6

A Framework for Enacting Equity Aims in Assessment Use

A Justice-Oriented Approach

Susan Lyons, Maria Elena Oliveri, and Mya Poe

Validity scholars have long called for independent evidentiary arguments to support test score interpretations and justify their use (Kane, 2006). Despite longstanding agreement within the measurement community that each use requires its own justification through the collection of evidence, the world of applied assessment development has either ignored this imperative or limited what counts as evidence. As Russell (2023) points out, the majority of scholars and practitioners in the educational measurement community have been silent on the social justice implications of assessment use and how those implications interact with socio-political realities such as race and racism in our country. As Au (2021) noted, one implication of this silence has been to ignore the effects of test use on curriculum, with negative effects going "directly against multicultural education and the educational experiences of our diverse student population" (p. 105). As measurement professionals, we must embrace a justice-oriented imperative to recognize the systems of oppression operating within our work. The products of educational measurement have the potential not only to reflect systemic inequities but also to perpetuate those inequities.

In this chapter, we offer a framework for justice-oriented assessment use to evaluate the implications of interpreting and using assessments within a culturally and linguistically diverse society. With this framework, we seek to advance a critical lens for assessment development, where outcomes that advance social justice are explicit priorities of testing programs. While many testing programs espouse commitments to equity, frameworks for enacting those commitments are in their nascent stages. Our framework is one attempt to make that commitment explicit.

The framework for justice-oriented assessment use creates an imperative to consider the direct relationships among the consequential impacts of the intended assessment interpretations, uses, and their potential effects on the assessed populations, particularly for historically marginalized learners and communities. The framework applies a critical theory perspective to "identify and locate the ways in which societies produce and preserve specific inequalities through social, cultural, and economic systems" (Martínez-Alemán et al., 2015, p. 8). We emphasize that—when gathering evidence to justify assessment uses—assessment developers and other program stakeholders jointly consider how the intended uses serve to create a more just society or, in contrast, further entrench existing inequities that continue to oppress already marginalized test takers.

DOI: 10.4324/9781003392217-7

Theoretical Grounding

Justice-Oriented Validity Arguments

The preeminent text that defines modern validity theory includes a discussion on the social consequences of testing. As Messick (1989) remarks:

> It will be recalled that the basic validity question was cast not in a logic of indicatives but in a logic of imperatives—"Should the test scores be interpreted and used in the manner proposed?" This was done deliberately as one means of raising consciousness about the ethical and not just the scientific underpinnings of testing and test validation.
>
> (p. 11)

The inclusion of social consequences acknowledges that testing is used to make inferences about groups, and those inferences have real—and often unequal—consequences. Implementing valid score-based inferences means attending to those consequences.

While educational researchers such as Hilliard (1991), Hood (1998), Smitherman (1992, 1993), and Madaus and Clark (2001) long ago pointed to the injustices in testing for minoritized students, recent social changes, and attention to culturally responsive assessment (e.g., Solano-Flores, 2019) in measurement has reanimated conversations about justice-oriented perspectives on assessment. Our work follows Randall's (2023b) definition of justice-oriented assessment in a recent presentation, which she defined as

> An approach to assessment design and development that (a) acknowledges the historical structures of oppression (such as racism, sexism, and colonialism) deeply embedded within our current assessment processes; (b) actively seeks to understand their ongoing consequences on marginalized populations; and (c) intentionally seeks to disrupt these negative processes and outcomes by centering the needs of these populations.

The Justice-Oriented Antiracist Validity (JAV) framework offered by Randall et al. (2022) seeks to reorient the design and use of assessments for creating the greatest benefits for the least advantaged members of society, recognizing the damaging role that assessment has served in reifying and exacerbating existing structures of harm for people of color and other marginalized groups. The JAV framework builds on Kane's (2013) Interpretation Use Argument approach, Oliveri et al.'s (2019) sociocognitive Evidence Centered Design approach, and previous work by Oliveri and colleagues (See "Justice-Oriented Perspectives on Assessment"). In this chapter, we extend the JAV framework to develop a heuristic for gathering evidence and evaluating uses for assessment programs with justice-oriented intentions.

A Framework for Justice-Oriented Assessment Use

Our proposed framework, shown in Table 6.1, consists of five steps: 1) identifying the framework completion team, 2) interpreting individual and group differences in assessment scores, 3) specifying uses, 4) gathering theoretical and empirical evidence to justify use, and 5) determining if use is justifiable.

1. Identifying the Framework Completion Team

The responsibility for the validation of test use typically rests with the test developer. Often an assessment vendor works to determine when and how it will collect evidence to support its assessment claims and intended uses, and then selects evidence to share with the test user in a technical

Table 6.1 Framework for Justice-Oriented Assessment Use

Identifying the Framework Completion Team	*Who comprises the team of people who are completing and discussing this framework?*	
Interpreting Individual and Group Differences in Scores	*What is the intended interpretation of the test scores?* *In what ways might the test scores reflect systemic oppression of marginalized examinees?* *How might we correctly interpret group differences in a multicultural society?*	
Specifying Use	*In what ways are the test scores intended to be used?* *What is the range of possibilities by which test scores might be used for additional, unintended purposes?*	
Gathering Evidence	*How might the intended and/or probable unintended uses of the test scores result in the further entrenchment of existing inequities that harm minoritized people and communities? What evidence supports this theory?*	Theory: Supporting Evidence:
	How might the intended and/or probable unintended uses of the test scores interrupt and reshape systemic factors to advance social justice? What evidence supports this theory?	Theory: Supporting Evidence:
Determination	*Given the interaction between the interpretation, use, and supporting evidence, is the intended use of the test scores justifiable from a social-justice perspective?*	Decision: Yes/No

report. Test users can influence this process by directly requesting additional validity evidence or indirectly by requesting it through a technical advisory committee. This standard operating practice serves to uphold the current power structures, positioning the test vendor as the expert on validity while placing users at a disadvantage, as they often have no way of accessing information about the validity of an assessment product outside of what the vendor chooses to share.

The framework for justice-oriented assessment use directly counters this procedure by requiring that the framework completion team be composed of an interdisciplinary group of stakeholders. The framework team is jointly responsible for each part of the framework—naming test interpretations and uses, collecting and interpreting evidence related to those uses, and ultimately, determining if the uses are justifiable from a social justice perspective. The team should include the test developers, test users, and, importantly, those who are most impacted by the test use—the examinees. The examinees at the table should overrepresent those who have the most potential for suffering the harms of ill-conceived test use, such as examinees of color, examinees with disabilities, and multilingual examinees. Bringing together stakeholders in this dynamic co-creation process creates more transparency in assessment validation, provides a platform for developing assessment literacy among community members, and builds community awareness among test designers who are often far removed from the very communities in which their tests are used. Additionally, the process for completing the framework should be facilitated by someone who is independent of the test vendor to help navigate and account for the potential power differentials and tensions that are likely to arise through the framework completion process.

2. Interpreting Test Scores with a Critical Lens

The psychological and educational sciences have a racist history of misinterpreting the results of standardized tests. For the majority of the 20th century, for example, group differences by

race and class on intelligence tests were viewed by many as scientific evidence of intellectual or cultural inferiority (see Jensen, 1969; Lewis, 1969). Race and ethnicity were falsely interpreted as the cause of group score differences rather than interpreting the group score differences as resulting from the intentional racialization of society that privileges Whiteness (Zuberi, 2000). In their landmark work exploring sociocultural perspectives on race, culture, and learning, Nasir and Hand (2006) offer two related but distinct concepts from the learning sciences and critical theories to help us interpret group score differences by race on educational tests:

1) Cultural and social processes play a central role in the activity systems that underlie human learning.
2) Cultural activity systems occur within sociopolitical contexts that mediate power by race and class.

The first point is now a well-understood tenet of the learning sciences (National Academies of Sciences, Division of Behavioral, Social Sciences, Board on Science Education, Board on Behavioral, Sensory Sciences, … & Practice of Learning, 2018). The sociocultural perspective on learning positions human learning as emerging through primarily social and cultural processes. Assessments are cultural artifacts; what and how we assess are inextricably shaped by the backgrounds, perspectives, and sociocultural identities of those with the power to write standards, design tests, develop items, and score responses (Bennett, 2023). Culture always shapes testing. In other words, culture is not construct irrelevant. As a result, understanding how culture shapes test design and score interpretation is fundamental to measurement. Given this, at a minimum, we should acknowledge explicitly that what is being measured is not an "objective" set of knowledge and skills, but—in the context of the United States—our academic achievements tests are measuring access and proximity to the knowledge, values, language, and ways of understanding of the White[1] dominant class (Randall, 2021).

Our current methods for exploring the relationships between race or ethnicity and the functioning of tests, such as differential item functioning (DIF), are not able to account for the structural embeddedness of culture within our assessments. While studies of intersectional DIF consider the "interaction effects that account for the simultaneous multiplicative influence (e.g., through structures, institutions, and power inequalities derived from those forces) associated with membership in multiple social categories" to offer a "a more contextualized and interpretive evaluation of fairness, equity, and validity," such studies have tended to remain focused on enriching the interpretation of test results, not interrogating tests themselves (Albano et al., 2024, p. 59). Most commonly used methods for detecting DIF rest on the assumption that the assessment itself is culturally neutral and seeks to flag a subset of anomalous items that appear to perform differently by student group.[2] When the entire assessment is a cultural representation of the values, language, priorities, and response processes of the dominant race and class, systematic discrepancies in test performance by race and ethnicity are not flagged as presenting an issue with the instrument.

While there is evidence to suggest that cultural differences can explain some of the group score differences on psychological and educational tests (Arbuthnot, 2020; Fagan & Holland, 2007; Gopaul-McNicol et al., 1999; Helms-Lorenz et al., 2003), we must also understand group score differences in the context of a society that actively values and facilitates the learning of some, while hindering the learning of others—circumstances that bear directly on the test score interpretations. This brings us to the second point we draw from Nasir and Hand (2006): Cultural activity systems occur within sociopolitical contexts that mediate power by race and class. Ecological Systems Theory invites us to consider the multiple ecosystems that influence child development and learning (Bronfenbrenner, 2000). According to this theory, the macrosystem represents the established society in which the child is

developing. In the case of the United States, our society is one where access to resources, healthcare, healthful food, secure housing, quality teaching, and safety is largely stratified by class and race (Feagin, 2013); each of these factors brings its own barriers to student learning and development for minoritized groups.

Given the two related factors discussed in this section, we can interpret systematic group differences by race and culture on psychological and educational tests as representing both:

1) The degree to which our society privileges the dominant class and culture in the content, language, format, and scoring of the assessments themselves; and
2) The systemic limitations and barriers our society places on opportunity and access for marginalized students.

These two factors can be more succinctly understood together as the degree of marginalization in our society. In other words, assuming there are no systematic differences in the capacity to learn (Sireci, 2021), group mean differences in test scores can, therefore, be interpreted as the degree of systematic oppression of a particular group that is othered by the assessments themselves and oppressed by our society at large.

Of course, these group-level systemic factors represent only a portion of the variance in any individual student's score. Each individual score includes variance attributed to the uniqueness of the student and their specific circumstances—i.e., their cultural, linguistic, cognitive, social, and emotional histories. We can understand any individual score as being comprised of three sources of variance (rather than the traditional two sources of variance from classical test theory):

Observed Score = **Individual factors** + **Systemic factors** + random error

In the interpretation of an individual's observed score, there is no way to reasonably disentangle the portion of the score due to individual factors and the portion attributable to systemic factors. Incorrect interpretations of test scores will lead to flawed theories for how to use assessment to advance social justice. When we interpret individual scores as reflecting both individual and systemic factors related to their sociocultural backgrounds, this then has major implications for how we might specify appropriate test score use.

3. Specifying Uses

Impoverished interpretations of test scores that fail to consider the effects of systemic oppression by race and class in our society often lead to assessment uses that are harmful, reifying the existing social strata. Educational assessments often function to reproduce the existing social order by codifying the values, priorities, and ways of being of those in power and enabling policies that continue to stratify society (Dixon-Román & Gergen, 2013). Assessments serve as gatekeepers in our society, perceived by many as an efficient and objective selection mechanism. It is the perceived objectivity of assessments that allows for them to simultaneously perpetuate the public perception of education as a meritocracy while also serving as a form of systemic oppression that disproportionately distributes access to opportunities (Au, 2016; Lyons et al., 2021; Mehan, 2008).

Our framework for justice-oriented assessment use calls on the framework completion team to explicitly state the intended and likely unintended uses of the assessment. Both the intended and any likely unintended uses of an assessment should be evaluated for their potential to advance or hinder social justice for marginalized examinees. For example, a seventh-grade math assessment may be developed to provide inferences about student achievement in the domain and may have the following intended and unintended uses:

- Monitor student progress over time in the domain (intended).
- Evaluate program effectiveness (intended).
- Place students into an Algebra 1 course in the 8th grade (unintended).

In the cases where there is a high probability of possible unintended use, as surfaced through user or market research or known common practices with similar assessment instruments, then it is incumbent on the framework completion team to gather the requisite evidence to justify that particular use from a social justice perspective.

4. Gathering Evidence Related to Assessment Uses

What counts as evidence and how to gather that evidence is at the core of justice-oriented approaches to validity argumentation. The *Standards* identify five sources of validity evidence for supporting the proposed interpretations and uses—content, response processes, internal structure, relationship to other variables, and consequences (AERA et al., 2014).

Randall (2023a) points to the ways that justice-oriented perspectives shift the conversation on content away from fear: "instead of simply warning item writers to handle important topics with care (a fear-based approach) that they be provided real guidelines about how to develop assessment tasks that are antiracist and justice-oriented" (p. 9). If content writers are to depict slavery and genocide accurately, include a wider range of biographical profiles, and include depictions of people typically not represented in test items (e.g., pregnant women), then test designers must rethink what and how they gather validity evidence related to content.

Response processes are a second source of validity evidence. To gather such evidence, researchers have used methods such as cognitive labs, think-aloud protocols, eye-tracking, focus groups, and interviews as validity evidence related to test-takers' response processes. Such data are used to interpret student comprehension of test items and address concerns about cognitive load. Solano-Flores and Trumbull (2003) used one such approach in analyzing test items for multilingual students. Their approach relied on "item microanalysis as the set of reasonings used to examine how the properties of items and students' linguistic, cultural, and socioeconomic backgrounds operate in combination to shape the ways in which students make sense of test items" (p. 4). Their analysis relied on *formal* properties of test design such as item wording, *empirical* properties based on test-takers' response processes, and *differential* properties in how these properties interact "in combination with the students' linguistic and cultural backgrounds to shape their interpretations of items." In another study, Solano-Flores and Li (2009) used cognitive interviews to "probe how [students] related a mathematics item to their personal lives" (p. 9). They found that students from different groups draw "on different sets of contexts of their personal lives to make sense of the item" (p. 9).

Two other types of validity evidence include internal structure and relationship to other variables. The connections between these relationships are only as just as the networks themselves. If the entire network of variables within and without the test is bankrupt, then the mapping project itself is bankrupt. For example, the SAT has long been used as a criterion variable for college-level writing placement testing when it is known that the SAT provides poor predictive validity evidence (Bridgeman et al., 1992; Isaacs & Molloy, 2010; Peckham, 2009). Not only the SAT is a poor predictor of college-level writing success because the kinds of writing tasks required in college are far different than those assessed on the SAT but also the criterion for that prediction—success in college—itself is quite different for historically minoritized students because of the additional cultural barriers they often face in predominantly White institutions (Welton & Martinez, 2014).

Finally, in a review of 283 test reviews in the Mental Measurements Yearbook (MMY), Cizek et al. (2008) looked at "indicators of the validity perspective represented in the 283 test reviews" (p. 403). They found that "validity information is not routinely provided in terms of modern

validity theory, some sources of validity evidence (e.g., consequential) are essentially ignored in validity reports" (p. 397). Because a justice-oriented perspective invites us to think much more expansively about what is meant by evidence, how it is gathered, and how it is used, it also demands careful consideration of consequence as a key source of validity evidence. For example, consequences are often closely associated with immediate concerns. But it need not be. The temporal play of experiential validity evidence is captured in the theory of action models applied to justice-oriented goals. For example, Oliveri et al. (2021) used the theory of action to hypothesize a range of consequences, including positive, negative, foreseeable, and unforeseeable consequences. In another work we are doing, we are exploring the notion of "experiential validity" in tracking the immediate and long-term consequences of placement testing with college students to help us understand the ways pathways are opened or closed based on cut score decisions.

The framework for justice-oriented assessment use requires careful consideration of evidence related to two specific kinds of assessment consequences:

1) How might the intended and/or probable unintended uses of the test scores result in the further entrenchment of existing inequities that harm minoritized people and communities? What evidence supports this theory?
2) How might the intended and/or probable unintended uses of the test scores interrupt and reshape systemic factors to advance social justice? What evidence supports this theory?

Our framework places a central focus on the consequences of test use for either further entrenching existing inequities or interrupting existing oppressive systems to advance social justice. Those using our framework must consider the theory and research-based evidence for both perspectives and, ultimately, use this evidence to support a social-justice-related justification for whether the assessment should be used as intended.

5. Determining Suitability of Proposed Uses for Advancing Social Justice

The final step in completing the framework for justice-oriented assessment use is determining whether the considered uses are justifiable from a social justice perspective. After considering the evidence, the answer may not be a clear-cut yes or no, but instead, the use may be justifiable under certain conditions that should be identified and explained. Rather than aiming to produce answers, we intend for the framework to lead to increased critical consciousness about how educational tests function within a society that is stratified by the effects of racism and other forms of systemic oppression.

Just as the validity of an assessment rests on its interpretation and use, so too does the viability of the assessment from a social justice perspective. We recognize the urgent need to reimagine the design and underlying psychometrics of standardized assessments to value and reflect the knowledge and skills of all students in a multicultural society. In the meantime, we also see the potential utility in the information currently surfaced by traditionally developed psychometric tools when interpreted and used correctly. Deficit interpretations about individuals and communities are not warranted or productive. Instead, aggregate score differences should be treated as a rallying call to action. Persistent score differences by group can and should be leveraged to demand action and funding for creating a more just society.

Applying the Framework to Uses in Local, State, and Workplace Assessment

In this final section, we attempt to apply our framework for justice-oriented assessment use to three distinct but common contexts for assessment: 1) course placement tests in higher education, 2) statewide achievement assessment for public school accountability, and 3) workplace

assessment in an agricultural setting. Tables 6.2–6.4 show how the framework could be applied in each of these situations. Please note that we did not assemble the diverse teams of program stakeholders to consider each aspect of the framework to work through these examples. Therefore, the completed frameworks included in this chapter are provided for illustrative purposes only. Despite this limitation, we believe that the issues surfaced by providing this set of completed examples are helpful for the reader to understand the nature and scope of the critical conversations raised by the framework.

College Writing Course Placement

Placement testing after college admissions is a common practice in the United States (Haswell, 2004). The two subject areas most commonly assessed for placement are math and first-year writing. As explained by Nastal et al. (2022), the rationale for placement hinges on the following argument:

1) Placement testing identifies students with the weakest writing abilities.
2) In order to boost those abilities, placement tests funnel students into specific classes or sections where instruction can be more manageable and students can learn better.
3) Therefore, placement testing leads to improved student learning, retention, and completion (p. 6).

In this closed system, students' cultural contexts are irrelevant because the skills assessed on the placement test are assumed to be universal, *and* the courses in which students are placed are assumed to teach those skills. When it comes to justice-oriented questions about placement testing, the conversation often degenerates into a debate about lowering standards for all students or "student need" with the assumption that university writing instructors are somehow "helping" students acquire those mythical universal writing skills that will help said students succeed in the workplace. But, justice-oriented questions in our framework offer different questions to guide placement testing decision-making.

In our example, we use Accuplacer®. Although many university and college writing programs have moved to local forms of placement testing (e.g., directed self-placement or multiple measures), the Accuplacer writing tests—the Next-Generation Reading and Writing test as well as WritePlacer®—remain popular commercial test options. The Next-Generation Accuplacer test is a multiple-choice exam. WritePlacer includes constructed response tasks—a prompt that yields a 300- to 600-word response. According to the College Board (2021), WritePlacer essays are given a holistic score that is based on six dimensions of writing: Purpose and Focus; Organization and Structure; Development and Support; Sentence Variety and Style; Mechanical Conventions; and Critical Thinking.

As evidenced in Table 6.2, the current paradigm for the use of WritePlacer to support placement into writing courses fails to meet the expectations for social justice. The College Board's Accuplacer "exhibits a narrow conceptualization of writing processes and genres" and "adversely impacts the placement of women and students of color" (Hughes & Li, 2019, p. 71).

While the current iteration of Accuplacer fails to meet social justice demands, placement processes can be reimagined to promote equity. Because placement decisions have a significant impact at the local level, meaningful participation of local stakeholders—such as advocated in the framework described in this chapter—is critical. Participation of local stakeholders may ensure alignment of placement and curriculum, as well as attention to intended and unintended consequences within college writing classrooms. For example, rapid assessment on the first day of class can augment placement decisions and offer an additional layer of evidence to

Table 6.2 Applying the Justice-Oriented Use Framework to ACCUPLACER (WritePlacer®) Course Placement Test

Identifying the Framework Completion Team	*Who comprises the team of people who are completing and discussing this framework?*	Students (particularly students of color, students with disabilities, and multilingual learners), college faculty (including writing teachers), college-university leadership, test developers, and technology support specialists.
Interpreting Scores	*What is the intended interpretation of the test scores?*	Student writing proficiency to determine which students need additional support for college-level literacy. Cut scores may be made based on vendor recommendations or local decisions.
	In what ways might the test scores reflect systemic oppression of marginalized examinees?	• Automated scoring algorithms may not have been trained on a wide range of linguistically and culturally diverse writing samples, resulting in inaccurate score outputs. • Scores that were generated from test prompts that assume narrow Western cultural norms in terms of prompt content and expected generic forms produced by students. • Training guides that rely on deficit-based language to provide rationales for scores and that conflate the ways humans read texts with the ways machine scoring "reads" texts. This conflation further obscures how scoring algorithms were designed.
	How might we correctly interpret group differences in a multicultural society?	Group differences by ethnicity/race and class can be interpreted as a manifestation of the degree of group-level marginalization within society. Individual test scores can be interpreted as representing the degree to which the student has access to the algorithmic representation of dominant race and class linguistic features and generic conventions.
Specifying Use	*In what ways are the test scores intended to be used?*	Student scores are used as the predominant factor for making placement decisions. Placement decisions may result in students taking additional writing courses (including non-credit earning courses).
	What is the range of possibilities by which test scores might be used for additional, unintended purposes?	• Test scores may be used for purposes beyond single course placement into developmental programs or ESL programs. • Scores may influence classroom teaching. • Scores may be used as a pretest measure and paired with an end-of-course post-test measure.
Gathering Evidence	*How might the intended and/or probable unintended uses of the test scores result in the further entrenchment of existing inequities that harm minoritized people and communities? What evidence supports this theory?*	Theory: Given that the placement scores are often closely related to race and language, using them as the only measure to place students into writing courses delays students' persistence and time-to-degree rates. When linked to other programs, such as developmental education and ESL programs, placement testing can result in additional coursework—often coursework that does not carry college graduation credit. Supporting Evidence: • Adverse impact on historically marginalized student populations (Hughes & Li, 2019; Molloy et al., 2020; Nastal et al., 2022). • Poor predictive validity of placement tests (Haswell, 2004; Scott-Clayton, 2012; Scott-Clayton et al., 2014). • Underplacement and misplacement of students (Barnett & Reddy, 2017; Toth, 2018). • Unequal outcomes related to remediation based on student background characteristics (Chen & Simone, 2016).

	How might the intended and/or probable unintended uses of the test scores interrupt and reshape systemic factors to advance social justice? What evidence supports this theory?	Theory: WritePlacer promises to efficiently and accurately place students based on general argument-based writing tasks and automated scoring algorithms. Placement tests may produce more equitable outcomes when used in combination with completion, through-course, and retention data. Supporting Evidence: • Community college cases evidence the justice-oriented potential of multiple measures for placement decisions (Nastal et al., 2022; Poe et al., 2019).
Determination	*Given the interaction between the interpretation, use, and supporting evidence, is the intended use of the test scores justifiable from a social justice perspective?*	Decision: No. While there is evidence to show that WritePlacer is efficient and can produce scores that align with human raters, there is no evidence that suggests its scoring processes result in justice for culturally and linguistically diverse students. Furthermore, because its training guides use language that suggests human ratings of student texts, WritePlacer obfuscates how its scoring algorithms were developed and on what data they were trained.

ensure the appropriateness of placement. Culturally aligned prompts that resonate with the institutional mission may provide another way to align placement processes to the course content that students will encounter, for example, in an HBCU context (Norment, 2021). And critically, any placement process that relies on automated writing evaluation must be transparent about the training materials on which scoring algorithms are based. There can be no justice through occlusion. Regardless of the method (purchased test, directed self-placement, or multiple measures), a justice-oriented use of placement testing demands that students be placed into credit-bearing courses. As Poe et al. (2019) have argued, an admitted student is a qualified student, and no placement process should result in the backsliding of students into zero-credit college courses.

Annual Statewide Testing for School Accountability

Statewide summative assessment programs in K-12 public education are federally mandated through Title 1 of the Elementary and Secondary Education Act (ESEA), most recently reauthorized as the Every Student Succeeds Act in 2015. The law requires that states administer statewide assessments, including annual measurement of student achievement in math and English language arts in grades 3–8 and once in high school, as well as once per grade span in science. The law goes on to require that the scores from those annual achievement tests be used, among other indicators, in a school accountability system. As stated in the law, "the purpose of this title [Title 1] is to provide all children with a significant opportunity to receive a fair, equitable, and high-quality education and to close educational achievement gaps." The intent of the assessment and accountability provisions in the law is to ensure that all schools are providing students with quality teaching and access to rigorous academic content. Civil rights advocates played a critical role in getting this legislation and its predecessor—No Child Left Behind—passed through Congress, which at the turn of the century was largely viewed as a victory for increasing opportunities to learn rigorous content for all learners. While educational equity is at the center of this goal, the way that the policy has played out in practice has failed to

effectively move the needle on addressing the substantial and longstanding educational opportunity gaps for low income students (see, for example, Hanushek et al., 2022).

The framework for justice-oriented assessment use can offer insight into why statewide assessment and accountability programs have failed at serving their intended goals. Additionally, the framework is useful for evaluating the likelihood of success of new policy proposals for the future of state assessment and accountability that are intended to overcome the limitations of the current policy frameworks.

As evidenced in Table 6.3, the current paradigm for the use of statewide assessment programs to support school accountability fails to meet expectations for social justice. While there is some evidence of modest success for the intended theory of action of Title 1 assessment and accountability, the evidence is mixed and far from meeting the originally intended equity aims. At the same time, there is strong evidence that the system has served to reify and further entrench existing injustices and cause harm to marginalized individuals and communities. Therefore, the use of statewide achievement assessment for the current version of ESEA school accountability is not supported from a justice-oriented perspective.

Table 6.3 Applying the Justice-Oriented Use Framework to Statewide Achievement Assessment

Identifying Framework Completion Team	*Who comprises the team of people who are completing and discussing this framework?*	Students (particularly students of color, students with disabilities, and multilingual learners), parents, educators, school leaders, state leaders, assessment developers, policy-makers, civil rights advocates, policy experts, and technical experts.
Interpreting Individual and Group Differences in Scores	*What is the intended interpretation of the test scores?*	Student proficiency level relative to state-adopted content standards.
	In what ways might the test scores reflect systemic oppression of marginalized examinees?	The state content standards and the assessments are cultural artifacts that reflect and represent the values, knowledge, language, and ways of being of the dominant race and class. Those students with greater proximity to the culture of the dominant race and class are more likely to have access to the knowledge, language, and anticipated response processes represented within the assessment. Individual test scores can be interpreted as representing the degree to which the student has access to and mastery of the knowledge, language, and ways of being of the dominant race and class.
	How might we correctly interpret group differences in a multicultural society?	Group differences by race and class can be interpreted as a manifestation of the degree of systemic marginalization within society.
Specifying Use	*In what ways are the test scores intended to be used?*	Student scores are used as the predominant factor for making inferences about school program quality in the state accountability systems. Schools are rank ordered on the basis of test performance, along with other indicators, to identify schools in need of comprehensive or targeted support. Schools that fail to show adequate improvement in student test scores face increasing scrutiny and state-determined consequences. Federal law also requires that individual student reports be provided to parents and educators to guide teaching and learning efforts.
	What is the range of possibilities by which test scores might be used for additional, unintended purposes?	While individual student-level scores are intended to provide criterion-referenced information about student achievement to students, educators, and families, there are some prevalent unintended uses that are problematic from a justice-oriented perspective. These include, but are not limited to, tracking low-performing students into remedial courses that limit opportunities disproportionately for students of color.

Gathering Evidence	*How might the intended and/or probable unintended uses of the test scores result in the further entrenchment of existing inequities that harm minoritized people and communities? What evidence supports this theory?*	Theory: Given that the assessment scores are closely related to race and class, using them as a primary factor to sort schools will result in ranking schools by their Whiteness and wealth. By using test scores as the primary indicator for evaluating school program quality, the state will be reifying systemic inequities through a multitude of mechanisms, including, but not limited to: • Decreased access to local funding and resources for low-performing schools. Local property taxes heavily supplement school funding, and families with access to resources are likely to make property-purchasing decisions, in part, based on perceptions of school quality. • Decreased ability for low-performing schools to attract and retain highly qualified educators. • Increased probability of harmful school closures in communities of color. • Cultural assimilation—the risk of losing important cultural and linguistic aspects of native and other non-dominant cultures. • Low-performing students are tracked into remedial classes focused on test preparation rather than research-based pedagogical practices. Supporting Evidence: • School districts mean scores on academic achievement tests are highly correlated with race and wealth (Reardon, 2016). • Title 1 school accountability increases segregation (Davis et al., 2015). • Effective educators tend to move from low-performing schools to high-performing schools (Boyd et al., 2008). • School closures disproportionately hurt communities of color (Tieken & Auldridge-Reveles, 2019). • Reduction of culturally and linguistically responsive pedagogical practices in the Navajo Nation (Balter & Grossman, 2009). • High-stakes accountability policies influence teachers to make bifurcated curricular and instructional decisions for students within different course "tracks," which are disproportionately stratified by race and class (Watanabe, 2008).
	How might the intended and/or probable unintended uses of the test scores interrupt and reshape systemic factors to advance social justice? What evidence supports this theory?	Theory: The equity-driven promise of test-based school accountability is that by holding all schools to the same high academic achievement standards, schools and educators will provide access to rigorous content for all students—addressing the persistent problem of low expectations for marginalized students. Additionally, providing resources and support to the lowest-performing schools will support program improvement where it is needed the most. Supporting Evidence: • Test-based accountability increases educator expectations for marginalized students (Spence, 2021). • Test-based accountability policy leads to modest improvements in student outcomes for all students, particularly for low-performing students (Loeb & Byun, 2019).
Determination	*Given the interaction between the interpretation, use, and supporting evidence, is the intended use of the test scores justifiable from a social-justice perspective?*	Decision: No. While there is some evidence of modest success for the intended theory of action of Title 1 accountability, the evidence is mixed and far from meeting the originally intended equity aims. At the same time, there is strong evidence that the system has served to reify and further entrench existing injustices and cause harm to marginalized individuals and communities. Therefore, the use of statewide achievement assessment for the current version of ESEA school accountability is not supported from a justice-oriented perspective.

Our assertion regarding ESEA's failure to provide a justice-oriented perspective is not an indictment against data gathering. In fact, Black activists have long pointed to the need to make visible the unjust educational practices and opportunities available to Black students. Making visible those Civil Rights claims is done both through analysis of disparate treatment as well as disparate impact: "[I]n this way, the Act's architects saw discrimination as located not only in individual action but also in institutional and social practices" (Poe & Cogan, 2016, "2.1 The Civil Rights Act of 1964 and Disparate Impact"). In the case of Civil Rights disparate impact analysis, the student does not bear the weight of proving his or her ability, rather, through the burden-shifting approach, the state is responsible for showing how accountability programs may achieve educational goals through alternative means with less of a discriminatory effect on the disproportionately affected group.

A reimagined, justice-oriented use of statewide achievement assessment is possible. As offered in Table 6.3, if we interpret group-level differences by race and class on academic achievement tests as representing the degree of systemic marginalization, then we could use the size of the group differences as a useful metric for monitoring advances toward a more equitable society. School accountability could be reimagined as state and federal accountability, where our state and federal representatives and officials are responsible for creating and carrying out policies and strategic investments that close opportunity gaps.

Creating Career Pathways for the Workplace

In our final use case, we illustrate the application of the framework to the design and development of assessments used to inform career pathways for the workplace with a specific focus on agricultural work. Unlike the other two examples, this use case highlights how an intended assessment use could be determined as justified and likely to advance social justice aims using our framework.

Agricultural workers are often migrant workers coming from different countries, with diverse educational backgrounds and different linguistic assets. Despite knowledge of the job, workers might not be promoted to receive higher pay or more advanced roles if they do not speak the language of most workers (e.g., Spanish or English). Many may have field experience but may not be able to demonstrate their knowledge in monolingual exams in English or Spanish.

Oliveri et al. (2023) and Suárez-Álvarez et al. (2023) outline the needs of adult learners with regard to novel forms of assessment that place adult learners at the forefront of design. Examples of innovations include assessments that use technology such as automated translation and language-select options to expand workers' opportunities to demonstrate field knowledge and expertise, leveraging their funds of knowledge and ways of learning from a culturally sustaining standpoint. Similarly, flexibility in assessment design, item type, and easy-to-use digital interfaces can reduce the barriers associated with the use of technology-based assessments. Technology-based assessments that are carefully designed with accessibility and universal design principles can help support social justice outcomes, increase opportunities for agricultural workers to demonstrate their knowledge, skills, and abilities, and develop career pathways for non-English or Spanish-speaking harvesters. These tests can be used to identify workers with work-relevant skills better to document their knowledge for the purpose of visa status and promotions.

As evidenced in Table 6.4, this example illustrates how tests in the agriculture sector can advance social justice by addressing language and access barriers. Agricultural workers, often migrant and linguistically diverse, face obstacles in demonstrating their knowledge and advancing in their careers due to language proficiency requirements. By embracing accessibility and universal design principles, technology-based assessments can not only identify workers with relevant skills but also support documentation for visa status and promotions, ultimately advancing equity in the agricultural sector.

Table 6.4 Applying the Justice-Oriented Use Framework to A Proposed Multilingual Agricultural Skills Assessment

Identifying the Framework Completion Team	*Who comprises the team of people who are completing and discussing this framework?*	Test developers of the agriculture certification, test purchasers (e.g., agricultural companies that plan to use the assessments), examinees—specifically those who represent the most marginalized groups of farm workers.
Interpreting Individual and Group Differences in Scores	*What is the intended interpretation of the test scores?*	Traditionally, farmers have valued factors such as English-speaking proficiency and residence status, but they have not had a systematic way to identify job-relevant skills. The proposed assessment would be designed to focus on job-relevant skills, regardless of language status, to facilitate promotion paths based on job-relevant skills.
	In what ways might the test scores reflect systemic oppression of marginalized examinees?	The scores for the proposed assessment would prioritize multilingualism and accessibility to counteract current traditional assessment practices that could privilege non-essential farming skills (e.g., knowledge of English) that are not construct-relevant (e.g., different from the needed farming skills to conduct a harvesting job), which could lead to systemic bias in the test for some groups of people (e.g., English versus non-English speakers).
	How might we correctly interpret group differences in a multicultural society?	Group score differences would reflect different levels of agricultural skills relevant to the needed skills to conduct a (harvesting) job relevant to the local area or geographic region of the United States where the job is conducted. Low scores would not be immediately assumed to denote lack of job-related knowledge, but alternative explanations would be carefully studied and examined for possible systematic biases.
Specifying Use	*In what ways are the test scores intended to be used?*	To promote farm workers with high levels of on-the-job skills, regardless of language status. To provide diagnostic, helpful feedback to agricultural workers in non-high-stakes situations—to identify areas of growth and areas that they need to develop/grow.
	What is the range of possibilities by which test scores might be used for additional, unintended purposes?	Scores might be used for high-stakes purposes other than the reasons for which the test was developed (e.g., dismiss workers without the assessed skills or deny workers opportunities to advance their knowledge).
Gathering Evidence	*How might the intended and/or probable unintended uses of the test scores result in the further entrenchment of existing inequities that harm minoritized people and communities? What evidence supports this theory?*	Theory: If the test fails to account for multilingualism and technological accessibility, it could result in further marginalization for groups of workers (e.g., those inexperienced with the test format, lack of knowledge of a particular product/commodity grown in one area of the country that differs from the products/commodities one group is accustomed to growing) specific subgroups can be systematically disadvantaged, receive low scores, and be promoted at different rates than other groups. Supporting Evidence: Research conducted by the Equitable Food Initiative (EFI) shed light on the inequities within the agricultural industry, particularly concerning farmworkers' learning and knowledge. Despite historically oppressive work structures that have traditionally disregarded farmworkers' expertise, it is evident that these structures have not hindered farmworkers from acquiring and applying their Indigenous farming knowledge. However, recognition of this learning by stakeholders across different levels has been lacking (Scully-Russ & Boyle, 2018); that is, the ways in which such knowledge and expertise have been collected neglect to fairly and accurately measure farmworkers' ways of knowing.

(*Continued*)

Table 6.4 Applying the Justice-Oriented Use Framework to A Proposed Multilingual Agricultural Skills Assessment (*Continued*)

	How might the intended and/or probable unintended uses of the test scores interrupt and reshape systemic factors to advance social justice? What evidence supports this theory?	Theory: The use of assessments for learning, the use of universal design for learning, and language select options, among other features, can help increase the utility of scores from workplace assessments (e.g., applied to the agricultural sector) and increase the diversity of the populations taking and using the assessments. Supporting Evidence: Prior literature reviews have been conducted on how technology can support the design of equitable assessments, for instance, funded by the Institute of Education Sciences (Oliveri et al., 2023; Soricone et al., 2024); these approaches have not yet been applied to the design of assessments in the agricultural sector. The research team has engaged in a co-design approach with the EFI organization and Jobs for the Future (JFF) to identify the challenges, hypothesize mechanisms, and seek outcomes from designing assessments that lead to more equitable scores. The supporting evidence is currently based on the use of anticipatory design frameworks (Oliveri et al., 2021) developed by the research team. It is also based on the results of focus groups and interviews conducted with various farms in California. Further research-to-practice connections are needed to continue collecting empirical data to support the utility, efficacy, and effectiveness of the proposed designs.
Determination	*Given the interaction between the interpretation, use, and supporting evidence, is the intended use of the test scores justifiable from a social-justice perspective?*	Decision: Yes, possibly. To date, there are no tests that systematically include the mentioned linguistic and technology-based features in practice. The results are currently inconclusive and experimental. Research suggests that the use of technology with adult learners has the potential to improve accessibility (Gregg, 2012; Rogghe et al., 2018; Rose & Meyer, 2002), particularly when universal design for learning is used at the start of design to render the educational technology being developed to be accessible to as many people as possible, reducing the need for retrofitting through accommodations (CAST, 2018).

Conclusion

As stated succinctly by Solano-Flores and Trumbull (2003), "Culture-free tests cannot be constructed because tests are inevitably cultural devices" (p. 9). A justice-oriented approach to assessment makes recognition of culture, participation by stakeholders, and attention to the potential misuse of tests central to test design. The justice-oriented assessment use framework provides a systematic way to evaluate the appropriateness of the intended uses of both new and existing assessment programs from a social justice perspective. We urge assessment developers, researchers, and applied psychometricians to consider it our responsibility to grapple with the potency of assessment interpretations and uses for both exacerbating existing inequities and advancing social justice. For assessment programs with explicit equity aims, we encourage the use of this framework to anticipate and monitor the impact of assessment use on disrupting inequities and creating a more just future.

Notes

1 By White, we mean the racial category of White while acknowledging that historically this would be White, upper-middle-class, able-bodied, male, heterosexual, and native English speaking.

2 Zumbo and Gelin (2005) depart from this tradition in considering how the sociological and ecological factors within communities might lead to what they term "differential domain functioning."

References

Albano, T., French, B. F., & Vo, T. T. (2024). Traditional vs intersectional DIF analysis: Considerations and a comparison using state testing data. *Applied Measurement in Education*, *37*(1), 57–70. https://doi.org/10.1080/08957347.2024.2311935

Arbuthnot, K. (2020). Reimagining assessments in the postpandemic era: Creating a blueprint for the future. *Educational Measurement, Issues and Practice*, *39*(3), 97.

Au, W. (2016). Meritocracy 2.0: High-stakes, standardized testing as a racial project of neoliberal multiculturalism. *Educational Policy*, *30*(1), 39–62.

Au, W. (2021). Testing for whiteness: How high-stakes, standardized tests promote racism, undercut diversity, and undermine multicultural education. In H. P. Baptiste & J. H. Writer (Eds.), *Evolution of multicultural education: 21st century* (pp. 99–113). Taylor Francisk.

Balter, A., & Grossman, F. D. (2009). The effects of the No Child Left Behind Act on language and culture education in Navajo public schools. *Journal of American Indian Education*, *48*(3), 19–46.

Barnett, E. A., & Reddy, V. (2017). *College placement strategies: Evolving considerations and practices* (CAPR Working Paper). Columbia University.

Bennett, R. E. (2023). Toward a theory of socioculturally responsive assessment. *Educational Assessment*, 1–22.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2008). *Who leaves? Teacher attrition and student achievement* (No. w14022). National Bureau of Economic Research.

Bridgeman, B., Hale, G. A., Lewis, C., Pollack, J., & Wang, M.-M. (1992). Placement validity of a prototype SAT with an essay test. *ETS Research Report Series*, *1992*(i-26). https://doi.org/10.1002/j.2333-8504.1992.tb01459.x

Bronfenbrenner, U. (2000). *Ecological systems theory*. Oxford University Press.

CAST. (2018). *The UDL guidelines* (Version 2.2). https://udlguidelines.cast.org/

Chen, X., & Simone, S. (2016). *Remedial coursetaking at U.S. public 2- and 4-year institutions: Scope, experiences, and outcomes* (NCES 2016–405). U.S. Department of Education. National Center of Education Statistics. https://nces.ed.gov/pubs2016/2016405.pdf

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, *68*(3), 397–412. https://doi.org/10.1177/0013164407310130

College Board. (2021). *WritePlacer®: Guide with sample essays*. Retrieved from: www.collegeboard.org.

Davis, T., Bhatt, R., & Schwarz, K. (2015). School segregation in the era of accountability. *Social Currents*, *2*(3), 239–259.

Dixon-Román, E. J., & Gergen, K. J. (2013). *Epistemology in measurement: Paradigms and practices*. The Gordon Commission.

Fagan, J. F., & Holland, C. R. (2007). Racial equality in intelligence: Predictions from a theory of intelligence as processing. *Intelligence*, *35*(4), 319–334.

Feagin, J. (2013). *Systemic racism: A theory of oppression*. Routledge.

Gopaul-McNicol, S., Reid, G., & Wisdom, C. (1999). The psychoeducational assessment of Ebonics speakers: Issues and challenges. *Journal of Negro Education*, *67*, 16–24.

Gregg, N. (2012). Increasing access to learning for the adult basic education learner with learning disabilities: Evidence-based accommodation research. *Journal of Learning Disabilities*, *45*(1), 47–63.

Hanushek, E. A., Light, J. D., Peterson, P. E., Talpey, L. M., & Woessmann, L. (2022). Long-run trends in the US SES—Achievement gap. *Education Finance and Policy, 17*(4), 608–640.

Haswell, R. (2004). Post-secondary entry writing placement: A brief synopsis of research. *WAC Clearinghouse*.

Helms-Lorenz, M., Van de Vijver, F. J., & Poortinga, Y. H. (2003). Cross-cultural differences in cognitive performance and Spearman's hypothesis: g or c? *Intelligence*, *31*(1), 9–29.

Hilliard, A. G., III. (1991). Introduction by Guest Editor. In *Testing African American students: Special re-issue of the negro educational review* (pp. 8–14). Third World Press.

Hood, S. (1998). Culturally responsive performance-based assessment: Conceptual and psychometric considerations. *The Journal of Negro Education*, *67*(3), 187–196. https://doi.org/10.2307/2668188

Hughes, S., & Li, R. (2019). Affordances and limitations of the ACCUPLACER automated writing placement tool. *Assessing Writing*, *41*, 72–75.

Isaacs, E., & Molloy, S. A. (2010). SATs for writing placement: A critique and counterproposal. *College English*, *72*(5), 518–538. http://www.jstor.org/stable/20749295

Jensen, A. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review*, *39*(1), 1–123.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education/Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Lewis, O. (1969). Review of C. Valentine's, *Culture and poverty: A critique and counter proposals*. *Current Anthropology*, *10*, 189–192.

Loeb, S., & Byun, E. (2019). Testing, accountability, and school improvement. *The ANNALS of the American Academy of Political and Social Science*, *683*(1), 94–109.

Lyons, S., Johnson, M., & Hinds, B. F. (2021). A call to action: Confronting inequity in assessment. *Lyons Assessment Consulting*.

Madaus, G., & Clarke, M. (2001). The adverse impact of high-stakes testing on minority Students: Evidence from one hundred years of test data. In G. Orfield & M. L. Kornhaber (Eds.), *Raising standards or raising barriers? Inequality and high-stakes testing in public education* (pp. 85–106). The Century Foundation Press.

Martínez-Alemán, A. M., Pusser, B., & Bensimon, E. M. (2015). *Critical approaches to the study of higher education: A practical introduction*. JHU Press.

Mehan, H. (2008). A sociological perspective on opportunity to learn and assessment. In P. A. Moss, D. C. Pullin, J. P. Gee, E. H. Haertel & L. J. Young (Eds.), *Assessment, equity, and opportunity to learn* (pp. 42–75). Cambridge University Press.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11.

Molloy, S., Fonville, S., & Salam, A. (2020). "Root and branch": Resisting a basic writing legacy system. *Journal of Basic Writing*, *39*(1), 5–35. https://doi.org/10.37514/JBW-J.2020.39.1.02

Nasir, N. I. S., & Hand, V. M. (2006). Exploring sociocultural perspectives on race, culture, and learning. *Review of Educational Research*, *76*(4), 449–475.

Nastal, J., Poe, M., & Toth, C. (Eds.) (2022). *Writing placement in two-year colleges: The pursuit of equity in postsecondary education*. University Press of Colorado/WAC Clearinghouse. https://doi.org/10.37514/PRA-B.2022.1565

National Academies of Sciences, Division of Behavioral, Social Sciences, Board on Science Education, Board on Behavioral, Sensory Sciences, . . ., & Practice of Learning. (2018). *How people learn II: Learners, contexts, and cultures*. National Academies Press.

Norment, N., Jr. (2021). Some results of using culture-referenced prompts for pre and post-test writing examinations at an HBCU. *CLA Journal*, *64*(2), 252–300. https://doi.org/10.1353/caj.2021.0036.

Oliveri, M. E., Green, A., Rudnyi, A., & Longe, B. (2023, April 14). Equity in learner and test-taker experience. Paper presented at the National Council on Measurement in Education Annual Meeting, Chicago, IL.

Oliveri, M. E., Lawless, R. R., & Mislevy, R. J. (2019). Using evidence-centered design to support the development of culturally and linguistically sensitive collaborative problem-solving assessments. *International Journal of Testing*, *19*(3). https://doi.org/10.1080/15305058.2018.1543308; https://www.tandfonline.com/eprint/kczjtG2hTJDDvKKpIfIz/full?target=10.1080/15305058.2018.1543308

Oliveri. M. E., Slomp, D., Rupp, A. A., & Mislevy, R. J. (2021). Principled development of workplace English communication part 2: Expanded evidence-centered design and theory of action frameworks. *The Journal of Writing Analytics*, *5*, 71–108. https://doi.org/10.37514/JWA-J.2021.5.1.03

Peckham, I. (2009). Online placement in first-year writing. *College Composition and Communication*, *60*(3), 517–540. http://www.jstor.org/stable/20457080

Poe, M., & Cogan, J. A. (2016). Civil rights and writing assessment: Using the disparate impact approach as a fairness methodology to evaluate social impact. *Journal of Writing Assessment*, *9*(1). https://escholarship.org/uc/item/08f1c307

Poe, M., Nastal, J., & Elliot, N. (2019). Reflection. An admitted student is a qualified student: A roadmap for writing placement in the two-year college. *Journal of Writing Assessment*, *12*(1).

Randall, J. (2021). Commentary: From construct to consequences: Extending the notion of social responsibility. *Educational Measurement: Issues and Practice*, *40*(4), 32–34. https://doi.org/10.1111/emip.12452

Randall, J. (2023a). 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. https://doi.org/10.1080/10627197.2023.2223924

Randall, J. (2023b, June 27). Get off my lawn breakout discussion: How do we get to justice? Presented at the National Conference for Student Achievement, New Orleans, LA.

Randall, J., Slomp, D., Poe, M., & Oliveri, M. E. (2022). Disrupting White supremacy in assessment: Toward a justice-oriented, antiracist validity framework. *Educational Assessment*, *27*(2), 170–178.

Reardon, S. F. (2016). *School district socioeconomic status, race, and academic achievement*. Stanford Center for Educational Policy Analysis.

Rogghe, S., Smeets, E., & De Witte, K. (2018). The added value of technology for adults with intellectual disabilities: A literature review. *Journal of Intellectual Disabilities*, *22*(2), 149–162.

Rose, D. H., & Meyer, A. (2002). *Teaching every student in the digital age: Universal design for learning*. Association for Supervision and Curriculum Development.

Russell, M. (2023). Shifting educational measurement from an agent of systemic racism to an anti-racist endeavor. *Applied Measurement in Education*, 1–26.

Scott-Clayton, J. (2012). *Do high-stakes placement exams predict college success?* (Working Paper No. 41). Community College Research Center: Columbia University.

Scott-Clayton, J., Crosta, P. M., & Belfield, C. R. (2014). Improving the targeting of treatment: Evidence from college remediation. *Educational Evaluation and Policy Analysis*, *36*(3), 371–393.

Scully-Russ, E., & Boyle, K. (2018). Sowing the seeds of change: Equitable food initiative through the lens of Vygotsky's cultural–historical development theory. *New Directions for Adult and Continuing Education*, *2018*(159), 37–52.

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

Smitherman, G. (1992). Black English, diverging or converging? The view from the national assessment of educational progress. *Language and Education*, *6*(l), 47–61.

Smitherman, G. (1993). "The Blacker the Berry, the Sweeter the Juice": African American student writers and the national assessment of educational progress. Paper presented at the Annual Meeting of the National Council of Teachers of English, Pittsburgh, PA.

Solano-Flores, G. (2019). Examining cultural responsiveness in large-scale assessment: The matrix of evidence for validity argumentation. *Frontiers in Education*, *4*(43). https://doi.org/10.3389/feduc.2019.00043

Solano-Flores, G., & Li, M. (2009). Generalizability of cognitive interview-based measures across cultural groups. *Educational Measurement: Issues and Practice*, *28*(2), 9–18. https://doi.org/10.1111/j.1745-3992.2009.00143.x

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher*, *32*(2), 3–13.

Soricone, L., Oliveri, M. E., & Atwell, A. (2024). Supporting learner success by keeping a focus on diversity, equity, inclusion, and accessibility. [Breakout session to be presented at COABE 2024]. *COABE 2024 National Conference*, Nashville, TN, United States.

Spence, C. (2021). *Teachers' expectations in the age of accountability* (Doctoral dissertation, State University of New York at Buffalo).

Suárez-Álvarez, J., Oliveri, M., Zenisky, A., & Sireci, S. G. (2023). Current assessment needs in adult education and workforce development: Summary report (Center for Educational Assessment Report No. 998). *Center for Educational Assessment*. Retrieved from https://drive.google.com/file/d/1h8JeVaUk8elZKLIRJmf4EykFrzo7Bn7y/view

Tieken, M. C., & Auldridge-Reveles, T. R. (2019). Rethinking the school closure research: School closure as spatial injustice. *Review of Educational Research*, *89*(6), 917–953.

Toth, C. (2018). Directed self-placement at "democracy's open door": Writing placement and social justice in community colleges. *Writing Assessment, Social Justice, and the Advancement of Opportunity*, 139–172.

Watanabe, M. (2008). Tracking in the era of high-stakes state accountability reform: Case studies of classroom instruction in North Carolina. *Teachers College Record, 110*(3), 489–534.

Welton, A. D., & Martinez, M. A. (2014). Coloring the college pathway: A more culturally responsive approach to college readiness and access for students of color in secondary schools. *The Urban Review*, *46*, 197–223.

Zuberi, T. (2000). Deracializing social statistics: Problems in the quantification of race. *The Annals of the American Academy of Political and Social Science*, *568*(1), 172–185.

Zumbo, B. D., & Gelin, M. N. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological/community moderated (or Mediated) test and item bias. *Journal of Educational Research and Policy Studies*, *5*, 1–2.