14

Realizing Fairness Through Accessibility for All Test Takers and for Specific Groups

Michael C. Rodriguez
University of Minnesota Twin Cities

Martha L. Thurlow
University of Minnesota Twin Cities (Retired)

Rodriguez, M. C., & Thurlow, M. L. (2025). Realizing fairness through accessibility for all test takers and for specific groups.
In L. L. Cook M. J. Pitoniak (Eds.), *Educational measurement* (5th ed., pp. 959–1014). Oxford University Press.
DOI: 10.1093/oso/9780197654965.003.0014

The *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA] et al., 2014), hereafter referred to as the *Standards*, are forward thinking in projecting fairness as a foundational concept for all educational and psychological measurement. The *Standards* described several general views of fairness, including equitable treatment during the testing process, absence of measurement bias, validity of individual test score interpretation and use, and access to the intended measured constructs and achievement domains. Increasingly, the concept of *accessibility* has been identified as a key aspect of realizing fairness in educational and psychological testing. Overall, a fair test maximizes construct and domain-relevant information, regardless of irrelevant individual characteristics and test contexts, to ensure valid test score interpretation and use for all test takers.

Zwick (this volume) notes the interconnectedness of the three foundational chapters in the *Standards*—validity, reliability, and fairness—as well as their interrelationships with concepts such as bias, construct-irrelevant variance, opportunity to learn, standardization, universal design for assessment (UDA), and others. In the foundational discussion of fairness in the *Standards* is the recognition that fairness is a broad concept that includes an overarching standard about the entire testing process being designed to minimize construct-irrelevant variance to promote valid score interpretations. The fairness standards are organized in four clusters that address (a) test design, development, administration, and scoring; (b) validity of score interpretations; (c) accommodations to remove construct-irrelevant barriers; and (d) safeguards against inappropriate score interpretations. Zwick moves beyond these concepts to consider fairness in the context of opportunity to learn and selection (admissions to higher education and employment). She also argues that fairness must be addressed through a larger validity argument which provides "a chain of reasoning using documented evidence to support the fairness of the intended uses and interpretations of test scores" (Zieky, 2016, as cited in Zwick, this volume, p. ##).

For the first time, the *Standards* eliminated chapters devoted to individuals with specific characteristics, such as those with disabilities and those from different cultural and linguistic backgrounds. Instead, a focus on all test takers was woven throughout the chapters with an emphasis on meeting the needs of all test takers throughout test design, development, implementation, reporting, interpretation, and use. Realizing accessibility is critical in this approach, which is much more than just providing accommodations to some students. Accessibility is a concept that has emerged slowly and been applied to educational and psychological tests as it has become clear that test takers include culturally and linguistically diverse groups of individuals.

The purpose of this chapter is to apply the theory and concepts of a fairness argument to all test-taking populations, both in general and for specific groups of test takers. We explore these through the lens of accessibility and what it means in general for all test takers, as well as through the perspective of specific groups, including individuals with disabilities, English learners, those from different racialized and ethnicized groups, and those living in poverty. We address fairness for all types of assessments, including tests of achievement most often used in preschool through Grade 12 (P–12) school settings (e.g., used for formative, interim, or summative purposes), higher education classroom settings, campus-wide assessments of student learning outcomes, admissions tests (e.g., ACT, SAT, GRE, Law School Admission Test [LSAT]), and licensure and certification exams. At the end of most sections, we provide implications for practice, or practice and research, and summarize some of the key guidance promoted and informed by the scholars reviewed in those sections.

Accessibility is the condition where test takers have "unobstructed opportunity to demonstrate their standing on the construct(s) being measured" (AERA et al., 2014, p. 49). A related approach, which has become a common component of test design principles, is UDA, borrowed from principles of architecture regarding "test design that seeks to maximize accessibility for all intended examinees" (AERA et al., 2014, p. 50), to be responsive to test-taker characteristics and the testing context(s). In addition, we expand the notions of UDA by including the principles of universal design for learning (UDL) created by CAST (https://www.cast.org/), a broader conceptualization of access that acknowledges that access to content and assessment tasks is necessary, but not sufficient; test takers need access to learning opportunities. Together, a fair test "reflects the same construct(s) for all test takers, and scores from it have the same meaning for all individuals in the intended population" (AERA et al., 2014, p. 50), regardless of racialized status, ethnicity, gender, age, socioeconomic status, special education status, or cultural or linguistic background. We use the term racialized status, rather than race, acknowledging that race is a social construct and that minoritized individuals are ascribed to racial groups, marginalized, and subjected to racism and oppression (in many ways, this includes ethnicity). As employed by Russell (2024), the term racialized denotes "the active process of creating categories into which humans are membered" (p. viii).

Furthermore, fairness is of concern in each stage of test design, development, administration, scoring, interpretation, and use. The fairness chapter in the *Standards* addresses four general views, including (a) fair and equitable treatment of test takers, (b) fairness as access to the measured constructs, (c) fairness as lack of measurement bias, and (d) fairness as validity of score interpretation and use; the latter two views are addressed comprehensively by Zwick (this volume). Finally, we explore the promise of culturally and linguistically responsive assessment to provide accessibility and achieve greater test fairness.

Fairness is often discussed in terms of test-taker characteristics needing additional or unique considerations in test development (see Huff et al., this volume), administration and scoring (see Shermis et al., this volume), and reporting (see Zenisky et al., this volume). However, it is important to acknowledge that fairness challenges are less about the test-taker characteristics and more about the quality, inclusiveness, and responsiveness of test design and implementation, where challenges result when test characteristics interact with test-taker characteristics in unintended (or unanticipated) ways (Kettler et al., 2018), thus the focus on accessibility and universal design principles. We agree that radical goals can be achieved "when educational tests focus on promoting the success of *all* students" (Sireci, 2021, p. 7).

FAIRNESS IN EDUCATIONAL TESTS AND ASSESSMENTS

For our purpose, we focus on fairness in the measurement of education achievement and do not directly address more purely psychological assessments, employment tests such as cognitive ability or personality tests, assessments of job performance, or other tests typically in the purview of clinical or industrial-organizational psychology. Both educational and psychological tests and assessments have fairness concerns (e.g., see Geisinger, 2015), and although many of the issues and contexts have parallels in psychological testing, our focus is more directly applied to testing in education settings or for education purposes.

Fairness and accessibility are concerns for all types of educational tests and assessments and all proposed purposes (including those for which validity evidence may or may not exist). For tests with more high-stakes outcomes, more rigorous evidence is needed to support fairness and accessibility, as with validity evidence. This includes tests of achievement for interim and summative purposes with implications for individuals (see Brookhart & DePascale, and Ho & Polikoff, both this volume); tests used for admissions to higher education, such as the ACT, SAT, GRE, LSAT, Medical College Admission Test [MCAT], GMAT, and others (see Camara et al., this volume); tests used for licensure and certification (see Margolis et al., this volume); and tests used for assessment of so-called social and emotional, character, behavioral, or intra- and interpersonal skills (see Kyllonen & Zu, this volume).

Our purpose is to apply the theory and concepts of fairness (see Zwick, this volume) to test-taking populations, both in general (i.e., accessibility for all test takers) and for specific groups (e.g., English learners, individuals with disabilities). "Regardless of the purpose of testing, the goal of fairness is to maximize, to the extent possible, the opportunity for test takers to demonstrate their standing on the construct(s) the test is intended to measure" (AERA et al., 2014, p. 51). Of course, such opportunity to demonstrate standing is provided by the test developer and administrator; the test should enable such access given the needs of the test taker.

A comment is needed at this point regarding terms used throughout the chapter, including test and assessment and construct and domain. The authors of the *Standards* described a test as a tool or procedure to systematically sample test-taker behavior in a specified domain in a standardized process. Assessment includes a broader set of systematic methods for obtaining information about a person's characteristics or performance; assessment includes processes to collect relevant evidence, interpret the resulting information, and inform a decision (Russell, 2022). Both include systematic processes and both require domain specifications to support intended interpretations and uses.

In this chapter, we use the word *test* when we mean a tool containing items and tasks administered and scored in systematic ways to measure (a quantitative activity) what a person knows and can do relative to the targeted domain; we use *assessment* when we mean the broader range of processes, both qualitative and quantitative, informal and formal, to similarly collect evidence of what a person knows and can do. In addition, a *construct* is a specific "concept or characteristic that a test is designed to measure" (AERA et al., 2014, p. 217), whereas a content *domain* is "the set of behaviors, knowledge, skills, abilities, attitudes, or other characteristics to be measured by a test, represented in detailed test specifications and often organized into categories by which items are classified" (AERA et al., 2014, p. 218). Although construct and domain appear to be used interchangeably in most educational measurement texts, we prefer the term domain in most education contexts because the tests we write about primarily include measures of achievement, with systematic reference (inference) to what people know and can do as reflected in test specifications (often including many constructs). We use the terms employed by other authors when representing their specific arguments and claims.

Fairness and the Choice to Test

The fact that a test is used presupposes the use of an inclusive decision process that determined that testing is an appropriate, meaningful, and useful method to collect evidence to support decision-making. At this early stage in the testing process, fairness considerations abound. Many decisions occur on a regular basis employing information about knowledge and skills across the life span, from early childhood through careers (e.g., progression through P–12 grade levels, placement in special education or gifted/talented programs, admissions to higher education institutions, or licensure and certification). Evidence can be collected through a wide range of informal to formal qualitative and quantitative approaches. Especially in accountability-based settings, decisions are improved when they are informed; yet what qualifies as evidence is value laden and culture bound. Drawing heavily on the work of Messick (1975, 1980, 1981), Kane (1992) argued that "the legitimacy of test use rests on the assumptions about the possible outcomes (intended and unintended) of the decision to be made and on the values associated with these different outcomes" (p. 530). Messick (1965) argued for attention to regulating ethical standards to justify test use.

Values enter our considerations in a number of ways, including the framing of the decisions that need to be made in education settings, as well as the inclusivity of who is at the table to determine the relevance of impending decisions. When testing is introduced as a means through which evidence can be collected, additional questions must be addressed. Those questions include what sources of evidence to gather to inform decisions; what we value as evidence can vary substantially and impact the determination of what to measure and how to measure it. Such questions are addressed in the final section, where we discuss accessibility through cultural and linguistic responsiveness.

Because fairness addresses the treatment of individuals and groups, a decision to test must consider the appropriateness of test use for all individuals and groups in the test-taker population, including those who may be tested in the near future. This includes opportunity to learn, prior test experience, access to test preparation, access to the test content itself in terms of cultural and linguistic accessibility and language of testing, and ability to respond to items or tasks as required by the test administration procedures.

"A proposal to use a test for a particular purpose, then, should be appraised in light of the probable future consequences of the testing, not only in terms of what it entails but also in terms of what it makes more likely" (Messick, 1975, p. 962). Zwick (this volume) discusses the possibility that, in some cases, the decision to not use a test is most appropriate. For example, if the use of a diagnostic reading test for the purpose of identifying students who need reading remediation identifies 80% to 90% of the students of color in a large population of mostly nonimmigrant children but only 10% of White students, a decision to not employ such a test is warranted, regardless of technical or psychometric quality.

Implications for Practice

Practitioners should secure diverse participation and include key stakeholders in the decision-making process regarding whether to use a test to inform certain decisions. Fairness requires consideration of whether testing is appropriate for all potential test takers and whether a test is of high enough technical or psychometric quality across the test-taker population to warrant use. Test developers and test users should collaboratively contend with the request of Messick (1975) to appraise the probable uses and future consequences of the testing program.

FAIRNESS AS ACCESSIBILITY FOR ALL TEST TAKERS

Accessibility Defined and Explored

"Optimal accessibility is implicitly promised to all" test takers (Kettler et al., 2018, p. 1), as is fairness in testing. Accessibility principles apply to instruction as the opportunity for individuals to learn the disciplinary knowledge and practices through an intended curriculum and to assessment as the opportunity for test takers to demonstrate the disciplinary knowledge and practices required by the test content (Kettler et al., 2009). Full access to instruction and assessment minimizes bias and increases fairness (Kettler et al., 2018), connecting the instruction and assessment accessibility principles directly to the notion of accessibility in the *Standards*.

Equity in access has been an important objective of federal education legislation (e.g., the Civil Rights Act of 1964 and Elementary and Secondary Education Act of 1965 [ESEA]) and case law (e.g., *Brown v. Board of Education*, 1954), particularly in the context of special education (e.g., Education for All Handicapped Children Act of 1975, reauthorized as the Individuals With Disabilities Education Act [IDEA] in 1990). A large body of legislative and federal laws and administrative rules protects inclusion and equitable access to instruction and assessments.

Prior to the 2014 *Standards*, researchers and practitioners were exploring and clarifying the meaning of accessibility. Willingham and Cole (1997) indicated that fair tests measure the same construct for all test takers with the same meaning for all individuals in the population. More specifically, test accessibility is "the degree to which a test and its constituent item set permit the test taker to demonstrate his or her knowledge of the target construct" (Beddow, 2012, p. 98). These definitions are consistent with the notions of fairness in the *Standards*, and here we embrace the definition offered by Beddow.

Test score error variance due to accessibility barriers could originate from physical, perceptive, receptive, emotive, and cognitive sources, sources that may be linked to test or item characteristics, administrator characteristics, or administration contexts and practices (Beddow, 2012; Elliott et al., 2018). In its most general sense, "accessibility—defined as the extent to which a product, environment, or system eliminates barriers and permits equal use of components and for a diverse population of individuals—is necessary for effective instruction and fair testing" (Elliott et al., 2018, p. 1). In the broader context of fairness, we include potential accessibility-limiting sources due to social, cultural, and linguistic contexts and other construct-irrelevant characteristics of test takers, not directly addressed in the sources previously discussed. In particular, accessibility principles that recognize the unique challenges of ensuring that assessments are accessible for Indigenous populations are promising. For example, Hawaiian educators developed principles for Hawaiian culture–based education focused on the purpose of the test, the design and content of the test, the context in which the test is conducted, and how the results are used (Nā Lau Lama, 2007). This work resulted in tests that embodied shared meanings, aligned with the Indigenous elements in the curriculum and language of instruction (in some cases, bilingual instruction), and sustained culturally grounded practices such as hōʹike, which includes meaningful authentic tasks (Nā Lau Lama, 2007).

Accessibility Factors That Affect Fairness

There are three components to accessibility relevant to test fairness, including access skills, target skills, and adaptations (Ketterlin-Geller, 2008; Kettler et al., 2018). Test takers need access skills to allow them to perform well on tests, skills that we often take for granted and assume are present prior to testing. In most cases, such skills are developed through test-taking experiences, including informal classroom tests, and sometimes through test preparation courses or training programs. In some cases, test preparation can help level the playing field and secure an equitable level of test-taking skills across test takers. A common recommendation when employing new item types or preparing test takers with little test-taking experience is to provide opportunities for students to practice so that they can know what is expected and how to interact and respond to test items. However, there are limits to test preparation, where some activities may cross ethical boundaries and cause inferences to be invalidated and, as such, not support fairness as accessibility. These include, as the most egregious, releasing and practicing on operational items or focusing effort on studying released test content and items over a deeper exploration of the learning objectives emerging from content standards on which the test is based. However, as suggested previously, some practice may be useful for those unfamiliar with the item formats or tasks.

Test takers need to explore, understand, and employ disciplinary knowledge and skills required to perform well on a specific test given the targeted knowledge and skills intended to be measured. Much of this chapter addresses accessibility to target knowledge and skills. There is a substantial literature on the accessibility of instruction to support opportunity to learn (Kettler et al., 2018). And in some cases, accommodations or modifications are needed to facilitate access to both instruction and testing. One area of persistent challenge is the degree to which the disciplinary knowledge and practices are culturally or linguistically relevant to the test taker—making explicit the limited degree to which a construct or domain is defined and represented in instruction, in practice, and on the test.

What do we mean by "the degree to which a test and its constituent item set permit the test taker to demonstrate his or her knowledge of the target construct" (Beddow, 2012, p. 98)? Our operational definition of accessibility has a hidden component of significant ambiguity—the target construct. We may simply accept the test developer's target domain (construct) as declared in the test specifications. Or, we might ask: Does the domain definition prohibit the possibility of allowing some test takers to demonstrate their disciplinary knowledge and skills from culturally and linguistically relevant disciplinary knowledge and practices? In the United States, much of the curriculum employed in P–12 and higher education is deeply entrenched in middle-class Euro-centered and colonialist history and perspectives of what people should know and be able to do. Consider the practice of wild rice cultivation (a centuries-long tradition in American Indian communities in Minnesota), which embodies disciplinary knowledge and practices from science, mathematics, history, and agriculture and provides rich opportunities to explore reading, writing, and storytelling; test items that place subject matter disciplinary knowledge in contexts known to students make them relevant and accessible (see, for example, Randall, 2021). Culturally and linguistically relevant disciplinary knowledge and practices are important sources to secure deeper accessibility and fairness in testing via fair domain definitions. Construct and domain definitions play a critical role in test fairness and validation (Randall et al., 2023).

Others also have argued that fairness is enhanced through principles of access via adaptations, including test modifications or accommodations (e.g., Rose et al., 2005; Thompson et al., 2004; Thurlow et al., 2016; Thurlow, Liu, et al., 2013; Zieky, 2016), whereas, in some cases, fairness and validity are reduced in the presence of barriers to access or sources of construct-irrelevant variance. We emphasize the applicability of accessibility principles across the life span, in the many settings and purposes of testing. In those contexts, there are multiple levels at which tests are administered, interpreted, and used, including interpretation and use at the individual (e.g., GRE) and group (e.g., National Assessment of Educational Progress [NAEP]) levels. An important component of accessibility, when no other modifications are possible, is the appropriate provision and use of accommodations.

There are many aspects of test design, administration, and reporting and use that potentially interact with test-taker characteristics that may call into question the fairness of test accessibility. Although test developers work to attend to every possible personal characteristic or experience in the testing process, some groups have historically faced discrimination and significant barriers to education and career opportunities. In a similar way, there are as many testing conditions and contexts under which fairness can be realized as there are potentials for barriers and interference for fair test interpretation and use. Here, we briefly describe several of the more prominent and potentially promising conditions under which fairness can be enhanced and supported. We also highlight the importance of accessibility features and accommodations for supporting fair testing and note the role of testing consequences.

Implications for Practice

Test developers and users should ensure that all test takers have adequate test-taking access skills and have access to appropriate test preparation that is available freely or at low cost. Test developers must make explicit the source of target domains and explore alternative sources of disciplinary knowledge and practices, vis-à-vis the diversity within the target test-taker population. As necessary, when additional adaptations are not possible, accommodations may be identified that fulfill the goals of access and fairness.

FAIRNESS AS ACCESSIBILITY FOR SPECIFIC GROUPS OF TEST TAKERS

Much of the history of thinking about accessibility as an avenue to fairness has grown out of concerns about fairness of testing results for specific groups of individuals. Considerations of fairness have been enhanced through principles of access (e.g., Nā Lau Lama, 2007; Rose et al., 2005; Thompson et al., 2004; Thurlow et al., 2016; Thurlow, Liu, et al., 2013; Zieky, 2016) that identify ways to reduce the presence of barriers to access or sources of construct-irrelevant variance. Accessibility principles can be applied across the life span, in the many settings and purposes of testing. In those contexts, there are multiple levels at which tests are administered, interpreted, and used, at both the individual and the group levels. An important component of accessibility has traditionally focused on the provision of accommodations and their appropriate use.

Initially, the push for thinking about how to make tests available to individuals grew out of the need to identify ways to allow individuals who were blind, who were deaf, or who had certain types of learning disabilities to take tests that would produce results that could be interpreted in the same way as the results of other individuals. We provide here an overview of the progression of work starting in the 1980s. It was not until the 1990s that discussions about accessibility began for English learners. Again, these efforts were focused almost always on accommodations. In the early 2000s, a number of states were engaged in enhanced assessment grants to transform accommodation practices and support item development and test delivery; these efforts were brought to scale in a large funding opportunity for consortia of states to develop college- and career-readiness assessments that were to be accessible to the greatest numbers of students (Race to the Top Fund, 2009), formalizing a paradigm shift in thinking about accessibility (Larson et al., 2020). Rather than a paradigm in which accommodations were the only avenue to accessibility, accessibility became the guiding goal for all tested individuals, with tiers of support from universal design for all, to features designated for any student by an adult or decision-making team, to accommodations for students with disabilities and English learners.

It is because of the historical progression in thinking about accessibility that we first address fairness through accessibility for individuals with disabilities and then for English learners. Following these two groups, we explore how accessibility has been part of the vision for other groups, including racialized and ethnicized groups, and individuals from low socioeconomic backgrounds. There is support for such attention in the *Standards*:

> For some test takers, factors related to individual characteristics such as age, race, ethnicity, socioeconomic status, cultural background, disability, and/or English language proficiency may restrict accessibility and thus interfere with the measurement of the construct(s) of interest. (AERA et al., 2014, p. 52)

However, we note that it is not racial stratification or other characteristics themselves that are the relevant factor, but the discrimination and racism that create disparities in outcomes; this is addressed more directly in the final section on accessibility through cultural and linguistic responsiveness.

Individuals With Disabilities

Early work focused on nonstandard administrations of higher education admissions testing (Laing & Farmer, 1984; Willingham et al., 1988) as the way to enable certain individuals with disabilities to participate in assessments. The researchers who examined nonstandard administrations of the ACT (Laing & Farmer, 1984) and the SAT and GRE (Willingham et al., 1988) focused on who might be eligible for accommodations and the effects of specific accommodations (e.g., statistical differences in predictive validity, reliability, factor structure, differential item functioning). During this early research, so-called nonstandard administrations (i.e., accommodations) included the use of braille, large type, and extended time.

Defining accessibility as the provision of accommodations continued for decades and early on expanded beyond college admissions exams to assessments administered to K–12 students with disabilities. Accommodations initially were defined as changes to the materials or procedures of testing that allowed students with disabilities to take tests but did not change the content or targeted knowledge and skills being measured (Pitoniak & Royer, 2001; Thurlow et al., 2003). These notions were later refined to mean changes that minimized construct-irrelevant variance and promoted the measurement of what students with disabilities know and can do, thus providing access while ensuring the validity of score interpretations and uses when accommodations are used (Bolt & Roach, 2009; Laitusis & Cook, 2007). In the *Standards*, test accommodations are

> adjustments that do not alter the assessed construct that are applied to test presentation, environment, content, format (including response format), or administration conditions for particular test takers, and that are embedded within assessments or applied after the assessment is designed. Tests or assessments with such accommodations, and their scores, are said to be *accommodated*. Accommodated scores should be sufficiently comparable to unaccommodated scores that they can be aggregated together. (AERA et al., 2014, p. 215)

Use of accommodations as a means to access for K–12 students was supported through policy changes, which resulted in part from analyses of state testing policies (Thurlow et al., 1993, 1995). The 1997 reauthorization of IDEA required that all students with disabilities participate in state- and district-wide assessments with accommodations as appropriate. This requirement heightened awareness of the role of accommodations during assessments and, in turn, their use during instruction and their implications for score interpretation and use (Heaney & Pullin, 1998). It also led, after considerable study, to the provision of accommodations in NAEP (Lawton, 1995; Olson & Goldstein, 1996).

Concern about access to the general education curriculum and assessments through the availability of accommodations in instruction and during testing has generated a rich body of research (e.g., Buzick & Stone, 2014; Cawthon & Leppo, 2013; Gregg & Nelson, 2012; Laitusis & Cook, 2007; Thurlow & Kopriva, 2015). Extensive work has improved the analyses of state assessment accommodations policies (e.g., Lazarus et al., 2009) and documentation of the percentage of students with individualized education program (IEP)-assigned accommodations (see Thurlow & Wu, 2019).

In 2004, the reauthorization of IDEA required states to report on the number of students with disabilities receiving accommodations during statewide assessments. These data revealed the large variation in the percentage of students with disabilities who are assigned accommodations across states (Thurlow & Wu, 2019), which, depending on the grade and subject area, could range from as little as 0.7% to as high as 89.9% of students with disabilities. They also provided information about differences in accommodations assigned for reading and mathematics tests and by grade level. Unfortunately, only state-level data for general education assessments are available; those data do not provide information on whether this type of variability exists for students who have the same disability category label.

Researchers, through hundreds of studies, have focused on a wide array of accommodations (e.g., read aloud, text to speech, sign language, word prediction) for these test takers and examined not only the effects of accommodations, but also the results of surveys about their use. A substantial body of research has been accumulated on the effectiveness of accommodations for individuals with disabilities. Phillips (1994), in presenting a case for defending the validity of accommodations, asked, "Would nondisabled examinees benefit if allowed the same accommodation?" (p. 104), where an affirmative answer would suggest that the accommodation may not be defensible. The differential-boost hypothesis (Fuchs et al., 2000) proposed that an accommodation that increases accessibility for individuals with unique needs should result in higher scores for individuals with disabilities but not for individuals without disabilities. Later, this hypothesis was replaced with the interaction hypothesis (Sireci et al., 2005), which proposed that the increase in performance would be greater for individuals with disabilities than for other individuals, but that the scores of all test takers might increase some.

Several meta-analyses and reviews have been conducted to aggregate the effects of accommodations for individuals with disabilities (Buzick & Stone, 2014; Gregg & Nelson, 2012; Rogers et al., 2014, 2016, 2019, 2020). Although it was generally reported that accommodations do in fact make a difference in the performance of individuals with disabilities, those effects varied in complex ways across test subject matter, the nature of the accommodation, and the ways in which accommodations were administered.

The singular focus on accommodations for access and fairness changed in the 1990s when advocates realized that accommodations were not a sufficient mechanism for providing access for some students with disabilities. In 1997, IDEA required that states develop (by the year 2000) alternate assessments for those students who were unable to participate in the general assessment with or without accommodations. The evolution in the development of these alternate assessments covered years in which there was confusion about the performance standards of other alternate assessment options that had been allowed by policy and regulations, such as alternate assessments based on grade-level achievement standards (Quenemoen, 2009; Wiener, 2005) and alternate assessments based on modified achievement standards (AA-MAS; Lazarus et al., 2015). The AA-MAS generated concerns about lower (inequitable) levels of expectation (Thurlow, Lazarus, & Bechard, 2013) and psychometric quality (Rodriguez, 2009, 2011); eventually, the regulation allowing it was rescinded (U.S. Department of Education, 2015), just months before the enactment of the Every Student Succeeds Act (ESSA, reauthorization of ESEA), which disallowed any alternate assessment other than the alternate assessment aligned to alternate academic achievement standards (AA-AAAS) for ESSA accountability purposes.

ESSA clarified that the AA-AAAS is appropriate only for students with the most significant cognitive disabilities. Because *students with the most significant cognitive disabilities* is not an IDEA disability category, states are required to develop their own guidelines and definitions (Thurlow et al., 2019). Researchers have found, though, that students who participate in AA-AAAS tend to have intellectual disabilities, autism, and multiple disabilities (though not all students in those categories have the most significant cognitive disabilities), and yet still vary widely in their characteristics (Erickson & Geist, 2016; Erickson & Quick, 2017; Kearns et al., 2011; Towles-Reeves et al., 2009).

Further, ESSA clarified that students with the most significant cognitive disabilities are to be taught and assessed on the same content standards as their peers without disabilities, although possibly at reduced depth, breadth, and complexity. The test development community has been challenged by the need to develop AA-AAAS appropriate for the range of students in this group, with the goal of producing results that support fair and valid test interpretations and uses. Despite these challenges, AA-AAAS are held to the same requirements as other tests for design and development that supports fair and valid test interpretations and uses (U.S. Department of Education, 2018). Much has been learned in the process of developing AA-AAAS that meet these requirements (e.g., Thurlow & Quenemoen, 2016), yet the expansion of what accessibility means for these students continues to be a challenge. The progression of thinking about accessibility for school-age students with disabilities leads to the argument that, for certain individuals, access could mean different standards for performance, in part because of the possibility of differences in disciplinary knowledge and skills covered in instructional practices and tailored approaches to learning to meet individualized learning goals, as well as extended learning periods through age 21 and beyond. This shift in thinking has not (yet) drifted into other assessments and other individuals being assessed.

As noted earlier, nonstandard administrations began in the 1980s with attention to accessibility for individuals with disabilities outside the K–12 school system. By 1990, the Americans With Disabilities Act required that any business or agency receiving federal funds had to make accommodations available. It stipulated

> (A) making existing facilities used by employees readily accessible to and usable by individuals with disabilities; and (B) . . . acquisition or modification of equipment or devices, appropriate adjustment or modifications of examinations, training materials or policies, the provision of qualified readers or interpreters, and other similar accommodations for individual with disabilities. (42 U.S.C. 12/11, Section 101[9])

Clarification of the individuals who qualified for accommodations was provided in the 2008 reauthorization of the Americans With Disabilities Act; this clarification was viewed as broadening the population of individuals with disabilities who were eligible for accommodations in higher education and the workplace. Because of concerns about the provision of these accommodations in practice, the U.S. Government Accountability Office (USGAO, 2011) issued a report calling for improved federal enforcement of the right to testing accommodations. It brought renewed attention to the role of accommodations regardless of age, as well as to the decisions made about whether to provide accommodations to individuals:

> Given the critical role that standardized tests play in making decisions on higher education admissions, licensure, and job placement, federal laws require that individuals with disabilities are able to access these tests in a manner that allows them to accurately demonstrate their skill level. While testing companies reported providing thousands of test takers with accommodations in the most recent testing year, test takers and disability advocates continue to raise questions about whether testing companies are complying with the law in making their determinations. (USGAO, 2011, p. 29)

Attention to college admissions testing continues to increase, in part because of state adoptions of college admissions tests since the NCLB act required high school statewide assessment. This attention is to a broader array of individuals with disabilities and a broader set of accommodations. Although testing companies report providing thousands of test takers with accommodations, test takers and disability advocates continue to question whether testing companies are complying with the law in making their determinations (USGAO, 2011).

Including the K–12 and higher education systems, individuals with disabilities comprise about 13.6% of the population (2021 American Community Survey; U.S. Census Bureau, 2023). These individuals are, at various times in their lives, taking tests. For example, the number of individuals with disabilities taking the Praxis II tests for educators is increasing (U.S. Department of Education, 2010). Estimates of overall participation of individuals with disabilities in credentialing exams are not available (van den Heuvel et al., 2016).

Attention also has been given to considerations for individuals with disabilities taking other types of tests, particularly those used for career decisions (Nester, 1993), including law exams (Eichorn, 1997; Thurlow et al., 1997) and medical exams (Little, 1999), to name a few. These early works almost exclusively focused on a concern about giving test takers an unfair advantage by providing them with accommodations during testing (Phillips, 1994).

We acknowledge that early employment of accommodations was an important vehicle to provide access to tests for many individuals. However, in the spirit of fairness through access, these earlier attempts were not sufficient for many to accurately and meaningfully represent their disciplinary knowledge and skills in ways that were relevant and responsive to the purposes of the tests.

The addition of alternate assessments was one way to provide access to assessments for those students with more significant or challenging disabilities. Still, over time, this access avenue was limited by federal law to a very small percentage of the student population, thus limiting their potential as a general accessibility approach. In addition, universal design principles have received partial attention rather than full application and use in a way that would eliminate much of the need for accommodations. Part of the hesitation may be due to the need to document the access needs of students with disabilities, which are generally documented on IEPs as accommodations. Further, it is only recently that other avenues to access, such as computer-adaptive tests, have attempted to conform to a person's ability level by focusing item selection to optimally locate the person on the ability continuum (reducing the need to administer items that are too hard or too easy and thus less informative). A major limitation with computer-adaptive tests is that they select from items that are limited in scope, modality, and format, as constrained in the item and test specifications. As of yet, there are not computer-adaptive tests that either select or build tasks (through automated item generation or the use of artificial intelligence) that acknowledge how the test taker might best engage in an assessment of their content knowledge and skills.

Universal Design

Universal design is largely known in the assessment industry through descriptions in the *Standards* and publications of the National Center on Educational Outcomes (Thompson et al., 2004, 2005), referred to as UDA throughout this chapter. In the *Standards*, UDA is described as a means to maximize fairness by defining constructs precisely, avoiding item or test characteristics that may introduce construct-irrelevant variance specific to some groups, and being intentional about selection of font size, administration time, and linguistic load, among others. In addition, UDA calls on test developers to "avoid item contexts that would likely be unfamiliar to individuals because of their cultural background" (AERA et al., 2014, p. 58).

CAST was founded in 1984 and has championed UDL. Although they have undergone revision over time, CAST promotes UDL guidelines to secure access and participation of all learners in challenging and appropriate learning opportunities. The guidelines are based on the why, what, and how of learning, through the provision of multiple means of engagement (stimulating and motivating learning), representation (presenting content in multiple ways), and action and expression (differentiating ways for students to express their knowledge and skills). In 2020, CAST initiated a community-driven process to revise the UDL guidelines (https://udlguidelines.cast.org/), with attention to equity, inclusion, and fairness. These guidelines are much broader than those for UDA and explicitly enable culturally and linguistically responsive approaches to assessment through the acknowledgment of multiple ways (means) of knowing and doing (we address this directly in the final section).

Implications for Practice and Research

Test developers and users should employ the principles of UDA and UDL for test and assessment development and administration and, as a corollary, encourage the use of UDL to enhance accessibility of instruction and learning experiences. Test developers and researchers should support research programs addressing the effectiveness of the application of UDA and UDL, particularly regarding the extent to which more intensive applications of these principles might reduce requests for accommodations. In addition, state and federal education leaders should intensify efforts to develop high-quality AA-AAAS as a positive direction to achieve accessibility and acknowledge important differences in how some individuals learn, access disciplinary knowledge and skills, and convey what they know and can do.

English Learners

English learners emerged as a group that required special consideration for inclusion in K–12 education testing programs almost immediately after concerns were initially raised about the exclusion of students with disabilities (see August & Hakuta, 1994; LaCelle-Peterson & Rivera, 1994). In general, in the United States English learners are students whose native language is not English and who are eligible for English language development services. They have difficulties in reading, writing, speaking, or listening in English that are severe enough to deny them the ability to achieve in classrooms where the language of instruction and assessment is English (National Center on Educational Outcomes, 2020). These characteristics result in difficulty accessing not only the content of instruction but also the assessments designed to measure their content knowledge and skills, as well as their English language proficiency.

In the K–12 education system, identified English learners are eligible to receive English language development services, with the expectation that eventually they will leave English learner status because they have gained proficiency in English reading, writing, speaking, and listening (Linquanti et al., 2013), thus being prepared to fully participate in English-instruction classrooms. Researchers have estimated a 3- to 7-year period to develop academic English proficiency to support schooling, with a slightly longer time for older immigrant English learners, with a shorter range of 3 to 5 years for oral proficiency and a longer time given socioeconomic background and age of immigration (Collier, 1989; Garcia, 2000; Hakuta et al., 2000).

English learners have been the subject of research on assessment accessibility. In part, this attention followed the attention given to students with disabilities, and much of the research has been carried out in the K–12 school system. Similar to individuals with disabilities, access to the curriculum and to assessments is a concern for English learners. Increasingly, researchers focused on the effects of various accommodations for English learners (e.g., Abedi & Hejri, 2004; Abedi et al., 2000, 2004; Kopriva et al., 2016; Kosak-Babuder et al., 2019). A substantial body of research has been accumulated on the effectiveness of accommodations, following a similar methodological path from a differential-boost hypothesis to an interaction hypothesis for these students. A number of meta-analyses and reviews also have been conducted to aggregate the effects of accommodations for English learners (Kieffer et al., 2009; Li, 2014; Liu et al., 2018; Vanchu-Orosco, 2012). Findings generally are less promising for English learners than for students with disabilities, but still vary in complex ways across test subject matter, the nature of the accommodation, and methods of accommodation administration. In a systematic review, Rios et al. (2020) reported that for the most part, accommodations used with English learners lacked evidence of effectiveness; most were accommodations employed for individuals with disabilities.

Researchers confirm that accessibility is a broad issue that applies beyond school years. With the National Education Longitudinal Study (National Center for Education Statistics [NCES], 1988), researchers revealed that 12 years after their eighth-grade year, only about 13% of English learners had earned a bachelor's degree, compared to 25% of non–English learners (Kanno & Cromley, 2013). Attention to the college-entrance assessments that these and other students take to gain access to postsecondary education is increasing (e.g., Moore et al., 2018), yet remains underdeveloped.

There is a growing body of work that supports the use of bilingual or multilingual assessments, in cases where English would be the language of instruction and assessment. Bilingual assessment can take multiple forms (Caesar, 2020). On a minimalist side, with digital delivery of tests, the test developer can embed a translation dictionary in the test by enabling a hover-over dictionary tool (or some other process to get immediate translation of each word in a test). A more encompassing approach is to provide a bilingual experience, where items are presented in a native language and English simultaneously (e.g., side by side). Bilingual assessment embodies the characteristics of accessibility by opening up access to the test and test questions by eliminating the barriers due to language (e.g., see efforts with the National Assessment of Educational Progress; NCES, 2023).

Another approach is to create assessments in multiple languages, as done with the international assessment survey programs (e.g., Trends in International Mathematics and Science Study [TIMSS] of the International Association for the Evaluation of Educational Achievement), especially in contexts with non-English instruction, including in learning settings that are engaging in language revitalization and immersion. Multilingual assessment efforts require additional layers of psychometric analyses to ensure score comparability because the explicit intention is for international comparisons of disciplinary knowledge and skills. Testing programs develop and deploy common content frameworks that are used to develop test specifications through which items are developed and translated into dozens of languages.

When the intent is to provide optimal access, translation makes sense. But in the United States, where the primary language of instruction is English and although there is no official national language, most state testing programs function under an assumption that an important outcome of education is proficiency in English (implied by federal school accountability regulations). However, there are states that test in multiple languages at various grade levels in both reading and mathematics to accessibly measure disciplinary knowledge and skills without the interference of English language proficiency. At least one other state, Hawaii, has encompassed Indigenous language knowledge and practices into state achievement content standards and thus has included tasks relative to Indigenous language in the state achievement tests. There are a growing number of examples of the inclusion of Indigenous language knowledge and practices in national assessment programs across the globe (e.g., Guatemala, New Zealand, Canada, and Australia).

The National Council on Measurement in Education (NCME, 2020) statement on testing with English learners promotes an approach to test development and administration that acknowledges specific factors relevant to these learners. The statement emphasized consideration of the following: (a) language is culturally grounded because culture is often expressed through language; (b) when possible, the language of the test should be consistent with the language of instruction; (c) test administration modes should be familiar for English learners; (d) English receptive and productive language skills should be strong enough to provide full access to the test; and (e) the presence of special needs or disabilities further complicates test accessibility. With greater understanding of the nature of English language development, test developers will achieve greater fairness in testing for all students.

Faulkner-Bond (2020) argued for the consideration of the language of test takers relative to disciplinary language that may be embedded in the test. She recommended oversampling English learners in test development and validation processes to gain deeper understanding of the role of language uses in local settings, in instruction, and within disciplines, as well as discipline-specific language use in the articulation of content standards (how we describe what test takers are expected to know and be able to do). The fundamental ways test takers come to learn and understand disciplinary content and practices are influenced by local contexts, sociocultural contexts, instruction (as practiced locally), and other ways. The operational assumption that test scores are not influenced by language proficiency not only ignores these contexts, but also ignores that we often know less about the language proficiency of non-English learners than we do of English learners. After reviewing the structure and results of several assessments of English language proficiency, regarding comparability, Faulkner-Bond suggested that the focus should be on comparability of uses and interpretations rather than psychometrics, as well as on how proficiency is defined and measured.

Finally, there is growing recognition that native language skills are important to not only sustain, but also continually develop, particularly in early childhood (Duran et al., 2019; Wackerle-Hollman et al., 2019, 2022). In addition, most states award a seal of biliteracy

> in recognition of students who have studied and attained proficiency in two or more languages by high school graduation . . . to help students recognize the value of their academic success and see the tangible benefits of being bilingual. (https://sealofbiliteracy.org/)

Sufficient evidence exists to support multilingual development, begging the question as to why so many tests continue to be administered in English only.

Implications for Practice and Research

Test developers and users should consider the role of language of instruction, language proficiencies in test and native languages, native and test language similarities, availability of validity evidence to support score interpretation and use, evidence regarding fairness, possible presence of specific disabilities, quality of English learner identification, and quality of language instruction and supports. They also should appreciate and recognize the sociocultural contexts in which students live and learn, from test and assessment design to score reporting, interpretation, and use. The research basis for the use of accommodations with English learners is inconsistent and limited, with stronger promise based on research with forms of bilingual assessment—an area that deserves additional attention, consistent with the goals of accessibility and UDL.

Racialized/Ethnicized Groups

Because of the persistence of disparities between racialized and ethnicized groups in education opportunities and outcomes, federal legislation required states to report on the education achievement levels of major racialized or ethnicized groups. We acknowledge that these persistent disparities are largely driven by oppression and racism, limiting education access and opportunities. Because of this, federal legislation has been used as a tool to ensure education access and opportunities. As part of President L. B. Johnson's War on Poverty, the ESEA authorized the federal government to direct federal dollars to the most disadvantaged children through state governments, resulting in wide expansion of state departments of education. Through the 1994 reauthorization of ESEA as the Improving America's Schools Act, states were required to implement challenging content and performance standards and state assessments of them; to include all students in them; to report on performance broken down by racialized and ethnicized status, gender, English proficiency, migrant status, disability, and economic status; and to hold schools accountable. Because many states either did not recognize the requirements or chose to ignore them, the 2001 reauthorization of ESEA, the No Child Left Behind Act (NCLB), significantly increased the Title I accountability requirements with sanctions for failure to do so. States were required to establish grade-level content standards and assessments of achievement of those standards; to report performance (achievement proficiency rates) of students by racialized and ethnicized status, and socioeconomic status, as well as English learners and students in special education; and to hold schools accountable for adequate progress toward the standards (see Ho & Polikoff, this volume).

When investigating fairness and equity in education opportunities and outcomes, racialized categories have been employed for decades (Bohrnstedt et al., 2015; Coleman et al., 1966; Cottrell et al., 2015), with some evidence that standardized testing may play a role in reinforcing segregation (Knoester & Au, 2015). In this work, the concept of race implies a history of housing segregation (resulting in education and occupation disparities) that is connected to occupation segregation (resulting in income and health disparities) that is connected to education segregation (resulting in occupation disparities)—cycles of segregation that limit access to education and career and life opportunities (Cottrell et al., 2015). We note that these cycles are the product of oppression and racism that further the reproduction of disparate opportunities and outcomes. Race and ethnicity are relevant to the extent that they reflect different levels of performance, as an outcome of racism, and such differences call into question qualities of the test itself or result in exacerbating existing disparities and inequality through unintended negative consequences. For example, in 1969, the Association of Black Psychologists called for a moratorium on testing Black children because of disproportionate placement in special education and developmental settings (a result of discrimination and racism), further excluding them from general education opportunities (Garrett Holliday, 2009).

Perhaps the most significant regulations regarding the role of racialized status in education outcomes came in the NCLB reauthorization of ESEA in 2001. President George W. Bush called on the country to end the soft bigotry of low expectations, acknowledging that achievement gaps have persistently fallen along racial and socioeconomic lines. However, Bush mistakenly argued that education achievement gaps produced discrimination, arguing for a causal direction that was inconsistent with the vast body of evidence indicating the opposite (Rubel & McCloskey, 2019). Through the rules and regulations from NCLB, states were required to report student group performance on academic standards-based tests based on racialized and ethnicized status, socioeconomic status, English learner status, and special education status. This was the first federal requirement for states to report on and respond to student group performance that was rigorously enforced, with a requirement that to be counted as achieving adequate yearly progress, 95% of students had to participate in the state's assessments. Prior to this, some students often were exempted from participation because of various exceptionalities, particularly for students receiving special education services (McDonnell & McLaughlin, 1997) or English language services (August & Hakuta, 1997).

Substantial research has addressed the associations between concentrated poverty and segregation in schools with disparities in education supports, lower expectations, increased disciplinary actions, and education achievement and school completion. Even though the nation's diversity continues to increase, schools in the United States are becoming increasingly more segregated, with some regions returning to pre–civil rights levels of segregation (Frankenberg et al., 2019; Reardon et al., 2019); again, we note that racial segregation is a direct result of oppression, discrimination, and racism. Using data from NAEP, it is clear that school achievement test performance is strongly associated with school composition or segregation (Bohrnstedt et al., 2015), where in the United States,

- on average, White students attended schools that were 9% Black, whereas Black students attended schools that were 48% Black;
- schools with greater than 60% Black students tended to be located in urban areas and in the South and the Midwest;
- achievement was lower for Black and White students in schools that had the highest Black density but achievement gaps were not different regarding levels of Black density of schools; and
- the magnitude of achievement gaps, given Black student density of schools, was smaller when accounting for socioeconomic status (SES) and other characteristics, but was still nonignorable.

Many of the challenges that exist regarding racial stratification and educational testing are a function of inappropriate test interpretation and use, which are too often intertwined with discrimination and racism. We acknowledge that, in most settings, test use can provide a deeper understanding of individual and group knowledge, skills, and abilities and inform nearly all education decision-making. However, to realize the potential advances brought about by the use of educational tests, fairness must be secured. Part of the challenges that exist regarding racial stratification and educational testing are the continued limitations of accessibility, particularly those associated with limited cultural and linguistic relevance of tests to culturally and linguistically diverse communities (this topic is discussed more fully in the section "Accessibility Through Cultural and Linguistic Responsiveness").

Implications for Practice and Research

Test developers and users should secure the highest levels of scrutiny of test score quality and fairness. However, our current approach to test development and analysis is limited in the accessibility sense because of the exclusion of considerations for cultural and linguistic backgrounds of intended test takers. Education disparities have a deep and long history rooted in racism, which results in segregation and cycles of poverty. Federal requirements to report achievement levels by race and ethnicity uncover these disparities to some extent, and tests such as NAEP have been important indicators of the nation's progress in addressing achievement disparities and the opportunity gaps producing them. To eliminate large-scale testing because it reflects disparities in opportunity would eliminate an important public policy indicator. However, to use such tests without acknowledging the contexts of racism and racial segregation could result in decisions that maintain segregation and continue to limit access to opportunities.

Socioeconomic Disadvantage

As with racial stratification, disparities in education opportunities and outcomes for students with different levels of SES have been a persistent challenge. SES also has a long-lasting trend connected to segregation and multigenerational poverty (Sharkey & Elwert, 2012). The authorization of ESEA was primarily focused on eliminating poverty through increased education opportunities, and reauthorizations through the Improving America's Schools Act, NCLB, and ESSA maintained the focus on students and schools that experienced persistent disadvantage.

James Popham has long argued for the instructional relevance of testing and assessment. In his work on instructional sensitivity of tests, he argued that instructional insensitivity could result from items that produced correlations between item response and SES, where wealthier students would be more likely to respond correctly (Popham & Ryan, 2012). Popham (2007) argued that the SES composition of schools explained more variance in student performance on accountability tests than the effectiveness with which students have been taught or teachers' instructional efforts. This position is consistent with decades of research where researchers reported that most of the variance in student achievement scores exists within schools rather than between schools, and most of the between-school variance is explained by school demographics resulting from segregation and not indicators of school quality (Coleman et al., 1966; Rodriguez & Nickodem, 2018).

We also acknowledge that SES and racial stratification are highly confounded because there exist significant disparities in wealth as a function of racial stratification, a trend that continues to grow, as has the wealth gap between the richest and poorest families in the United States (Schaeffer, 2020). SES is clearly a construct-irrelevant factor, yet to the extent that it explains variation in access to opportunities and resources, it explains variation in achievement. Once again, elements of segregation and oppression are intertwined with cycles of poverty and limited access to opportunities that may provide greater education and social mobility.

Test Takers Facing Multiple Access Needs

Aspects of test development and administration that limit accessibility to some individuals with certain characteristics associated with education inequity and disparities tend to affect performance for multiple reasons, in part because of the multilayered and intersectional nature of disparities resulting from discrimination and oppression. For example, in 2015–2016, the percentage of students served through IDEA (special education services) was highest among American Indian/Alaskan Native students (17%) and Black students (16%) relative to White students (14%; NCES, 2019). Similarly, in 2016, 24% of Black and Latino children under the age of 18 were in families at or below the poverty level (based on the supplemental poverty measure), relative to 8% of White children (NCES, 2019). In addition, in fall 2017, 45% of the nation's Black and Latino students attended high-poverty schools (more than 75% eligible for free or reduced-priced lunch), whereas 8% of White students did so; 31% of White students attended mid- to high-poverty schools, whereas 75% of Black and Latino students did so (NCES, 2021). Note the differences between poverty rates and attendance in high-poverty schools as another indicator of segregation. There are persistent disparities in SES, employment, income, stable housing, access to healthcare, and other economic indicators across racialized and ethnicized groups, all of which have implications for education opportunities and outcomes.

Although much of the discussion about fairness for specific groups of test takers focuses on one group at a time, in reality, most of the individuals in these groups also belong to other groups for which there are concerns about fairness. For example, individuals with disabilities are frequently among the group of individuals in poverty; the U.S. Census Bureau reported that the poverty rate of individuals with disabilities aged 18–64 in 2018 was 26.9%, compared to 9.5% of individuals without disabilities (Semega et al., 2020); similarly, the U.S. Department of Labor (2021) noted that "across all age groups, persons with disabilities were much less likely to be employed than those with no disabilities" (p. 1). Among school-age children, researchers also have noted the connection between disability status and some disability categories with poverty (e.g., Schifter et al., 2019). Similarly, being an English learner is often associated with poverty and may be associated with immigration status.

Interest continues to grow about how best to assess the school-age population of English learners who have disabilities because of the unique characteristics they present in achieving a fair assessment process, characteristics that most testing programs are not designed to adequately address. The population of school-age English learners with disabilities is increasing at a higher rate than the population of students with disabilities overall (Wu et al., 2021). Historically, the access needs of these students have been addressed in one of two ways. First, these students often have been treated as having special education needs in ways that eclipsed their need for English language development services and accessibility supports (National Council on Disability, 2018). This meant that when they were assessed, consideration was given to the disability access needs but not to their limited English needs. Second, as the access needs of English learners have been recognized, educators have been asked to think of a student's needs as though the student were two individuals—addressing the disability needs first, then addressing the English language needs; if there were conflicts in accessibility supports, then, again, their disability needs would be treated as more important than their English language development needs.

For too long, attention to the access needs of English learners with disabilities only focused on those English learners who did not have significant cognitive disabilities (National Center on Educational Outcomes, 2014). Questions about how to provide assessments that meet their access needs have been raised, with little action until federal law required that these students participate in both alternate achievement assessments and English language proficiency assessments (Christensen et al., 2018; Thurlow et al., 2017; Winter et al., 2018).

Educational measurement practices regarding accessibility for students with multiple categories of access needs continue to improve. Most states provide assessment accessibility manuals that address both students with disabilities and English learners, consistent with the recommended approaches of the Council of Chief State School Officers (Lazarus et al., 2021). However, these efforts are less likely to be occurring beyond testing in K–12 education settings, such as higher education admissions assessments (Lazarus & Thurlow, 2016) and certification and licensure assessments (Lazarus et al., 2017).

Implications for Practice

Test developers and users should analyze the role of disabilities, English learner status, racial stratification, living in poverty, and other public policy relevant characteristics, simultaneously or in combination to obtain a clear picture of item and test performance in complex social contexts. For example, they could employ or create methods to disentangle certain learning disabilities and English learner status because the natural process of being a multilingual learner may present behaviors that appear to indicate language delays. Additional recommendations are provided in the following sections regarding the role of differential item functioning and measurement invariance analyses. In the prevailing empirical research paradigm, we test interaction terms before testing and interpreting main effects. If the association between *X* and *Y* depends on *Z*, then the main effect of *X* is misinterpreted without considering the level of *Z*. In some cases, the multiple effects of *X* and *Z* on *Y* are additive or even multiplicative, and possibly nonlinear.

REALIZING FAIRNESS THROUGH TEST DEVELOPMENT

Test Development

Fairness finds its introduction in the earliest test development stages, from conceptualization to administration. Principled assessment design (PAD; see Huff et al., this volume) provides a flexible and explicit comprehensive approach to test development that acknowledges the important roles of UDA and fairness principles. As described by Huff et al., PAD requires the definition of the construct or domain to be based on disciplinary learning science evidence of how individuals learn and develop disciplinary knowledge and skills. Second, PAD requires "the explicit articulation of all assumptions, design decisions, and rationales for those decisions" (p. 446) through the employment of a coherent set of design tools. Third, PAD also makes explicit the reality that assessment is a process of reasoning from imperfect evidence, but when the evidence is grounded in principled design, tests can meet their most challenging purposes. Huff and colleagues acknowledge the relevance and connections between accessibility and culturally and linguistically responsive approaches to test development, noting that "without the precision and transparency demanded by the PAD process, incorporating accessibility and culturally and linguistically responsive features could jeopardize the validity of the inferences about what students know and can do" (p. 494). The authors also argue that evaluation of cultural and linguistic responsiveness is not done one item at a time, but across an item pool or an entire test form. In addition, accessibility and cultural and linguistic responsiveness are associated with motivation and engagement, which further enhance accessibility for all students.

Embedded in PAD are principles of UDA (Ketterlin-Geller et al., 2015; Thompson et al., 2002, 2004), which require test developers to fully consider and embrace all possible test takers during each stage of test development, beginning with the clarity of the construct definition or domain. Particularly when considering the purposes driving the development of a test and the construct definition stages, the extent to which purpose and construct definition apply consistently to all test takers will support fairness. Principles of accessible measurement (Beddow, 2012) also require greater attention to construct definition, to prevent many instances of differential item functioning and promote measurement invariance. In addition, while addressing the assessment of achievement with a focus on classroom assessment, Shepard et al. (2018a) reminded us of the importance of context:

> All learning is fundamentally social, involving the individual's use of shared language, tools, norms and practices in interaction with his or her social context. . . . Sociocultural theory offers a powerful, integrative account of how motivational aspects of learning—such as self-regulation, self-efficacy, sense of belonging, and identity—are completely entwined with cognitive development. (p. 23)

This position has implications for domain specification and test development. Construct or domain definitions are often underdeveloped (naturally limited because of the lack of inclusion of diverse developers or narrow specification of disciplinary knowledge and practices), from classroom and state achievement tests to professional certification assessments. In their argument centering on the role of sociocultural learning theory as a way to bring coherence to curriculum, instruction, and assessment, Shepard et al. (2018a, 2018b) argued that assessments designed to facilitate inferences about learning must be constructed with research-based models or theories of learning (the first component of PAD).

Regarding test development, fairness can be improved through explicit attention to fairness guidelines (Zieky, 2016) and quality assurance procedures (Allalouf, 2007, 2011, 2017) in the professional practices that produce test blueprints and specifications, item specifications, item-writing and -editing guidelines, item-writer training, item review processes, test assembly and review, scoring rules and computational procedures, and item and test analyses procedures of field-test and operational administration results (Oliveri & von Davier, 2016). Item-writing guidelines exist that specifically promote greater accessibility and fairness (Haladyna & Rodriguez, 2013; Rodriguez, 2011; Rodriguez et al., 2014; Zieky, 2016; see also ETS, 2022). Given the many approaches and stages of test development, the process should be a collaborative effort, rather than a linear staged process, through which fairness practices are shared among individuals with diverse experiences and expertise (Rodriguez, 2016).

As an example of accessibility-focused test development (or redesign), Wolfe and Gitomer (2001) demonstrated how domains can be made more accessible through improved assessment design and scoring. They focused their work on the National Board for Professional Teaching Standards' certification assessment of accomplished teachers, which includes a portfolio component where candidates provide evidence of accomplished practice, as defined in the standards. To do this, they set out to simply reduce the guessing candidates often engage in, trying to figure out (guess) what will be scored and what makes a difference. By explicitly describing how responses are scored and providing guidance to optimize decisions regarding the construction and content of the portfolio, they brought depth and clarity to the domains being assessed. In addition, they introduced additional structure to the assessment prompts and improved scoring by allowing for more benchmarks and training samples, rater training on bias, and rubric improvements.

Implications for Practice

Test developers and users should employ the principles of UDA and UDL through the PAD approach, perhaps in ways that go beyond surface features of tests and address the core aspects of domain specifications and inclusion of culturally and linguistically relevant disciplinary knowledge and practice (see final section). PAD holds great promise as a test development process that centers accessibility; accessibility features in PAD should be amplified.

Innovations in Item and Task Development

With widespread use of computerized testing and remote or cloud-based delivery, test developers have explored many alternatives for item formats. We use the term technology enhanced (TE) to describe such items, but other terms essentially describe the same notions, including innovative and interactive items. In part, the purposes of TE item development are to create opportunities for test takers to interact with items and tasks to deepen the domain representation, tap difficult-to-measure higher order cognitive processes, represent pedagogically relevant features, include multimedia components expanding the sources of stimulus materials and reduce text dependence, employ more authentic contextualized elements in the test, allow for a wide variety of response modes, and increase test-taker motivation and engagement. These not only are goals of item development, but also become components or assumptions underlying the interpretation of item responses and test performance and introduce questions of accessibility and fairness.

There are too many forms of TE items to create an exhaustive list (see Bennett et al., this volume). However, some TE item features have become more commonplace in large-scale testing, including the use of drag and drop (structured or freeform), hot spots (clickable regions on images and displays or among listed options), figural drawing, ordering and sorting, and the use of a wide range of interactive tools such as balance beam, measuring devices, thermometer, telescope, compass, and even science lab equipment (virtual simulations). Russell and Moncaleano (2019) reported on the prevalence of various item types. They found that approximately 40% of 236 TE items that were released by large-scale K–12 education test programs aligned with a high level of domain fidelity, whereas nearly 40% provided low domain fidelity. This indicated potential challenges of TE items to meet the construct and domain definition clarity required by UDL and PAD.

One tool to support these efforts is the Technology-Enhanced Item Utility Framework (Russell, 2016; Russell & Moncaleano, 2019), designed to examine the fidelity with which TE items represent the intended construct. The framework enables TE item evaluation through structured human judgment regarding usability and accessibility, as well as construct fidelity.

Validity research has lagged behind the advances in item and test delivery innovations, in part because item developers often create tasks and sometimes deploy them without sufficient validation beyond psychometric item analysis. Early investigations of TE items addressed questions of psychometric properties and efficiency (Jodoin, 2003; Zenisky & Sireci, 2002). Some efforts have been successful in identifying TE items that may or may not contribute to domain representation and support test score interpretation. This work includes, as examples, think-aloud investigations of mathematics and English language arts items (Dolan et al., 2011), examination of associations with other measures and predictive evidence for a carpal tunnel release surgery simulation (Shanedling et al., 2010), and the detection of construct-irrelevant variance in a test involving the creation of mathematical expressions (Gallagher et al., 2002).

There are critical fairness and accessibility issues that must be addressed in the design and use of TE items (Stone et al., 2015; Strain-Seymour et al., 2009), such as accessibility for individuals with vision and fine-motor disabilities. Consider, for example, drag-and-drop items—this item type requires visual, fine-motor, and hand–eye coordination skills. Most TE item formats have similar requirements, requiring the test taker to interact with elements presented on screen, thus presenting significant accessibility challenges. A team of researchers in the Accessibility for Technology-Enhanced Assessments Project (Shaftel et al., 2015) spent 3 years investigating questions of accessibility. Employing the principles of evidence-centered design (Mislevy & Haertel, 2006; Huff et al., this volume) and UDA (Thompson et al., 2002, 2004), including UDL for technology-enhanced assessments (Dolan et al., 2013), they engaged a panel of vision and motor skill experts to consider several categories of processing: perceptual (vision, hearing, sensory modalities), linguistic (oral, written, sign, braille), cognitive, motoric, executive (engagement, attention, motivation), and affective (test-taker psychological states and moods). The researchers studied TE items in original and accessible formats, engaged students in cognitive labs, and field tested items in 11 states.

The team found that "item accessibility can be improved for all students on all item types by providing access to assistive technologies" (Shaftel et al., 2015, p. 128). They recommended that (a) instructions should explicitly declare what test takers are expected to do; (b) graphic and font features should be visually simple and large and avoid dependence on colors; and (c) tasks requiring fine-motor skills, such as scrolling, should be minimized. "These studies permit preliminary conclusions about the effectiveness and fairness of alternative item presentations" (Shaftel et al., 2015, p. 165) for students with disabilities.

Researchers addressing the accessibility challenges of TE items illustrate that cognitive complexity is not fixed and depth of knowledge interacts with specific abilities of test takers. Braille and print versions of TE items may not retain the intended cognitive demands. Similarly, as noted earlier, familiarity with item formats and response modes is an essential element for fairness and accessibility, and this may differentially impact English learners and others with limited or interrupted formal education.

Implications for Practice and Research

Test developers should secure collaboration among item writers, accessibility experts, subject matter experts, diversity specialists, psychometricians, and others, since this is essential in the design of TE items. They should engage in rigorous evaluation of item and test quality, with explicit examination of accessibility and fairness, with a focus on the extent to which the expansion of TE items and other innovations obtain responses relative to targeted knowledge and skills. Decisions about item format choice naturally have fairness implications, including standard (non-TE) item formats—test developers choosing among item formats must consider principles of fairness and access (Albano & Rodriguez, 2018; Rodriguez, 2002).

A Sociocultural Framing of Differential Item Functioning

As a standard step toward fairness, in the field-test or pilot stages of item and test development and periodically with operational data, differential item functioning (DIF) is conducted to identify items for bias and sensitivity review. The fundamentals of DIF and the uses of DIF in fairness are described by Zwick (this volume). We also acknowledge that many existing DIF practices are based on test-taker groups with heterogeneous groupings, typically based on racialized and ethnicized status, English learner status, and special education status. Within each of these groups, there may be substantial heterogeneity that masks more important DIF results based on national original, language group or linguistic dialect group, and specific types of disabilities (Ercikan & Oliveri, 2013), as well as the needs described in the section "Test Takers Facing Multiple Access Needs." Such efforts are typically limited because of the smaller, more specific groups of test takers resulting in samples sizes too small to support DIF methods. However, empirical approaches to identifying sources of DIF help protect us against ad hoc interpretations (Alavi & Karami, 2010; Karami & Salmani Nodoushan, 2011).

For a deeper treatment of DIF in the context of intersectionality, see Russell et al. (2022). Intersectionality provides a theoretical framework (Crenshaw, 1989), grounded within Black feminist theory, that addresses the intersections of racialized groups, gender, and other social identities, focused on the social positions within a hierarchy of social power—jointly shaping human experience (Bauer et al., 2021). In addition, Russell (2024) provided a review of DIF procedures in the context of multiple conceptualizations of justice.

Here, we explore the sociocultural context of DIF. To do so, we review three DIF studies that addressed the content and contexts of items, relative to test-taker characteristics. Social models of learning acknowledge that individual learning and cognition occurs though social interaction. Although focused on classroom assessment, Penuel and Shepard (2016) provided a summary of social and sociocultural models they deemed appropriate in the context of assessment. Sociocultural perspectives of learning extend social models into cultural spaces and acknowledge what cultural and linguistic traditions, knowledge, and practices test takers bring to the testing space (Penuel & Shepard, 2016).

An Example in Primary Education

Banks (2006) employed two culturally grounded taxonomies regarding Latino and Black cultures in the United States to identify common elements through which cultural bias could be examined in tests, while acknowledging the within-culture heterogeneity. She identified items from several large-scale tests that exemplified cultural elements. Using the Terra Nova test performance data from Latino, Black, and White fifth-grade students, Banks bundled items based on the presence of culture-specific elements in correct or incorrect options and evaluated for differential bundle functioning (correct-option culture-based item) and differential distractor functioning (incorrect-option culture-based item). Banks reported that a small number of items illustrating each cultural element limited statistical detection power. However, Banks did conclude that DIF was more likely to occur for distractors than for correct options. She also demonstrated how "cultural aspects illustrated in correct and incorrect options" (p. 131) may influence item functioning. She encouraged researchers and test developers to create more comprehensive means for classifying the cultural features of items.

An Example in Secondary Education

Stricker and Emmerich (1999) investigated the Advanced Placement (AP) Psychology examination and potential gender DIF. In their review of prior DIF research, they summarized potential sources of DIF, including familiarity with the item content, including exposure, experience, and cultural loading, as well as test-taker interest in the content and potential emotional reactions to item content. To further study such potential sources, the authors asked 717 students in 19 high school classes to rate items from the AP Psychology exam on familiarity, interest, and unpleasant affect with four-point rating scales (*not at all* to *very*). The standardized mean difference between females and males (*d*) was computed for each rating variable of each item and compared to Mantel–Haenszel values across items. Across the 13 content areas, gender differences in ratings for familiarity and unpleasantness were not significantly different. There were gender differences in interest ratings, where motivation, developmental psychology, and abnormal psychology were of more interest to females and biological psychology and social psychology were of more interest to males. However, these mean *d* effects for familiarity, interest, and unpleasantness were associated with Mantel–Haenszel means across content categories. The more that familiar and interesting items were in a content category to females relative to males, the easier the items were for females (DIF advantaging females); the more that unpleasant items were in a content category, the more difficult the items were for females (DIF disadvantaging females). The authors argued that the gender-based differences found in the AP Psychology items likely developed through differential socialization of females and males at home, at school, and in the community.

An Example in Postsecondary Education

O'Neill et al. (1993) examined the GMAT regarding the potential Black/White and female/male DIF relative to the content and characteristics of the items. Although the proportion of items with moderate to large DIF was low, they found interesting results. Regarding verbal items, they found that reading comprehension items with stimulus materials referring to Black-membered and Latino-membered people (*membered* denoting the socially derived identification of people) and those with social science content (acknowledging overlap in these characteristics) were easier for Black test takers than for the matched group of White test takers, with the opposite being true for general material (material with no reference to specific ethnicized contexts) and items with humanities and science content. Regarding quantitative items, Black test takers performed better on items focused on calculations or formulas, whereas White test takers performed better on word items requiring test takers to set up the problem and translate the words to numerical expressions to solve. In addition, Black test takers performed worse on items with business-related contexts than the matched White test takers. Black test takers also performed worse on items with 50 words or more in the stem/stimulus, whereas Black test takers performed better on items with fewer than 50 words in the stem than the matched White test takers. The authors suggested the possible effects of differences in reading ability and the relevance of item characteristics and the enacted curriculum experienced by test takers. They noted that internal and external advisory committees should review such results to identify implications for test content.

O'Neill and McPeek (1993) acknowledged the relevance of social issues regarding DIF interpretation. They commented on differential opportunities afforded to students of color by less-resourced schools and differential treatment by teachers, as well as lower expectations for students of color, "even when all students are enrolled in the same class. Until we recognize that these education inequities exist, we will not be able to understand DIF results in the proper context" (O'Neill & McPeek, 1993, p. 276).

Implications for Practice and Research

Test developers and users should provide greater attention to the sociocultural context in which tests purport to measure academic achievement to enhance fairness with respect to accessibility, where such achievement is also embedded in sociocultural contexts (see Ercikan & Solano-Flores, this volume). An important function of DIF analyses is gained through the aggregation of lessons learned, lessons that should be turned into item and test development guidance. Test developers should design DIF analyses in ways that address the heterogeneity within groups and the relevant features of items and tests and supplement the typical item-by-item DIF flagging procedures undertaken by most testing programs. They should employ DIF analyses to achieve justice-oriented goals (Russell, 2024).

ACCESSIBILITY THROUGH CULTURAL AND LINGUISTIC RESPONSIVENESS

Given the deepening understanding of the role of sociocultural contexts in teaching and learning, and in testing and assessment (Shepard et al., 2018a), we support the use of culturally and linguistically responsive (CLR) assessment to increase accessibility.

The Role of Cultural and Linguistic Backgrounds

> Accessibility can best be understood by contrasting the knowledge, skills, and abilities that reflect the construct(s) the test is intended to measure with the knowledge, skills, and abilities that are not the target of the test but are required to respond to the test tasks or test items. For some test takers, factors related to individual characteristics such as age, race, ethnicity, socioeconomic status, cultural background, disability, and/or English language proficiency may restrict accessibility and thus interfere with the measurement of the construct(s) of interest. (AERA et al., 2014, p. 52)

The first point is what Ketterlin-Geller (2008) would call differentiating target skills from access skills—although such skills may be interdependent, particularly when the construct of interest is imbedded in real-world scenarios. The second point suggests that our humanity (individual and group characteristics and lived experiences) may restrict accessibility to the extent the construct of interest may reflect the humanity of some and not others.

The authors of the *Standards* offered general examples of how tests may limit access to the construct for some by including idiomatic phrases and regional vocabulary unrelated to the target construct or stimulus contexts unfamiliar to test takers given their cultural background (AERA et al., 2014, pp. 52–53). These test characteristics were not further explained and specific examples were not offered. Most test developers now do consider person characteristics such as, for example, blindness, dyslexia, and limited English proficiency, and do consider the potential barriers facing these test takers. These characteristics are in part the basis for UDA principles to improve accessibility, but they are limited and do not explicitly require test developers to consider the sociocultural contexts of test takers.

A threat to fairness is in test content, vis-à-vis test-taker culture and linguistic histories. As an example, the authors of the *Standards* argued that critical reading

> should not include words and expressions especially associated with particular occupations, disciplines, cultural backgrounds, socioeconomic status, racial/ethnic groups, or geographical locations, so as to maximize the measurement of the construct (the ability to read critically) and to minimize confounding of this measurement with prior knowledge and experience that are likely to advantage, or disadvantage, test takers from particular subgroups. (AERA et al., 2014, p. 54)

The authors continued:

> Differential engagement and motivational value may also be factors in exacerbating construct-irrelevant components of content. Material that is likely to be differentially interesting should be balanced to appeal broadly to the full range of the targeted testing population (except where the interest level is part of the construct being measured). In testing, such balance extends to representation of individuals from a variety of subgroups within the test content itself. For example, applied problems can feature children and families from different racial/ethnic, socioeconomic and language groups. (p. 55)

We contend, as do the scholars and measurement specialists exploring CLR, that such a balance is contradictory with the earlier advice of what to avoid and destroys the value of measures of critical reading, when the readings themselves are based on narrowly defined content domains void of the cultural and linguistic realities of students. How is it possible to "feature children and families from different racial/ethnic, socioeconomic and language groups" (AERA et al., 2014, p. 55) while avoiding "words and expressions especially associated with particular occupations, disciplines, cultural backgrounds, socioeconomic status, racial/ethnic groups" (p. 54), and other contexts that make representation of different groups authentic? Giving extensive guidance about the avoidance of sensitive materials to item writers results in little to no representation of diversity, and when representation is present, it is only in surface features of items, resulting in items and context material with diverse names and places only. The neutral content approach to item and test development is insufficient.

In fact, Randall (2021, 2023) argued that neutralizing items by eliminating any reference or expressions associated with cultural backgrounds or other contexts specific to some test takers places the item within the dominant cultural context, essentially White, middle-class contexts; neutrality is unattainable. Randall (2021) asked test developers to develop a commitment to justice, consider their own positionality, and consider the sociocultural and other identities of the test audience, including the voices of the intended test audience in the test development process.

Finally, the authors of the *Standards* presented opportunity to learn as a relevant context factor, which is "the extent to which individuals have had exposure to instruction or knowledge that affords them the opportunity to learn the content and skills targeted by the test" (AERA et al., 2014, p. 56). Furthermore, disparities in school resources in some settings, including those with large populations of linguistically, racially, and ethnically diverse communities, and possibly rural and isolated communities, affect the quality and content of teaching and learning. Opportunity to learn is a restricted concept because it presumes the target content domain is appropriate and (exhaustively) comprehensive. Does the (specified) content domain include cultural and linguistic disciplinary content knowledge and practices? If we fully adopted the principles of UDL, we would acknowledge many ways of knowing and doing through providing multiple means of engagement, representation, and action and expression in instruction and assessment.

To extend the discussion in the *Standards*, Standard 3.2 requires test developers to minimize the possible effects of "construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics" (AERA et al., 2014, p. 64). Much of the attention to such effects focuses on evidence such as DIF. However, when the entire test presents a uniform barrier, DIF will not uncover the effect on one or more items because the item performance is conditioned on the remaining total test score (which could be uniformly depressed as a result of a narrowly specified content domain). Similarly, a number of standards address the language of the test, relative to the native language or language proficiency of the test takers. Although a test may be administered in a relevant and appropriate language given the test purpose, as required in Standard 3.13, if the content is not relevant and appropriate, significant barriers to access remain. Another common method to address these barriers is external sensitivity reviews of test specifications and items—which typically focus on what items contain that may be problematic, not what items are missing in terms of CLR. Even though nearly every large-scale testing program has sensitivity review processes in place (many employ multiple stages of review), questionable items do find their way to operational tests (Dee & Domingue, 2021).

Content relevance is also addressed in the *Standards* regarding workplace testing and credentialing. An example is given where

> "*salt* is to *pepper*" may be the correct answer to the analogy item "*white* is to *black*" in a culture where people ordinarily use black pepper, but the item would have a different meaning in a culture where white pepper is the norm. (AERA et al., 2014, p. 181)

In the context of developing educational tests, the authors of the *Standards* asserted that

> focus is placed on measuring the knowledge, skills, and abilities of all examinees in the intended population without introducing any advantages or disadvantages because of individual characteristics (e.g., age, culture, disability, gender, language, race/ethnicity) that are irrelevant to the construct the test is intended to measure. (p. 187)

Standard 12.3 requires test developers and users of educational assessments to promote access to the construct for all intended test takers throughout all steps of test design. However, at this point, there is no discussion about the nature of the content reflected by test specifications, other than an acknowledgment that to allow access to test content, context, and response formats, test accommodations and modifications may be needed—again, a response to what is there, not an acknowledgment of what might be missing.

Although the authors of the *Standards* do not employ the language "culturally and linguistically responsive assessment," the notions and elements of such an approach are present throughout. Authors of the "Fairness in Testing" chapter in the *Standards* interpreted fairness as "responsiveness to individual characteristics and testing contexts so that test scores will yield valid interpretations for intended uses" (AERA et al., 2014, p. 50), noting that such a definition exceeds what may be required legally. Similarly, they noted that strategies (accommodations) can be employed with tests and testing procedures "to be responsive to the needs of test takers with disabilities and those with diverse linguistic and cultural backgrounds" (p. 60) as well as other groups. Item and test developers are left with a challenge: to be responsive to the diversity of test takers regarding cultural and linguistic histories, but to avoid culturally and linguistically specific words and expressions and contexts that may not be familiar to some.

CLR Construct and Domain Definitions

Decisions about what to measure and how to measure pervade educational measurement and have done so since its beginning. Such decisions are often made not simultaneously, but sequentially, and not necessarily in a particular order. In education settings, many data collection strategies often default to a test as the tool of choice (how to measure), without substantial regard for the construct or domain of interest (what to measure). Certainly, some constructs and domains are better assessed through other means than a test. Such definitions are often made by policy boards (e.g., state school board or professional associations), without the inclusion of broad representation from the test-taker population or their advocates or measurement specialists.

As described in the test design and development chapter in the *Standards*, the association between the test content and established content standards is critical—that "content specifications must clearly describe the content and/or cognitive categories to be covered so that evidence of the alignment of the test questions to these categories can be gathered" (AERA et al., 2014, p. 75). For most purposes, test specifications include descriptions of how the construct or content domain is represented on the test, which are often refined through the test development process (see Huff et al., this volume). Test specifications are defined broadly, including documentation of purpose and comprehensive details about content decisions. The test content should be fully described in the content domain, content specifications, or content frameworks. These often come from content standards defined by professional associations, education authorities (informed by subject matter experts), and job task or requirements analyses (see Margolis et al., this volume). Content decisions detailed in test specifications should include the sources and characteristics of reading passages, stimulus materials (e.g., illustrations, figures, graphs), practice-based cases, scenarios, vignettes, and contexts related to diversity and local, regional, or national relevance (Rodriguez, 2016).

In education settings, questions have been raised about the appropriateness or fairness of domain specifications and the identification of disciplinary ways of knowing and doing. Some of the clearest messages regarding fairness of construct or domain definitions come from the work of Indigenous educators. Tests should reflect relevant content and modes of understanding and learning to adequately assess what we expect individuals to know and be able to do. Although education researchers have long argued that culture and society play a role in cognitive development (e.g., Vygotsky, 1978), this acknowledgment has been late to influence assessment (for an example in the context of culture and language revitalization and assessment, see Kūkea Shultz & Englert, 2021, 2023). Some reviewers of this chapter argued that CLR principles may be more appropriate for classroom assessment than most large-scale testing programs; however, those engaged in the CLR work seek to influence testing across venues, purposes, and uses (Randall, 2021).

Some assessment researchers have promoted principles of culturally grounded validity evidence (what they called cultural validity) in the area of science assessments. Solano-Flores and Nelson-Barber (2001) argued that what they recommend would require a radical change in current test development and administration procedures, but the added effort may be necessary to achieve greater equity and fairness. They called into question the skills needed to develop fair assessments, such that test development teams should include cultural anthropologists and others who can bring a sociocultural perspective to the team. Specific disciplinary examples include mathematics in cultural contexts (Parker Webster & Yanez, 2007), mathematics in an authentic bicultural context (Lipka et al., 2007), and science assessments acknowledging Aboriginal scientific knowledge and practices (Friesen & Ezeife, 2009).

A broad approach to this challenge is exemplified by the work of Canadian educators to incorporate Aboriginal content and epistemology throughout the education experiences of students (Claypool & Preston, 2011). An important fairness outcome of this work is the assertion that academic assessment must occur in a holistic context, recognizing physical, emotional, and spiritual forms of assessment. In a similar attempt to provide for fair testing of Aboriginal and Torres Strait Islander students in Australia, researchers are studying empirical data from large-scale testing, acknowledging the increasing cultural, ethnic, social, and linguistic diversity of the test-taker population. These researchers recommend greater balance of assessment types and alternative assessment approaches that are inclusive and participatory, particularly in the context of classroom assessment (Klenowski, 2016).

Another approach resulted in the transformation of large-scale assessments, including the Hawaiian state assessment program, supported by statewide collaboratives addressing Indigenous assessment. "Hawaiian culture-based education is based on a holistic view of the world and deep appreciation of interconnectedness" (Nā Lau Lama, 2007, chap. 2, p. 39), acknowledging that relevant knowledge and ways of knowing consider connectedness of worldviews and the assertion that "assessment within a Hawaiian context inherently includes the dimension of spirituality" (p. 38). Two principles of Hawaiian culture-based assessment include

> *E kuahui like i ka hana* (Let everybody pitch in and work together): Assessment is strengths-based, respectful and constructive, looking for the particular attributes, contributions and potentials of the individuals or groups assessed, with particular emphasis on how they contribute to the larger community. Implicit in this approach is a respect for the "funds of knowledge" of students and their communities, an emphasis on growth, and continuous improvement that results from diligent effort. (p. 40)
>
> *Ma ka hana ka 'ike* (In working, one learns): Assessment is personal in that it is appropriate to a particular individual, place and time. There is an emphasis on engagement as well as application of knowledge and skills in authentic ways. (p. 41)

Principles such as these directly impact the conceptualization of constructs and domains. These efforts extend what Moll and colleagues (1992) called funds of knowledge, the "historically accumulated and culturally developed bodies of knowledge and skills essential for household or individual functioning and well-being" (p. 133).

Test and Item Specifications in Support of CLR

Large-scale testing programs have item and task specifications describing sensitivity and bias concerns. Such specifications describe item content that should be avoided because of issues related to representation of diversity in content, illustrations, and contexts, with a focus on elements of stereotyping that should be avoided. Item specifications also declare elements of language to avoid, including controversial or emotionally charged topics. In most cases, this includes avoidance of culturally embedded traditions, holidays, current events, politics, religion, and many other topics that might otherwise bring the test to life and make it relevant. Such sensitivity and bias-free practices not only leave the test context neutral (resolving to White, middle-class contexts), but also thus render it void of relevance to many test takers.

Haladyna and Rodriguez (2013) provided guidance for comprehensive item specifications, which also employ aspects of UDA and are supportive of the goals and components of PAD, beginning with precise descriptions of the construct, domain, knowledge and skills to be assessed, and specification of intended claims and relevant evidence. Item specifications also describe allowable item formats and their structures. Most important, item specifications describe the allowable sources and characteristics of reading passages, characteristics of stimulus materials (illustrations, figures, graphs), and characteristics of practice-based cases, scenarios, and vignettes. Item specifications lose ground regarding CLR in the descriptions (often long lists) of what to avoid, particularly regarding issues related to diversity; local, regional, or global affairs; and most topics that make school subjects, work, and life interesting.

> The process of item development requires collaboration among a number of individuals and groups, including general test developers, item writers who typically are content or subject matter experts, measurement specialists (who may be psychometricians) and relevant specialists in areas such as culture, language development, gender issues and cognitive, emotional/behavioral or physical disabilities. (Rodriguez, 2016, p. 263)

Another potentially powerful contributor to item development is the test taker, where critical insight could be obtained through including student voices in K–12 education assessment development (Roach & Beddow, 2011), as well as student and community voices more generally. This collaborative process is rarely fully realized because item development and review typically occur sequentially, sometimes iteratively, but rarely collaboratively, with multiple specialists reviewing items, just not together. The *Standards* contains many instances of standards articulated with conditions such as "if appropriate" or "if possible." For example, Standard 4.6 states,

> When appropriate to documenting the validity of test score interpretations for intended uses, relevant experts external to the testing program should review the test specifications to evaluate their appropriateness for intended uses of the test scores and fairness for intended test takers. (AERA et al., 2014, p. 87)

We argue that such external review would be appropriate in all cases, particularly in the early stages of test specification development.

Assessment for Justice

There exists another movement in the educational testing and assessment arena that pushes CLR even further. It provides for a grander purpose than responsiveness. Some have long argued that testing and assessment can serve an educational purpose, that individuals learn preparing for the test, taking the test, and reviewing the test (Mehrens & Lehman, 1991). There exist strong examples of tests of, for, and as learning (Bennett, 2015). To push these ideas in a progressive direction, educators are promoting the use of assessment to achieve justice goals—following the direction of efforts to promote social justice–oriented instruction (McLaren, 2000; Papa et al., 2016; Williamson et al., 2007). The introduction of social justice theory and goals in instruction and assessment has as its goal advancing opportunity, particularly for students who have been historically disenfranchised.

In the assessment realm, researchers have illuminated the important connections between assessment, justice, and opportunity (Gardner et al., 2009; McArthur, 2016). In the case of writing assessment, researchers, in collaboration with educators, ask the question, "How can we ensure that writing assessment leads to the advancement of opportunity?" (Poe et al., 2018, p. 379). They argue, and demonstrate, that by disrupting the historic disciplinary isolation of writing, writing assessment can promote student agency and advance opportunities for all students, consistent with the justice orientation of the field of writing and English education. This includes a reorientation of instructional approaches to writing, eliminating barriers to the potential source material and construction of writing exercises, and giving authentic purposes to writing and writing assessments, essentially unconstraining construct representation in writing assessment "to produce evidence related to fairness of comparable consequences for all" (Poe et al., 2018, p. 15).

In a state high school test of writing skills during the early 2000s, the following seemingly neutral and uncontroversial prompt was given to students: If you could change one thing about your life, what would it be and why? Unexpectedly, many students wrote stories about horrific neglect and abuse and illegal behaviors, both as victims and as perpetrators. Readers scoring the essays began to ask whether they were considered mandatory reporters and whether they were required to report cases to authorities—the identities of the students were known because the essay was part of the state accountability testing program. Were these truthful stories or were students writing fiction to display their skills? In addition, the state school counseling association took issue with the potential negative focus of the question because their collective mission was one of positive youth development with a focus on positive assets and opportunities for students. A seemingly neutral writing prompt required mass destruction of the written responses and the provision of counseling for test-response debriefing supports.

In the context of writing, justice prompts abound. Gonchar (2018) offered over 1,000 writing prompts, followed by 300 additional argument writing prompts. Among them were the following:

> Should kids be social media influencers? Should all Americans receive anti-bias education? Should all companies require anti-bias training for employees? Is it offensive for sports teams and their fans to use native American names, imagery, and gestures? Do you support affirmative action in college admissions? What

> rules should apply to transgender athletes when they compete? Why is race so hard to talk about?

Such prompts provide space to display writing skill, argumentation skills, and opportunity to express voice and to do so in a context that allows for multiple perspectives. Even among the GRE analytical writing prompts we find opportunities for students to explore issues of justice, possibly not because ETS is intentionally adopting issues of justice as part of a progressive mission, but because such issues are relevant to adults globally and acknowledge that there are multiple views, experiences, and perspectives. For example, one issue prompt the test employs asks test takers to discuss the statement: "The well-being of a society is enhanced when many of its people question authority" (https://www.ets.org/gre/revised_general/prepare/analytical_writing/issue/pool). An example of an argument prompt is based on a scenario of an anthropologist studying childrearing in an island society that identifies two very different approaches to childrearing when comparing observation-centered versus interview-centered methodologies—calling into question the validity of the interpretations of one set of results (https://www.ets.org/gre/revised_general/prepare/analytical_writing/argument/pool). In both examples, individuals who live in or experience multiple cultural settings will find the prompts relevant and be able to express their voice on issues of justice, perhaps motivating relevant engagement in the measurement procedure.

Although there is not a strong presence in the assessment literature, some are introducing practices targeting antiracist education practices, including instruction and assessment that promote the recognition and analysis of injustice and introduce disciplinary approaches to change (Inoue, 2015, 2019; Kishimoto, 2018). Through antiracist practices in teaching, learning, and assessment, assessment tasks are developed to disrupt racist beliefs in disciplinary knowledge and practices by directly confronting economic, structural, and historical roots of inequality (McGregor, 1993; although McGregor wrote about this in the context of teaching). In the assessment arena, antiracist approaches to assessment ensure that tasks are developed to sustain test-taker cultural and linguistic disciplinary knowledge and practices, rather than ignore or eradicate those ways of knowing and doing (Baker-Bell, 2020). Randall et al. (2023) introduced a justice-oriented antiracist framework for validation. They applied this framework to address construct articulation, data analysis, and score interpretation, uncovering racist logics within standard processes of assessment design. Randall (2023) provided a justice-oriented antiracist perspective on bias and sensitivity review, processes that occur at multiple stages of item development and review, including, for example, following DIF analyses.

Russell (2024) provided a comprehensive review of the role of oppression and racism, as well as White supremacy, to move the field toward antiracist approaches and increase fairness. Russell demonstrated how the White racial frame structures test development theories and practices in ways that perpetuate racialized social structures.

Although these scholars are leading the way, test development, administration, analysis, and use practices remain grounded in prior limited notions and applications of fairness.

Implications for Practice and Research

Test developers need to manage the somewhat conflicting guidance in the *Standards* by revising long-standing item-writing guidelines and sensitivity review procedures to focus not only on what to avoid in item and test content and contexts, but also what to include to embrace CLR. Researchers should develop agendas to evaluate the impact of such approaches. A CLR assessment research agenda should include the application of UDL guidelines that embrace the principles of multiple means in terms of the why, what, and how of learning and doing. Agencies, professional associations, and other entities charged with defining content domains and content standards regarding what people are expected to know and do must engage their constituencies in discussions of essential, necessary, and CLR content. To promote fair tests, from the purpose of testing, test domain specifications, item-writing guidelines, test construction methods, field-test procedures, and operational administration and scoring, the roles of cultural and linguistic disciplinary content knowledge and practices should be considered. We must begin to ask questions about the oppressive and racist frameworks under which we have worked in the past to promote antiracist and fair approaches for a better future to promote social equity, access, and opportunity; testing and assessment more broadly can be an important tool to achieve greater justice.

CONCLUDING REMARKS

Fairness in educational measurement is promoted through consistent commitment to inclusion, access, and rigorous attention to content coverage and relevance and the elimination of barriers to access and sources of construct-irrelevant variance. There are many conditions under which fair and valid score interpretation and use are supported and enhanced, beginning with the way decisions are made about the need and role for testing. A preliminary condition for fair test interpretation and use requires a comprehensive analysis of opportunity to learn, because unfair, discriminatory, or exclusionary practices prohibit fair and uniform score interpretation (AERA et al., 2014, see p. 54). This includes the biases employed in the definition and specification of constructs and content domains, as well as every phase of test development, administration, scoring, reporting, and use. For some test takers, fairness also depends on the use of accommodations because even the most accessible test may not be accessible to all. In addition, fairness regarding access to the construct or domain is about what is *not* included in the content domain as much as it is about what *is*.

"The National Council on Measurement in Education is a community of measurement scientists and practitioners who work together to advance theory and applications of educational measurement to benefit society" (https://www.ncme.org/about/mission). In his NCME presidential address, Sireci (2021) proposed five values, based on scholars and theorists in the field, and consistent with the *Standards*, including the following:

> 1. Everyone is capable of learning. 2. There are no differences in the capacity to learn across groups defined by race, ethnicity, or sex. 3. All educational tests are fallible to some degree. 4. Educational tests can provide valuable information to (a) improve student learning, and (b) certify competence. 5. All uses of educational test scores must be sufficiently justified by validity evidence. (p. 12)

These are fundamental values that undergird fairness in testing. The NCME mission statement calls on us to "serve the common good" (Sireci, 2021), a focus that was elevated during Sireci's leadership of the organization. To promote that goal, he called on the organization to

> (a) ensure we enforce adherence to our AERA et al. (2014) *Standards*, particularly as they relate to the provision of validity evidence to defend test use; (b) de-emphasize norm-referenced competitiveness in educational testing except in those rare instances where examinees actually are competing for a benefit; (c) reorient our practices so that we value students more than the score scale; (d) engage with teachers and other educators to collaboratively develop tests and interpret test scores; (e) reconceptualize our notions of standardization to make tests more flexible to students' needs and funds of knowledge; (f) design test score reports for students that emphasize their strengths, rather than their weaknesses; and (g) take full advantage of technology to allow assessments to tailor themselves to the needs of each specific examinee, foster engagement in the testing process, and to be fully aligned with and integrated into instruction. (p. 14)

These steps offer another way to engage in the principles of accessibility.

Relying on the *Standards* (AERA et al., 2014) and the principles of fairness discussed throughout the chapter, a number of concerns and guidance include broader attention to balanced assessment and policy design (Brookhart et al., 2019; Chattergoon & Marion, 2016; Martineau et al., 2018). The primary focus of balanced assessment efforts requires a careful articulation of the purposes of assessment, followed by the identification or creation and selection of assessments to meet those purposes, where the more important and valued purposes encompass the majority of assessment efforts, reducing emphasis on less essential or informative assessment practices. In addition, no decision resulting in consequences for test takers, groups, or organizations should be made based on a single test score, especially without retake opportunities. A long-standing principle of good measurement practice is to employ multiple measures, which underlies much of the discussion throughout this chapter. Multiple measures capture a richer picture of the intended construct or content knowledge and skills and enhance our understanding of what people know and can do. Consistent with fairness concerns, multiple measures may help to balance intraindividual measurement errors across assessments and allow for multiple ways for test takers to display their knowledge and skills.

Research-informed policies have strong grounding and promote fairness. As such, testing policies must be research informed to provide fairness for test takers from culturally and linguistically diverse backgrounds, as well as those with disabilities. Testing policies and their application will support fairness when they are research based, including policies regarding identification of individuals as English learners and those with disabilities.

A comprehensive test evaluation plan will include a comprehensive review of fairness in every step of the test development and operational stages. Additional focus needs to be given to the evaluation of item and test score functioning vis-à-vis cultural and linguistic characteristics, for example, evaluation of measurement invariance, differential item and distractor functioning, evidence from response processes, differential predictive validity evidence, and the search for unintended consequences among test-taker groups, acknowledging within-group heterogeneity.

A much larger constituency must continue to learn about the principles of testing and test interpretation and use in diverse contexts, particularly test takers with disabilities and English learners who may not be proficient in the test language, including test takers themselves, educators, families, community members, policy makers, teacher educators, workforce development and human resource specialists, and test developers. Test developers and users benefit greatly from reaching out to and engaging with members of diverse communities where tests will be administered and used.

Applications of fairness principles in score interpretation and use continue through the evaluation of potential intended and unintended consequences of testing. Of particular concern in the arena of test fairness is the potential reproduction of disparities and limited opportunities, which are also a function of policies outside the arena of education and employment, including housing, transportation, health, economic development, law enforcement and corrections, and others. The interpretation and use of educational and credentialing tests are not isolated, and public policies interact in unique and unintended ways. The role of consequences continues to be a debated issue (see Ercikan & Solano-Flores, Lane & Marion, and Zwick, all this volume, for additional discussions). Although we agree that the underlying disparities are the result of much larger and historical trends, policies, and social structures, our concern is about the extent to which tests may contribute to the maintenance of social stratification.

Principles of fairness are known and are being implemented and used in many settings. Educational measurement specialists and test developers and users have available to them a wide range of tools and resources to promote fairness in testing through accessibility for specific groups who have experienced challenges in test accessibility and, ultimately, for all test takers.

ACKNOWLEDGMENTS

We are grateful for the attention to detail and wise advice from reviewers throughout the process, including Michael Kane, Maria Araceli Ruiz-Primo, and Michael Russell, a special review by Julio Caesar, and the incredible support and patience of the editors, Linda Cook and Mary Pitoniak.

REFERENCES

Abedi, J., & Hejri, F. (2004). Accommodations for students with limited English proficiency in the National Assessment of Educational Progress. *Applied Measurement in Education, 17*(4), 371–392.

Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research, 74*(1), 1–28.

Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice, 19*(3), 16–26.

Alavi, S. M., & Karami, H. (2010). Differential item functioning and ad hoc interpretations. *Teaching English Language, 4*(1), 1–18.

Albano, A. D., & Rodriguez, M. C. (2018). Item development research and practice. In S.N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of accessible instruction and testing practices* (pp. 181–198). Springer.

Allalouf, A. (2007). Quality control procedures in the scoring, equating, and reporting of test scores. An NCME instructional module. *Educational Measurement: Issues and Practice, 26*(1), 36–46.

Allalouf, A. (2011). *ITC guidelines for quality control in scoring, test analysis, and reporting of test scores*. International Test Commission.

Allalouf, A. (2017). Quality control for scoring tests administered in continuous mode: An NCME instructional module. *Educational Measurement: Issues and Practice, 36*(1), 58–68.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. https://www.testingstandards.net/open-access-files.html

August, D., & Hakuta, K. (1994). *Evaluating the inclusion of L.E.P. students in systemic reform.* U.S. Department of Education, Office of the Under Secretary, Planning and Evaluation Service. https://web.stanford.edu/~hakuta/Publications

August, D., & Hakuta, K. (Eds). (1997). *Improving schooling for language minority children: A research agenda*. National Academies Press.

Baker-Bell, A. (2020). We been knowin: Toward an antiracist language & literacy education. *Journal of Language & Literacy Education, 16*(1), 1–12. https://files.eric.ed.gov/fulltext/EJ1253929.pdf

Banks, K. (2006). A comprehensive framework for evaluating hypotheses about cultural bias in educational testing. *Applied Measurement in Education, 19*(2), 115–132.

Bauer, G. R., Churchill, S. M., Mahendran, M., Walwyn, C., Lizotte, D., & Villa-Rueda, A. A. (2021). Intersectionality in quantitative research: A systematic review of its emergence and applications of theory and methods. *SSM—Population Health, 14*, 1–11. https://doi.org/10.1016/j.ssmph.2021.100798

Beddow, P. A. (2012). Accessibility theory for enhancing the validity of test results for students with special needs. *International Journal of Disability, Development and Education, 59*(1), 97–111.

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education, 39*, 370–407.

Bohrnstedt, G., Kitmitto, S., Ogut, B., Sherman, D., & Chan, D. (2015). *School composition and the Black–White achievement gap* (NCES 2015-018). U.S. Department of Education, National Center for Education Statistics.

Bolt, S., & Roach, A. T. (2009). *Inclusive assessment and accountability: A guide to accommodations for students with diverse needs*. Guilford Press.

Brookhart, S., Stiggens, R., McTighe, J., & Wiliam, D. (2019). *The future of assessment practices: Comprehensive and balanced assessment systems*. Learning Sciences International.

Buzick, H., & Stone, E. (2014). A meta-analysis of research on the read aloud accommodation. *Educational Measurement: Issues and Practice, 33*(3), 17–30.

Caesar, J. (2020). *Standardized bilingual assessments: A means to reduce construct-irrelevant variance and ethnic/racial stereotype threat* [Doctoral dissertation]. University of Minnesota Digital Conservancy. https://conservancy.umn.edu/handle/11299/216165

Cawthon, S., & Leppo, R. (2013). Assessment accommodations on tests of academic achievement for students who are deaf or hard of hearing: A qualitative meta analysis of the research literature. *American Annals of the Deaf, 158*(3), 363–376.

Chattergoon, R., & Marion, S. (2016). Not as easy as it sounds: Designing a balanced assessment system. *State Education Standard, 16*(1), 6–9. https://eric.ed.gov/?id=EJ1087555

Christensen, L. L., Shyyan, V. V., & Stewart, K. (2018). *Developing a request for proposals for an alternate assessment of English language proficiency*. Alternate English Language Assessment. https://altella.wceruw.org/pubs/Developing-Request-for-Proposals.pdf

Claypool, T. R., & Preston, J. P. (2011). Redefining learning and assessment practices impacting Aboriginal students: Considering Aboriginal priorities via Aboriginal and western worldviews. *in education, 17*(3), 84–95.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Department of Health, Education, and Welfare, Office of Education. https://files.eric.ed.gov/fulltext/ED012275.pdf

Collier, V. P. (1989). How long? A synthesis of research on academic achievement in a second language. *TESOL Quarterly, 23*(3), 509–531.

Cottrell, J. M., Newman, D. A., & Roisman, G. I. (2015). Explaining the Black–White gap in cognitive test scores: Toward a theory of adverse impact. *Journal of Applied Psychology, 100*(6), 1713–1736.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum, 1989*(1), 139–167. http://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8

Dee, T. S., & Domingue, B. W. (2021). Assessing the impact of a test question: Evidence from the "Underground Railroad" controversy. *Educational Measurement: Issues and Practice, 40*(2), 81–88. https://doi.org/10.1111/emip.12411

Dolan, R. P., Burling, K., Harms, M., Strain-Seymour, E., Way, W., & Rose, D. H. (2013). *A universal design for learning-based framework for designing accessible technology-enhanced assessments*. Pearson. https://eric.ed.gov/?id=ED576691

Dolan, R. P., Goodman, J., Strain-Seymour, E., Adams, J., & Sethuraman, S. (2011). *Cognitive lab evaluation of innovative items in mathematics and English language arts assessment of elementary, middle, and high school students*. Pearson Education. http://images.pearsonassessments.com/images/tmrs/cognitive_lab_evaluation_of_innovative_items.pdf

Duran, L. K., Wackerle-Hollman, A., Kohlmeier, T. L., Brunner, S. K., Palma, J., & Callard, C. H. (2019). Individual Growth and Development Indicators–Español: Innovations in the development of Spanish oral language general outcome measures. *Early Childhood Research Quarterly, 48*, 155–172.

ETS. (2022). *ETS guidelines for developing fair tests and communications*. https://www.ets.org/content/dam/ets-org/pdfs/about/fair-tests-and-communications.pdf

Eichorn, L. (1997). Reasonable accommodations and awkward compromises: Issues concerning learning disabled students and professional schools in the law context. *Journal of Law and Education, 26* (1), 31–63.

Elliott, S. N., Kettler, R. J., Beddow, P. A., & Kurz, A. (Eds.). (2018). *Handbook of accessible instruction and testing practices: Issues, innovations, and applications* (2nd ed.). Springer.

Ercikan, K., & Oliveri, M. E. (2013). Is fairness research doing justice? A modest proposal for an alternative validation approach in differential item functioning (DIF) investigations. In M. Chatterji (Ed.), *Validity and test use: An international dialogue on educational assessment, accountability, and equity* (pp. 69–86). Emerald Group.

Erickson, K. A., & Geist, L. A. (2016). The profiles of students with significant cognitive disabilities and complex communication needs. *Augmentative and Alternative Communication, 32*(3), 187–197.

Erickson, K., & Quick, N. (2017). The profiles of students with significant cognitive disabilities and known hearing loss. *Journal of Deaf Studies and Deaf Education, 22*(1), 35–48.

Faulkner-Bond, M. (2020). Comparability when assessing English learner students. In A. I. Berman, E. H. Haertel, & J. W. Pellegrino (Eds.), *Comparability of large-scale educational assessments: Issues and recommendations* (pp. 150–175). National Academy of Education. https://naeducation.org/wp-content/uploads/2020/06/Comparability-of-Large-Scale-Educational-Assessments.pdf

Frankenberg, E., Ee, J., Ayscue, J. B., & Orfield, G. (2019). *Harming our common future: America's segregated schools 65 years after Brown*. Center for Education and Civil Rights, University of California, Los Angeles. https://www.civilrightsproject.ucla.edu

Friesen, J. B., & Ezeife, A. N. (2009). Making science assessment culturally valid for Aboriginal students. *Canadian Journal of Native Education, 32*(2), 24–37.

Fuchs, L. S., Fuchs, D., Enton, S. B., Hamlett, C. L., & Karns, K. M. (2000). Supplementing teacher judgements of mathematics test accommodations with objective data sources. *School Psychology Review, 29*(1), 65–85.

Gallagher, A., Bennett, R. E., Cahalan, C., & Rock, D. A. (2002). Validity and fairness in technology-based assessment: Detecting construct-irrelevant variance in an open-ended, computerized mathematics task. *Educational Assessment, 8*(1), 27–41.

Garcia, G. N. (2000). *Lessons from research: What is the length of time it takes limited English proficient students to acquire English and succeed in an all-English classroom?* National Clearing House for Bilingual Education, Issue Brief No. 5. https://files.eric.ed.gov/fulltext/ED450585.pdf

Gardner, J., Holmes, B., & Leitch, R. (2009). *Assessment and social justice: A Futurelab literature review*. Futurelab. https://www.nfer.ac.uk/assessment-and-social-justice/

Garrett Holiday, B. (2009). The history and visions of African American psychology: Multiple pathways to place, space, and authority. *Cultural Diversity and Ethnic Minority Psychology, 15*(4), 317–337.

Geisinger, K. F. (Ed.). (2015). *Psychological testing of Hispanics: Clinical, cultural, and intellectual issues* (2nd ed). American Psychological Association.

Gonchar, M. (2018, April 12). Over 1,000 writing prompts for students [Student opinion piece]. *The New York Times*. https://www.nytimes.com/2018/04/12/learning/over-1000-writing-prompts-for-students.html

Gregg, N., & Nelson, J. (2012). Meta-analysis of the effectiveness of extra time as a test accommodation for transitioning adolescents with learning disabilities: More questions than answers. *Journal of Learning Disabilities, 45*(2), 128–138.

Hakuta, K., Goto Butler, Y., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* The University of California Linguistic Minority Research Institute, Policy Report 2000-1. https://escholarship.org/uc/item/13w7m06g

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

Heaney, K. J., & Pullin, D. C. (1998). Accommodations and flags: Admissions testing and the rights of individuals with disabilities. *Educational Assessment, 5*(2), 71–93.

Inoue, A. B. (2015). *Antiracist writing assessment ecologies: Teaching and assessing writing for a socially just future*. WAC Clearinghouse. https://open.umn.edu/opentextbooks/textbooks/293

Inoue, A. B. (2019). Classroom writing assessment as an antiracist practice: Confronting White supremacy in the judgments of language. *Pedagogy, 19*(3), 373–404.

Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement, 40*(1), 1–15.

Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin, 112,* 527–535.

Kanno, Y., & Cromley, J. G. (2013). English language learners' access to and attainment in postsecondary education. *TESOL Quarterly, 47*(1), 89–121.

Karami, H., & Salmani Nodoushan, M. A. (2011). Differential item functioning (DIF): Current problems and future directions. *International Journal of Language Studies, 5*(3), 133–142.

Kearns, J. F., Towles-Reeves, E., Kleinert, H. L., Kleinert, J. O., & Thomas, M. K.-K. (2011). Characteristics of and implications for students participating in alternate assessment based on alternate academic achievement standards. *The Journal of Special Education, 45*(1), 3–14.

Ketterlin-Geller, L. R. (2008). Testing students with special needs: A model for understanding the interaction between assessment and student characteristics in a universally designed environment. *Educational Measurement: Issues and Practices, 27*(3), 3–16.

Ketterlin-Geller, L. R., Johnstone, C. J., & Thurlow, M. L. (2015). Universal design in assessment. In S. E. Burgstahler (Ed.), *Universal design in higher education: From principles to practice* (2nd ed., pp. 163–175). Harvard Education Press.

Kettler, R. J., Elliott, S. N., & Beddow, P. A. (2009). Modifying achievement test items: A theory-guided and data-based approach for better measurement of what students with disabilities know. *Peabody Journal of Education, 84,* 529–551.

Kettler, R. J., Elliott, S. N., Beddow, P. A., & Kurz, A. (2018). Accessible instruction and testing today. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of accessible instruction and testing practices: Issues, innovations, and applications* (pp. 1–16). Springer International Publishing. https://doi.org/10.1007/978-3-319-71126-3

Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research, 79*(3), 1168–1201.

Kishimoto, K. (2018). Anti-racist pedagogy: From faculty's self-reflection to organizing within and beyond the classroom. *Race Ethnicity and Education, 21*(4), 540–554.

Klenowski, V. (2016). Fairer assessment for Indigenous students: An Australian perspective. In S. Scott, D. Scott, & C. Webber (Eds), *Leadership of assessment, inclusion, and learning* (pp. 273–285). Springer International.

Knoester, M., & Au, W. (2015). Standardized testing and school segregation: Like tinder for fire? *Race Ethnicity and Education, 20*(1), 1–14.

Kopriva, R., Thurlow, M. L., Perie, M., Lazarus, S. S., & Clark, A. (2016). Test takers and the validity of score interpretations. *Educational Psychologist, 5*(1), 108–128.

Kosak-Babuder, M., Kormos, J., Ratajczak, M., & Pizorn, K. (2019). The effect of read-aloud assistance on the text comprehension of dyslexic and non-dyslexic English language learners. *Language Testing, 36*(1), 51–75.

Kūkea Shultz, P., & Englert, K. (2021). Cultural validity as foundational to assessment development: An Indigenous example. *Frontiers in Education, 6,* 1–11.

Kūkea Shultz, P., & Englert, K. (2023). The promise of assessments that advance social justice: An Indigenous example. *Applied Measurement in Education, 36*(3), 255–268.

LaCelle-Peterson, M. W., & Rivera, C. (1994). Is it real for all kids? A framework for equitable assessment policies for English language learners. *Harvard Educational Review, 64*(1), 55–75.

Laing, J., & Farmer, M. (1984). *Use of ACT assessment by examinees with disabilities* (Research Report No. 84). ACT. https://eric.ed.gov/?id=ED249261

Laitusis, C. C., & Cook, L. L. (2007). *Large-scale assessment and accommodations: What works?* (pp. 67–79). Council for Exceptional Children.

Larson, E. D., Thurlow, M. L., Lazarus, S. S., & Liu, K. K. (2020). Paradigm shifts in states' assessment accessibility policies: Addressing challenges in implementation. *Journal of Disability Policy Studies, 30*(4), 244–252. https://doi.org/10.1177/1044207319848071

Lawton, M. (1995, September 6). NAEP to cut exams, test more disabled pupils. *Education Week,* 27.

Lazarus, S., Goldstone, L., Wheeler, T., Paul, J., Prestridge, S., Sharp, T., Hochstetter, A., & Warren, S. (2021). *CCSSO accessibility manual: How to select, administer, and evaluate use of accessibility supports for instruction and assessment of all students*. Council of Chief State School Officers.

Lazarus, S. S., & Thurlow, M. L. (2016). *2015–16 high school assessment accommodations policies: An analysis of ACT, SAT, PARCC, and Smarter Balanced* (NCEO Report 403). National Center on Educational Outcomes, University of Minnesota. https://nceo.umn.edu/docs/OnlinePubs/Report403/NCEOReport403.pdf

Lazarus, S. S., Thurlow, M. L., Lail, K. E., & Christensen, L. (2009). A longitudinal analysis of state accommodations policies: Twelve years of change 1993–2005. *Journal of Special Education, 43*(2), 67–80.

Lazarus, S. S., Thurlow, M. L., Ysseldyke, J. E., & Edwards, L. M. (2015). An analysis of the rise and fall of the AA-MAS policy. *Journal of Special Education, 48*(4), 231–242.

Lazarus, S. S., van den Heuvel, J. R., & Thurlow, M. L. (2017). Using K–12 lessons learned about how to balance accessibility and test security to inform licensure, credentialing, and certification exam policies. *Journal of Applied Testing Technology (JATT), 18*(1), 1–11.

Li, H. (2014). The effects of read-aloud accommodations for students with and without disabilities: A meta-analysis. *Educational Measurement: Issues and Practice, 33*(3), 3–16.

Linquanti, R., Cook, H. G., Bailey, A. L., & McDonald, R. (2013). *Toward a more common definition of English learners: Collected guidance for states and multi-state assessment consortia*. Council of Chief State School Officers.

Lipka, J., Sharp, N., Adams, B., & Sharp, F. (2007). Creating a third space for authentic biculturalism: Examples from math in a cultural context. *Journal of American Indian Education, 46*(3), 94–115.

Little, D. (1999). Learning disabilities, medical students, and common sense. *Academic Medicine, 74,* 622–623.

Liu, K. K., Thurlow, M. L., Press, A. M., & Lickteig, O. (2018). *A review of the literature on measuring English language proficiency progress of English learners with disabilities and English learners* (NCEO Report 408). National Center on Educational Outcomes, University of Minnesota. https://nceo.umn.edu/docs/OnlinePubs/NCEOReport408.pdf

Martineau, J., Dewsbury-White, K., Roeber, E., Vorenkamp, E., Snead, S., & Flukes, J. (2018). *District assessment system design toolkit*. Center for Assessment. https://www.nciea.org/featured-resources

McArthur, J. (2016). Assessment for social justice: The role of assessment in achieving social justice. *Assessment & Evaluation in Higher Education, 41*(7), 967–981.

McDonnell, L. M., & McLaughlin, M. J. (1997). *Educating one & all: Students with disabilities and standards-based reform*. National Academies Press. https://nap.nationalacademies.org/login.php?record_id=5788

McGregor, J. (1993). Effectiveness of role playing and antiracist teaching in reducing student prejudice. *Journal of Educational Research, 86*(4), 215–226.

McLaren, P. (2000). *Che Guevara, Paulo Freire, and the pedagogy of revolution*. Rowman & Littlefield.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). Wadsworth.

Messick, S. (1965). Personality measurement and the ethics of assessment. *American Psychologist, 20,* 136–142.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30,* 955–966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012–1027.

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher, 10,* 9–20.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25,* 6–20. https://doi.org/10.1111/j.1745-3992.2006.00075.x

Moll, L. C., Amanti, C., Neff, D., & Gonzalez, N. (1992). Funds of knowledge for teaching: Using a qualitative approach to connect homes and classrooms. *Theory Into Practice, 31*(2), 132–141.

Moore, J., Huang, C., Nooree, H., Li, T., & Camara, W. (2018). *Testing supports for English learners: A literature review and preliminary ACT research findings* (Working Paper 2018-1). ACT. https://files.eric.ed.gov/fulltext/ED593176.pdf

Nā Lau Lama. (2007). *Nā Lau Lama community report*. Kamehameha Schools. http://www.ksbe.edu/_assets/spi/pdfs/reports/na-lau-lama/Executive_Summary_Final.pdf

National Center for Education Statistics. (1988). *National Education Longitudinal Study of 1988 (NELS:88)*. U.S. Department of Education. https://nces.ed.gov/surveys/nels88/

National Center for Education Statistics. (2019). *Status and trends in the education of racial and ethnic groups (as of February 2019)*. U.S. Department of Education. https://nces.ed.gov/programs/raceindicators/index.asp

National Center for Education Statistics. (2021). *Percentage distribution of public school students, for each racial and ethnic group, by school poverty level: Fall 2017*. U.S. Department of Education. https://nces.ed.gov/fastfacts/display.asp?id=898

National Center for Education Statistics. (2023). *NAEP accommodations increase inclusiveness*. https://nces.ed.gov/nationsreportcard/about/accom_table.aspx

National Center on Educational Outcomes. (2014). *Exploring alternate ELP assessments for ELLs with significant cognitive disabilities* (NCEO Brief 10). University of Minnesota. https://nceo.umn.edu/docs/OnlinePubs/briefs/brief10/NCEOBrief10.pdf

National Center on Educational Outcomes. (2020). *English learners (ELs)*. University of Minnesota. https://nceo.info/student_groups/english_language_learners

National Council on Disability. (2018). *English learners and students from low income families*.

National Council on Measurement in Education. (2020). *Position statement on testing English learners*. https://www.ncme.org/resources-publications/position-statements/equity

Nester, M. A. (1993). Psychometric testing and reasonable accommodations for persons with disabilities. *Rehabilitation Psychology, 38*(2), 75–85.

Oliveri, M. E., & von Davier, A. A. (2016). Psychometrics in support of a valid assessment of linguistic minorities: Implications for the test and sampling designs. *International Journal of Testing, 16*, 220–239.

Olson, J. F., & Goldstein, A. A. (1996). Increasing the inclusion of students with disabilities and limited English proficient students in NAEP. *Focus on NAEP, 2*(1) 1–5. https://files.eric.ed.gov/fulltext/ED400339.pdf

O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255–276). Lawrence Erlbaum Associates.

O'Neill, K. A., McPeek, W. M., & Wild, C. L. (1993). *Differential item functioning on the Graduate Management Admissions Test* (ETS Research Report No. RR-93-35). ETS. https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.1993.tb01546.x

Papa, R., Eadens, D. M., & Eadens, D. W. (Eds.). (2016). *Social justice instruction: Empowerment on the chalkboard*. Springer.

Parker Webster, J., & Yanez, E. (2007). Qanemcikarluni tekitnarqelartuq [One must arrive with a story to tell]: Traditional Alaska native Yup'ik Eskimo stories in a culturally based math curriculum. *Journal of American Indian Education, 46*(3), 116–131.

Penuel, W. R., & Shepard, L. A. (2016). Social models of learning and assessment. In A. A. Rupp & J. P. Leighton (Eds.), *Handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 146–173). John Wiley. https://doi.org/10.1002/9781118956588.ch7

Phillips, S. E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education, 7*(2), 93–120.

Pitoniak, M. J., & Royer, J. M. (2001). Testing accommodations for examines with disabilities: A review of psychometric, legal, and social policy issues. *Review of Educational Research, 71*(1), 53–104. https://doi.org/10.3102/00346543071001053

Poe, M., Inoue, A. B., & Elliot, N. (Eds.). (2018). *Writing assessment, social justice, and the advancement of opportunity*. University Press of Colorado.

Popham, W. J. (2007). Instructional insensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan, 89*(2), 146–150. http://www.pdkmembers.org/members_online/publications/Archive/pdf/k0710pop.pdf

Popham, W. J., & Ryan, J. M. (2012, April 12–16). *Determining a high-stakes test's instructional sensitivity* [Paper presentation]. National Council on Measurement in Education Annual Meeting, Vancouver, BC, Canada.

Quenemoen, R. F. (2009). The long and winding road of alternate assessments: Where we started, where we are now, and the road ahead. In W. D. Schafer & R. W. Lissitz (Eds.), *Alternate assessments based on alternate achievement standards: Policy, practice, and potential* (pp. 127–153). Paul H. Brookes.

Race to the Top Fund, (34 CFR Subtitle B, Chapter II) 74 Fed. Reg. 59687. (November 18, 2009). https://www.govinfo.gov/content/pkg/FR-2009-11-18/pdf/E9-27426.pdf

Randall, J. (2021). "Color-neutral" is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice, 40*(4), 32–34.

Randall, J. (2023). It ain't near 'bout fair: Re-envisioning the bias and sensitivity review process from a justice-oriented antiracist perspective. *Educational Assessment, 28*(2), 68–82.

Randall, J., Poe, M., Oliveri, M. E., & Slomp, D. (2023). Justice-oriented, antiracist validation: Continuing to disrupt White supremacy in assessment practices. *Educational Assessment, 29* (1), 1–20. https://doi.org/10.1080/10627197.2023.2285047

Reardon, S. F., Weathers, E. S., Fahle, E. M., Jang, H., & Kalogrides, D. (2019). *Is separate still unequal? New evidence on school segregation and racial academic achievement gaps* (CEPA Working Paper No.19-06). Stanford Center for Education Policy Analysis, Stanford University. http://cepa.stanford.edu/wp19-06

Rios, J. A., Ihlenfeldt, S. D., & Chavez, C. (2020). Are accommodations for English learners on state accountability assessments evidence-based? A multistudy systematic review and meta-analysis. *Educational Measurement: Issues and Practice, 39*(4), 65–75.

Roach, A. T., & Beddow, P. A. (2011). Including student voices in the design of more inclusive assessments. In S. Elliott, R. Kettler, P. Beddow, & A. Kurz (Eds.), *Handbook of accessible achievement tests for all students* (pp. 243–254). Springer.

Rodriguez, M. C. (2002). Choosing an item format. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 213–231). Lawrence Erlbaum Associates.

Rodriguez, M. C. (2009). Psychometric considerations for alternate assessments based on modified academic achievement standards. *Peabody Journal of Education, 84*(4), 595–602.

Rodriguez, M. C. (2011). Item-writing practice and evidence. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of accessible achievement tests for all students: Bridging the gaps between research, practice, and policy* (pp. 201–216). Springer.

Rodriguez, M. C. (2016). Selected-response item development. In S. Lane, M. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 259–273). Routledge.

Rodriguez, M. C., Kettler, R. J., & Elliott, S. N. (2014). Distractor functioning in modified items for test accessibility. *Sage Open, 4*(4), 1–10.

Rodriguez, M. C., & Nickodem, K. (2018, April 12–16). *Comprehensive partitioning of student achievement variance to inform equitable policy design* [Paper presentation]. National Council on Measurement in Education Annual Meeting, New York, NY, United States. https://conservancy.umn.edu/handle/11299/195229

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

Rogers, C. M., Lazarus, S. S., & Thurlow, M. L. (2016). *A summary of the research on the effects of test accommodations: 2013–2014* (NCEO Report 402). National Center on Educational Outcomes, University of Minnesota. https://nceo.info/Resources/publications/OnlinePubs/Report402/

Rogers, C. M., Lazarus, S. S., Thurlow, M. L., & Liu, K. K. (2020). *A summary of the research on the effects of K–12 test accommodations: 2017* (NCEO Report 418). National Center on Educational Outcomes, University of Minnesota. https://nceo.umn.edu/docs/OnlinePubs/NCEOReport418.pdf

Rogers, C. M., Thurlow, M. L., Lazarus, S. S., & Liu, K. K. (2019). *A summary of the research on effects of test accommodations: 2015–2016* (NCEO Report 412). National Center on Educational Outcomes, University of Minnesota. https://nceo.umn.edu/docs/OnlinePubs/NCEOReport412.pdf

Rose, D. H., Meyer, A., & Hitchcock, C. (2005). *The universally designed classroom: Accessible curriculum and digital technologies*. Harvard Education Press.

Rubel, L., & McCloskey, A. V. (2019). The "soft bigotry of low expectations" and its role in maintaining White supremacy through mathematics education. *Critical Mathematical Inquiry, 41*, 113–128. https://educate.bankstreet.edu/occasional-paper-series/vol2019/iss41/10

Russell, M. (2016). A framework for examining the utility of technology-enhanced items. *Journal of Applied Testing Technology, 17*(1), 20–32. http://www.jattjournal.com/index.php/atp/article/view/89189/67798

Russell, M. (2022). Clarifying the terminology of validity and the investigative stages of validation. *Educational Measurement: Issues and Practice, 41*(2), 25–35.

Russell, M. (2024). *Systemic racism and educational measurement: Confronting injustice in testing, assessment, and beyond*. Routledge.

Russell, M., & Moncaleano, S. (2019). Examining the use and construct fidelity of technology-enhanced items employed by K–12 testing programs. *Educational Assessment, 24*(4), 286–304.

Russell, M., Szendey, O., & Li, Z. (2022). An intersectional approach to DIF: Comparing outcomes across methods. *Educational Assessment, 27*(2), 115–135.

Schaeffer, K. (2020). *6 facts about economic inequality in the U.S*. Pew Research Center. https://www.pewresearch.org/fact-tank/2020/02/07/6-facts-about-economic-inequality-in-the-u-s/

Schifter, L. A., Grindal, T., Schwartz, G., & Hehir, T. (2019). *Students from low-income families and special education*. The Century Foundation. https://tcf.org/content/report/students-low-income-families-special-education/

Semega, J., Kollar, M., Creamer, J., & Moharity, A. (2020). *Income and poverty in the United States: 2018, Current population reports*. U.S. Census Bureau. https://www.census.gov/content/dam/Census/library/publications/2019/demo/p60-266.pdf

Shaftel, J., Benz, S., Boeth, E., Gahm, J., He, D., Loughran, J., Mellen, M., Meyer, E., Minor, E., & Overland, E. (2015). *Accessibility for technology-enhanced assessments: Report of project activities*. Center for Educational Testing and Evaluation, University of Kansas. https://ateassessments.atlas4learning.org/documents

Shanedling, J., Van Heest, A., Rodriguez, M. C., Putnam, M., & Agel, J. (2010). Validation of an online assessment of orthopedic surgery residents' cognitive skills and preparedness for carpal tunnel release surgery. *Journal of Graduate Medical Education, 2*(3), 435–441.

Sharkey, P., & Elwert, F. (2012). The legacy of disadvantage multigenerational neighborhood effects on cognitive ability. *American Journal of Sociology, 116*(6), 1934–1981.

Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018a). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational Measurement: Issues and Practice, 37*(1), 21–34.

Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018b). Classroom assessment principles to support learning and avoid the harms of testing. *Educational Measurement: Issues and Practice, 37*(1), 52–57.

Sireci, S. G. (2021). NCME presidential address 2020: Valuing educational measurement. *Educational Measurement: Issues and Practice, 40*(1), 7–16.

Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75*(4), 457–490.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching, 38*(5), 553–573.

Stone, E., Cahalan Laitusis, C., & Cook, L. L. (2015). Increasing the accessibility of assessments through technology. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 217–234). Routledge.

Strain-Seymour, E., Way, W. D., & Dolan, R. P. (2009). *Strategies and processes for developing innovative items in large-scale assessments*. Pearson Education. https://images.pearsonassessments.com/images/tmrs/StrategiesandProcessesforDeveloping-InnovativeItems.pdf

Stricker, L. J., & Emmerich, W. (1999). Possible determinants of differential item functioning: Familiarity, interest, and emotional reaction. *Journal of Educational Measurement, 36*(4), 347–366.

Thompson, S. J., Johnstone, C. J., Anderson, M. E., & Miller, N. A. (2005). Considerations for the development and review of universally designed assessments. National Center on Educational Outcomes, University of Minnesota. https://rtc3.umn.edu/docs/OnlinePubs/TechReport42.pdf

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). National Center on Educational Outcomes, University of Minnesota. http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html

Thompson, S. J., Thurlow, M. L., & Malouf, D. (2004). Creating better tests for everyone through universally designed assessments. *Journal of Applied Testing Technology, 6*(1), 1–15.

Thurlow, M. L., Elliott, J. L., & Ysseldyke, J. E. (2003). *Testing students with disabilities: Practical strategies for complying with district and state requirements* (2nd ed). Corwin.

Thurlow, M. L., Elliott, J. L., Erickson, R. N., & Ysseldyke, J. E. (1997). Learning disabilities and accommodations: Best practice for bar exams. *The Bar Examiner, 66*(4), 17–30.

Thurlow, M. L., & Kopriva, R. J. (2015). Advancing accessibility and accommodations in content assessments for students with disabilities and English learners. *Review of Research in Education, 39*, 331–369.

Thurlow, M. L., Lazarus, S. S., & Bechard, S. (Eds). (2013). *Lessons learned in federally funded projects that can improve the instruction and assessment of low performing students with* disabilities. National Center on Educational Outcomes, University of Minnesota. https://nceo.umn.edu/docs/OnlinePubs/LessonsLearned.pdf

Thurlow, M. L., Lazarus, S. S., Albus, D. A., Larson, E. D., & Liu, K. K. (2019). *2018–19 participation guidelines and definitions for alternate assessments based on alternate academic achievement standards* (NCEO Report 415). National Center on

Educational Outcomes, University of Minnesota. https://nceo.umn.edu/docs/OnlinePubs/NCEOReport415.pdf

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 400). National Center on Educational Outcomes, University of Minnesota. https://nceo.umn.edu/docs/OnlinePubs/Report400/NCEOReport400.pdf

Thurlow, M. L., Liu, K. K., Ward, J. M., & Christensen, L. L. (2013). *Assessment principles and guidelines for ELLs with disabilities. Improving the validity of assessment results for English language learners with disabilities (IVARED)*. National Center on Educational Outcomes, University of Minnesota. https://nceo.umn.edu/docs/OnlinePubs/ivared/IVAREDPrinciplesReport.pdf

Thurlow, M. L., & Quenemoen, R. F. (2016). Alternate assessments for students with disabilities: Lessons learned from the National Center and State Collaborative. In C. Wells & M. Faulkner-Bond (Eds.), *Educational measurement: From foundations to future* (pp. 416–432). Guilford.

Thurlow, M. L., Shyyan, V. V., Lazarus, S. S., & Christensen, L. L. (2017). *Providing English language development services to English learners with disabilities: Approaches to making exit decisions* (NCEO Report 404). National Center on Educational Outcomes, University of Minnesota. https://nceo.umn.edu/docs/OnlinePubs/NCEOreport404.pdf

Thurlow, M. L., & Wu, Y.-C. (2019). *2016–2017 APR snapshot #20: Students in special education assigned assessment accommodations*. National Center on Educational Outcomes, University of Minnesota. https://ici.umn.edu/products/sLyOfa4fS6yH-ubiml21iQ

Thurlow, M. L., Ysseldyke, J. E., & Silverstein, B. (1993). *Testing accommodations for students with disabilities: A review of the literature* (Synthesis Report 4). National Center on Educational Outcomes, University of Minnesota. https://nceo.umn.edu/docs/OnlinePubs/archive/Synthesis/SynthesisReport004.pdf

Thurlow, M. L., Ysseldyke, J. E., & Silverstein, B. (1995). Testing accommodations for students with disabilities. *Remedial and Special Education, 16* (5), 260–270.

Towles-Reeves, E., Kearns, J., Kleinert, H., & Kleinert, J. (2009). An analysis of the learning characteristics of students taking alternate assessments based on alternate achievement standards. *The Journal of Special Education, 42*(4), 241–254.

U.S. Census Bureau. (2023). *Anniversary of Americans With Disabilities Act: July 26, 2023*. Press Release No. CB23-FF.06. https://www.census.gov/newsroom/facts-for-features/2023/disabilities-act.html

U.S. Department of Education. (2010). *Recent trends in mean scores and characteristics of test-takers on Praxis II licensure tests*. Office of Planning, Evaluation, and Policy Development, Policy and Program Studies Service. https://www2.ed.gov/rschstat/eval/teaching/praxis-ii/report.pdf

U.S. Department of Education. (2015, August 21). Improving the academic achievement of the disadvantaged: Assistance to states for the education of children with disabilities. *Federal Register, 80*(162), 50773–50785.

U.S. Department of Education. (2018). *A state's guide to the U.S. Department of Education's assessment peer review process*.

U.S. Department of Labor. (2021). *Persons with a disability: Labor force characteristics—2020* (USDL-21-0316). Bureau of Labor Statistics. https://www.bls.gov/news.release/pdf/disabl.pdf

U.S. Government Accountability Office. (2011). *Higher education and disability: Improved federal enforcement needed to better protect students' rights to testing accommodations* (GAO-12-40). https://www.gao.gov/assets/590/587367.pdf

Vanchu-Orosco, M. (2012). *Meta-analysis of testing accommodations for students with disabilities: Implications for high-stakes testing* [Doctoral dissertation, University of Denver]. Digital Commons @DU. https://digitalcommons.du.edu/etd/668/

van den Heuvel, J. R., Lazarus, S. S., & Thurlow, M. L. (2016, June 13). Are accessibility and exam security mutually exclusive aims? *Certification Magazine*. http://certmag.com/accessibility-exam-security-mutually-exclusive-aims/

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wackerle-Hollman, A., Durán, L., Brunner, S., Palma, J., Kohlmeier, T., & Rodriguez, M. C. (2019). Developing a measure of Spanish phonological awareness for preschool age children: Spanish Individual Growth and Development Indicators. *Educational Assessment, 24*(1), 33–56.

Wackerle-Hollman, A., Duran, L. K., Rodriguez, M. C., Chavez, C., Miranda, A., & Medina Morales, N. (2022). Understanding how language of instruction impacts early literacy growth for Spanish speaking children. *School Psychology Review, 51*, 406–426.

Wiener, D. (2005). *One state's story: Access and alignment to the GRADE-LEVEL content for students with significant cognitive disabilities* (Synthesis Report 57). National Center on Educational Outcomes, University of Minnesota. https://nceo.umn.edu/docs/OnlinePubs/Synthesis57.pdf

Williamson, J. A., Rhodes, L., & Dunson, M. (2007). A selected history of social justice in education. *Review of Research in Education, 31*, 195–224.

Willingham, W. W., & Cole, N. S. (Eds.). (1997). *Gender and fair assessment*. Routledge.

Willingham, W. W., Rogasta, M., Bennett, R. E., Braun, H., Rock, D. A., & Powers, D. E. (Eds). (1988). *Testing handicapped people*. Allyn & Bacon.

Winter, P. C., Karvonen, M., & Christensen, L. L. (2018). *Developing item templates for alternate assessments of English language proficiency. Alternate English Language Assessment (ALTELLA)*. Wisconsin Center for Education Research, University of Wisconsin–Madison. https://altella.wceruw.org/pubs/ALTELLA_Report_Item_083108.pdf

Wolfe, E. W., & Gitomer, D. H. (2001). The influence of changes in assessment design on the psychometric quality of scores. *Applied Measurement in Education, 14*(1), 91–107.

Wu, Y.-C., Thurlow, M. L., & Liu, K. K. (2021). *Understanding the characteristics of English learners with IEPs to meet their needs during state and districtwide assessments.*

National Center on Educational Outcomes, University of Minnesota. https://files.eric.ed.gov/fulltext/ED616099.pdf

Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*(4), 337–362.

Zieky, M. J. (2016). Developing fair tests. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 81–99). Routledge.