions that allow translanguaging are better than accommodations for linguistic minorities. For example, Gándara and Randall (2019) pointed out,

> Because multilingual students do not behave as multiple monolinguals, translated tests are not a satisfactory solution. Test developers should produce assessments that enable multilingual students to use their entire linguistic repertoires and engage in their natural linguistic practices" (p. 58).

Gándara and Randall (2019) evaluated the use of translanguaging to assess the mathematics competencies of girls in the Democratic Republic of the Congo. Using bilingual test administrators, they allowed the students tested to access test instructions and respond to test items using the Lingala or French languages. They concluded the multilingual test administration and scoring resulted in much more appropriate measurement than a monolingual administration of the test.

Clearly, the logic of translanguaging can be extended to other types of funds of knowledge to make testing conditions

more flexible. It is not the purpose of this article to lay out all of the different options that could or should be embedded into test administration conditions because such ideas are only now emerging and any such list would be incomplete. The important point here is we should investigate the atypical resources nonmajority students draw from to reason and make inferences when solving complex problems, and expand test administration conditions to allow for accessing those resources.

Other Means for Introducing Flexibility into Standardized Testing Conditions

Ideas for more flexible test administration conditions are in their infancy. Some include allowing students to create their own avatar when taking a computer-based test, and to switch avatars as they like. The older notion of *self-adaptive testing* (Wise, Plake, Johnson, & Roos, 1992), where students get to choose whether they want an easier or more difficult question than the one they previously answered, is also an example of building flexibility into standardized testing procedures. Allowing students to bring their own device (e.g., tablet, laptop, cell phone) to access the test also allows flexibility that can be helpful to students. Such flexibility may evoke immediate protests from psychometricians stuck in 20th-century ways of thinking where comparability of test scores and the test score scale are sacrosanct. My response to these protests is the students are more important than the scale, and the scores derived from more flexible assessments that allow students to better demonstrate their skills will be more valid than those from more rigid test administration conditions that hold score comparability above all else. Another idea is to allow students to choose the reading passages or writing prompts to which they respond. Again this may introduce some lack of comparability across test scores (Lukhele, Thissen, & Wainer, 1994), but the small sacrifice in test score comparability may lead to great gains in test score validity.

The key aspect of standardization is providing a level playing field. Allowing all students to have the same, flexible options is fair. Just like some baseball players use batting gloves and some do not, all students should be allowed to change languages when taking an assessment (assuming the assessment is not a measure of proficiency in a specific language), access a calculator (assuming computation is not the proficiency measured), have test directions read aloud or presented in a different language, and other types of supports that could benefit any single student. Allowing all students these options is within the boundaries of uniform testing conditions, even if all students do not take advantage of the option.

Conclusions and Recommendations

The advent of modern educational and psychological measurement focused on evidence that such measurement was possible, and could be quantified on a numerical scale. Now that the utility of educational tests has been established, it is time to put the focus on the students who are tested, and how assessment results can be used to help them achieve success (Gordon, 2020). A key way to do that is to move away from traditionally rigid test administration conditions and allow for flexibility in test administration to accommodate the diverse experiences, talents and needs of contemporary students. We are 20 years into the 21st century. It is time we improve upon the standardization procedures that were established in the 19th and 20th centuries. As Montenegro and Jankowski (2017) described, "There is a difference between assessing all students in the same way in relation to a specific outcome of interest and making sure assessments are appropriate and inclusive of all students" (p. 5).

In this article, I introduced the concept of understandardization, which is the process of establishing flexible standardized testing conditions that are appropriate for the diversity within the student population tested. The understandardization process involves three steps: (a) conducting research to understand the full diversity of students in the population, (b) identifying ways in which anticipated testing conditions may negatively interact with the diversity of the student population, and (c) adjusting the standardization process to be sufficiently flexible so that any negative interactions with student characteristics are mitigated. Understandardization is consistent with universal test design, but adds an additional task of conducting research to understand the student population, which leads to universal design for all types of student diversity, not just diversity due to language proficiency or disability.

Before closing I must acknowledge the task of understandardization is not easy. It is far easier to continue to do testing the way we have done for over 120 years. However, as the other articles in this special volume indicate, those antiquated testing practices are limiting the potential of educational tests, which can propagate the marginalization of historically underrepresented students. I recommend we take the harder path. By better understanding the students we test, and by making standardized testing conditions more flexible, we will arrive at more valid, more just, and more effective measurement for minority and nonminority students alike.

Note

[1]Consider the example with which Professor Gordon began his article in this volume. It was the conversations and observations he had with Ms. Haeussermann, the test administrator, from which he gained his understanding of the greater potential value of educational assessments.

Acknowledgments

The author thanks Edmund Gordon for initiating the important conversations that led to the formulation of the concepts laid out in this article and for his inspiration and support for writing them, and Jennifer Randall for providing helpful references to culturally responsive assessment, and for even more helpful conversations about how we can improve educational assessments for historically marginalized students.

References

Abedi, J., & Ewers, N. (2013). *Accommodations for English language learners and students with disabilities: A research-based decision algorithm*. Smarter Balanced Assessment Consortium. Retrieved from https://portal.smarterbalanced.org/library/en/accommodations-for-english-language-learners-and-students-with-disabilities-a-research-based-decision-algorithm.pdf.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

Arbuthnot, K. (2020). Reimagining assessments in the post pandemic era: Creating a blueprint for the future. *Educational Measurement: Issues and Practice*, *39*, 97–99.

Berman, A., Haertel, E., & Pellegrino, J. (Eds.) (2020). *Comparability issues in large-scale assessment* (pp. 177–204). Washington, DC: National Academy of Education Press.

Bunch, M. & Clauser, B. (Eds.) (in press). *The history of educational measurement: Key advancements in theory, policy, and practice*. New York: Routledge.

Council of Chief State School Officers (2012). *Distinguishing formative assessment from other educational assessment labels*. Washington, DC: Author.

Faulkner-Bond, M. F., & Sireci, S. G. (2015). Validity issues in assessing linguistic minorities. *International Journal of Testing*, *15*, 114–135.

Faulkner-Bond, M., & Soland (2020). Comparability when assessing English learner students. In A. Berman, E. Haertel, & J. Pellegrino (Eds.) *Comparability issues in large-scale assessment* (pp. 149–175). Washington, DC: National Academy of Education Press.

Gándara, F. (2017). Evaluating a translingual administration of the EGMA assessment in the Democratic Republic of the Congo. Unpublished doctoral dissertation, University of Massachusetts Amherst.

Gándara, F., & Randall, J. (2019). Assessing mathematics proficiency of multilingual students: The case for translanguaging in the Democratic Republic of the Congo. *Comparative Education Review*, *68*, 58–78.

Geisinger, K. F. (2000). Psychological testing at the end of the millennium: A brief historical review. *Professional Psychology: Research and Practice*, *31*, 117–118.

Gordon Commission (2013). *To assess, to teach, to learn: A vision for the future of assessment (technical report)*. Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/gordon_commission_technical_report.pdf.

Gordon, E. W. (2020). Toward assessment in the service of learning. *Educational Measurement: Issues and Practice*, *39*, 72–78.

González, N., Moll, L. C., & Amanti, C. (Eds.). (2005). *Funds of knowledge: Theorizing practices in households, communities, and classrooms*. Mahwah, NJ: Lawrence Erlbaum.

Hood, S. (1998). Culturally responsive performance-based assessment: Conceptual and psychometric considerations. *The Journal of Negro Education*, *67*(3), 187. https://doi.org/10.2307/2668188

Individuals With Disabilities Education Act (IDEA). (2004). 20 U.S.C. § 1400.

Kaestle, K. (2012) *Testing policy in the United States- A historical perspective*. http://www.gordoncommission.org/rsc/pdf/kaestle_testing_policy_us_historical_perspective.pdf.

Koelsch, N., Estrin, E., Farr, B. (1995) *Guide to developing equitable performance assessments*. Office of Educational Research and Improvement. WestEd

Koenig, J. A., & Bachman, L. F., (Eds.). (2004). *Keeping score for all: The effects of inclusion and accommodation policies on large-scale educational assessments*. Washington, DC: National Academies Press.

Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal*, *32*, 465–491.

Lee, C. D. (1998). Culturally Responsive Pedagogy and Performance-Based Assessment. *The Journal of Negro Education*, *67*(3), 268. https://doi.org/10.2307/2668195

Lehman, N. (1999). *The big test*. New York: Farrar, Straus, & Giroux.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement*, *31*, 234–250.

Montenegro, E., & Jankowski, N. A. (2017, January). *Equity and assessment: Moving towards culturally responsive assessment (Occasional Paper No. 29)*. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA).

National Center for Education Statistics (2020). *The condition of education 2020*. Washington, DC: Author. Retrieved from https://nces.ed.gov/pubs2020/2020144.pdf.

No Child Left Behind (NCLB) Act of 2001. (2002). Pub. L. No. 107-110, § 101, Stat. 1425.

Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy of Sciences.

Sireci, S. G., & Banda, E., & Wells, C. S. (2018). Promoting valid assessment of students with disabilities and English learners. In Elliott, S. N., Kettler, R. J, Beddow, P. A., & Kurz, A. (Eds.), *Handbook of accessible instruction and testing practices: Issues, innovations, and application* (pp. 231–246). Cham, Switzerland: Sage.

Sireci, S. G., & O'Riordan, M. (2020). Comparability issues in assessing individuals with disabilities. In A. Berman, E. Haertel, & J. Pellegrino (Eds.) *Comparability Issues in Large-Scale Assessment* (pp. 177–204). Washington, DC: National Academy of Education Press.

Sireci, S. G., & Randall, J. (in press). Evolving notions of fairness in testing in the United States. In M. Bunch & B. Clauser (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice*. New York: Routledge.

Sireci, S. G., Scarpati, S., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research*, *75*, 457–490.

Sireci, S. G., Wainer, H., & Braun, H. (1998). Psychometrics, overview. In *Encyclopedia of biostatistics*. New York: John Wiley & Sons.

Sireci, S. G., Wells, C., Hu, H. (2014, April). *Using internal structure validity evidence to evaluate test accommodations*. Paper presented at the annual meeting of the National Council on Measurement in Education, Philadelphia.

Thompson, S., Blount, A., & Thurlow, M. (2002). A summary of research on the effects of test accommodations: 1999 through 2001 (Technical Report No. 34). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M. L., Lazarus, S. S., Christensen, L. L., & Shyyan, V. (2016). *Principles and characteristics of inclusive assessment systems in a changing assessment landscape (NCEO Report No. 400)*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Winter, P. (Ed.), (2010). *Evaluating the comparability of scores from achievement test variations*. Washington, DC: Council of Chief State School Officers.

Wise, S., Plake, B., Johnson, P., & Roos, L. (1992). A comparison of self-adapted and computerized adaptive tests. *Journal of Educational Measurement*, *29*(4), 329–339.

