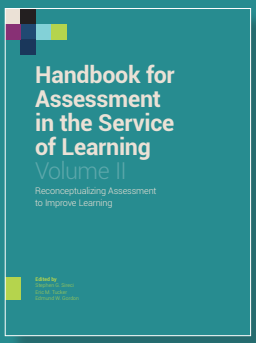
Social Justice in Educational Assessment: A Blueprint for the Future

Stephen G. Sireci, Sergio Araneda, and Kimberly McIntee

UMassAmherst
University Libraries



Series Editors:
Edmund W. Gordon, Stephen G. Sireci, Eleanor Armour-Thomas, Eva L. Baker, Howard T. Everson, and Eric M. Tucker

© 2025 by Stephen G. Sireci, Sergio Araneda, and Kimberly McIntee

The Open Access version of this chapter is licensed under a Creative Commons Attribution–NonCommercial–NoDerivatives 4.0 International License (CC-BY-NC-ND 4.0).

ISBN: 978-1-945764-33-2

Suggested Citation:
Sireci, S. G., Araneda, S., & McIntee, K. (2025). Social justice in educational assessment: A blueprint for the future. In S. G. Sireci, E. M. Tucker, & E. W. Gordon (Eds.), *Handbook for assessment in the service of learning, Volume II: Reconceptualizing assessment to improve learning.* University of Massachusetts Amherst Libraries.

Social Justice in Educational Assessment: A Blueprint for the Future

Stephen G. Sireci, Sergio Araneda, and Kimberly McIntee

This chapter has been made available under a CC BY-NC-ND license.

Abstract

Social justice is a popular topic in education, but it is rarely addressed in educational assessment. Historical attempts to engage the educational assessment community in issues of social justice include several pioneering researchers who pointed out the potential and realized adverse societal effects of testing, as well as the *Standards for Educational and Psychological Testing's* validity evidence based on testing consequences (American Educational Research Association et al., 2014). Drawing from these sources, and from criticisms of the negative effects educational tests have had on individuals and society, we identify several actions that can be taken to promote more socially just testing practices. These actions begin at the earliest stages of test development and extend to test administration, scoring, score reporting, and validation. By considering issues of social justice in test development, we can design assessment systems that better serve education and society, and help end oppressive practices in education, which are central to the goals of assessments that serve learners.

Social Justice in Educational Assessment: A Blueprint for the Future

> *"I don't want no peace. I need equal rights and justice."*
>
> —Peter Tosh

According to Merriam-Webster's online dictionary, the word "just" has several definitions. When used with respect to the social concept of *justice*, the definitions provided are "acting or being in conformity with what is morally upright or good" and "being what is merited." These definitions begin to portray what is meant by the term "social justice," a term first credited to Luigi Taparelli (1840) who defined it as the ability to "maximize individual freedom to associate at all levels" (cited in Boyles et al., 2009; p. 32). This definition was based on Taparelli's belief that smaller communities in society should work together for the common good of larger communities. Today, we may think of Taparelli's vision as a system of "grass roots" efforts; however, the concept of social justice in education has taken on deeper and multifaceted meanings (Porfilio, Strom, & Lupinacci, 2019). In fact, as Boyles et al. pointed out,

> *there are groups promoting education reform in order to perpetuate status quo norms of power and privilege acting in the name of social justice. Yet, and at the same time, there are other groups who wish to dismantle such a privilege under the auspices of social justice.* (p. 30)

For most educators and educational researchers, social justice entails pursuing equity and access to high-quality education for all students regardless of race, sexual orientation, religion, age, and other sociopolitical or sociological characteristics. Although many researchers argue the concept of "social justice" in education cannot be defined in one way (Jean-Marie, Normore, & Brooks, 2009), in discussing how to teach about social justice in education, Bell (1997) characterized social justice education as both a goal and a process. As she described, "The goal of social justice education is full and equal participation of all groups in a society that is mutually shaped to meet their needs…[while]… the process for attaining the goal of social justice… should be democratic and participatory, inclusive and affirming of human agency and human capacities for working collaboratively to create change" (pp. 3–4). In addition to participation as both a goal and a process, common themes across different conceptualizations of social justice in education

include confronting issues of racism, sexism, homophobia, xenophobia, and other manifestations of oppression to create and maintain equitable schooling so all students (broadly defined) can achieve their educational and occupational goals (e.g., Boyles et al., 2009; Jean-Marie et al., 2009; Porfilio et al., 2019).

Although considerable attention has been devoted to social justice in education, the concept has received little attention in the educational testing community. This lack of attention is unfortunate, in that justice is a critical component for educational assessments to serve learners. Our chapter uses themes of social justice to advance key principles of assessment in the service of learning—namely, making assessment processes transparent and fair (Principles 1 and 6) and ensuring high-quality, credible uses of assessments (Principle 7, see Baker et al., 2025). In this chapter, we argue social justice in educational assessment should be front-of-mind for the entire measurement community; and by properly addressing issues of social justice, we can eliminate the contribution of educational tests to systemically racist and unjust educational practices.

Educational tests are an integral and enduring part of contemporary society. They have enormous consequences that may affect the quality of education an individual may receive and the degree to which individuals are able to reach their aspirations. Although there have been criticisms against the widespread use of tests (e.g., Bertrand & Marsh, 2021a, 2021b; Koljatic et al., 2021a), and some measurement scholars have discussed the importance of values and consequences in testing (e.g., Messick, 1989; Shepard, 1993; Sireci, 2021), there has not yet been a comprehensive discussion of social justice in educational assessment. Thus, the purpose of this chapter is to end this deficit in the psychometric literature by (a) tracing the history of concerns for social justice in educational assessment, and (b) proposing actions for more socially just educational assessment practices.

To accomplish these purposes, we first review the guidance provided by professional testing standards. Next, we review seminal literature written by educational measurement researchers and practitioners who addressed issues of social justice. Drawing from our review, we recommend several steps to promote testing practices that are consistent with and useful for an educational system with social justice at its core.

Concerns for Social Justice in Educational Assessment: A Brief History

> *It was not until I was long out of school and indeed after the (first) World War that there came the hurried use of the new technique of psychological tests, which were quickly adjusted so as to put black folk absolutely beyond the possibility of civilization (W.E.B. DuBois, 1940, as quoted in Guthrie,* 1998, p. 55).

As Sireci and Randall (2021) described, the history of contemporary educational testing is often traced back to the early work of Alfred Binet, who developed tests to identify children in Paris in need of special education. Although this purpose was laudable and led to the inclusion of children in schools who otherwise would have been denied an education, there is also a darker side to the history of educational testing. It is this darker side that is reflected in W.E.B. DuBois's quote at the beginning of this section. In the early 20th century, Binet's assessment techniques were transported to the United States, where they fueled an Eugenics movement that tracked minoritized students to less challenging educational experiences, and privileged white students to more challenging and rewarding experiences. This practice continues to this day, as evidenced by the recent lawsuit in New York City (IntegrateNYC vs. State of NY, 2021). Before describing present day concerns, we will first describe the work of those who sounded the alarm long ago, beginning with the development of professional guidelines for the testing profession.

Standards for Educational and Psychological Testing

In 1952, the American Psychological Association (APA) released *Technical Recommendations for Psychological Tests and Diagnostic Techniques: A Preliminary Proposal* (APA, 1952). This document was the first attempt at promoting professional guidelines on test development, use, and evaluation. Two other organizations joined with APA to transform the proposal into the first formal version of standards for the testing industry: the American Educational Research Association (AERA), and the National Council on Measurements Used in Education (which dropped "used" and became NCME in 1961). The first product of this joint effort was *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (APA, 1954). At the time of this writing (July 2025), there have been six versions of these *Technical Recommendations*, the most recent of which are the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014; see also AERA et al., 2018 for the Spanish-language version of the

Standards). Despite their efforts to promote more socially just uses of tests, none of these versions can be cited as a helpful treatise on social justice in educational assessment. Nevertheless, we include this work because a review of the evolution of the content of these *Standards* over the past seven decades illustrates increasing concern for the effects of tests on groups of test takers and society.

In the first version of what we today call the *Standards* (APA et al., 1954), it is difficult to find content related to social justice. The words "fairness," "race," and "adverse impact" do not even appear in the document, and "bias" only appears in relation to sampling and statistics. However, these early *Standards* did take a stand against Eugenics by pointing out that using intelligence test to infer innate ability was a misconception. As they put it,

> *another common misconception is that intelligence tests are measures of inherent native ability alone. It would be desirable for manuals of such tests to caution against this interpretation.* (APA et al., 1954, p. 11)

Today, this acknowledgement of this "misconception" seems vacuous, or at best tepid. However, two versions later, APA, AERA, and NCME (1974) explicitly addressed issues of unfairness in test use, and the acknowledgement of such issues required revision of the *Standards.* In describing reasons for the 1974 revision, they claimed,

> *part of the stimulus for revision is an awakened concern about problems like an invasion of privacy or discrimination against members of groups such as minorities or women. Serious misuses of tests include, for example, labeling Spanish-speaking children as mentally retarded on the basis of scores on tests standardized on "a representative sample of American children," or using a test with a major loading on verbal comprehension without appropriate validation in an attempt to screen out large numbers of Blacks from manipulative jobs requiring minimal verbal communication.* (APA, 1974, p. 1)

Thus, by the mid-1970s, issues of social justice in assessment were formally being confronted by the assessment community. With respect to subsequent versions of the *Standards*, the next version (APA, AERA, & NCME, 1985) included two chapters that addressed concerns for "linguistic minorities" and "people who have handicapping conditions," which evolved into a separate chapter on "Fairness

in Testing and Test Use" in the subsequent version (AERA, APA, & NCME, 1999). The 1999 version couched social justice concerns under the rubric of fairness and acknowledged the complexity of "fairness" by stating,

> *concern for fairness in testing is pervasive, and the treatment according to the topic here cannot do justice to the complex issues involved…The Standards cannot hope to deal adequately with all these broad issues, some of which have occasioned sharp disagreement among specialists and other thoughtful observers.* (p. 73)

The current version of the *Standards* (AERA et al., 2014) retained the "fairness in testing" chapter, and described four perspectives of fairness: (a) in treatment during the testing process, (b) as lack of measurement bias, (c) in access to the construct(s) measured, and (d) as validity of individual test score interpretations for the intended uses. Perhaps the biggest fairness issue related to social justice addressed by both the 1999 and 2014 versions of the *Standards* was adverse impact, meaning differential outcomes for subgroups of students based on test scores (e.g., when tests are used for selection into jobs, schools, or competitive programs). In most situations where tests are used for "high-stakes" purposes, African American, Hispanic/Latino, Native American, and other historically minoritized groups have substantially lower acceptance rates, which has led to truncated participation of individuals from these groups in the rewards associated with higher test scores. For this reason, the use of test scores to make such awards has been criticized as supporting systemic racism in education (Hobson, Szostek, & Griffin, 2021; IntegrateNYC vs. State of NY, 2021; Sireci, 2021).

> The AERA et al. (2014) *Standards* addressed this issue by claiming,
>
> *…the* Standards' *measurement perspective explicitly excludes one common view of fairness in public discourse: fairness as the equality of testing outcomes for relevant test-taker subgroups. Certainly, most testing professionals agree that group differences in testing outcomes should trigger heightened scrutiny for possible sources of test bias…However, group differences in outcomes do not in themselves indicate that a testing application is biased or unfair.* (p. 54)

The logic underlying this view is, observed mean test score differences across groups defined by racial or other demographic variables may reflect true

differences across groups, rather than imperfections of the measurement properties of the test. However, Helms (2006) pointed out when test scores reflect such differences without explaining the reasons for them in a construct-relevant manner, evidence of *invalidity* persists, and such evidence argues against test use. In a subsequent section, we provide a recommendation for AERA, APA, and NCME for revising this position to emphasize the pursuit of more socially just measurement practices. In our view, the preceding excerpt from the *Standards* sidesteps an important social justice issue that deserves more discourse in the assessment community.

In addition to issues of fairness, the concept of test validity also overlaps with issues of social justice in assessment. As the evolution of the *Standards* indicates, validity evidence can focus on data related to issues of social justice. In fact, in the last two versions of the *Standards* (AERA et al. 1999; 2014), an explicit source of validity evidence based on testing consequences was introduced to expand the conceptualization of validity and validation. Including the consequences of testing as an essential source of validity evidence opened the door to further consideration of social justice issues in assessment.

Sources of validity evidence

It is important to note that the concept of validity and test validation evolved over the six versions of the *Standards*, and the past two versions (AERA et al., 1999; 2014) specified five sources of validity evidence that can be used to evaluate the use of a test for a particular purpose. Three of these sources of validity evidence were discussed in some form since the first version (APA, 1954): validity evidence based on test content, internal structure, and relations to other variables. A fourth source of validity evidence, based on response processes, refers to confirming the cognitive processes intended to be measured are in fact being invoked by the test items. It is the fifth source, however, *validity evidence based on testing consequences*, that overlaps with social justice issues. Ideally, this source of validity evidence should require testing agencies and other test users to provide evidence that the intended consequences of a testing program are being realized, and that (un)intended negative consequences, to individuals or to society, based on the use of test scores are identified and eliminated, or at least minimized. However, the *Standards* fall short of that ideal by requiring only that construct-irrelevant sources of bias be investigated as threats to fairness.

Although concerns over the consequences of testing have been codified into the *Standards* for over 22 years, some researchers have rejected the requirement that consideration of consequences is part of a reasonable validation effort or pertains to validity at all (See Koljatic et al. 2021a, 2021b; Sireci, 2016a, 2016b for discussions). Such a position is at odds with incorporating a social justice perspective into educational assessment, which is why we remain grateful the AERA et al. (1999, 2014) *Standards* rejected such dismissiveness. However, even given the discussions of testing consequences and fairness, the *Standards* remain incomplete with respect to evaluating and providing guidance on social justice issues in educational assessment. Several prominent measurement specialists paid closer attention to these issues and encouraged the profession to do so. It is to these scholars we turn next.

Pioneers of Social Justice in Educational Assessment

We did not complete an exhaustive search of the educational testing literature to identify all discussions of social justice in assessment; however, as students of validity theory and as practitioners who conduct validity studies, several theorists stand out as pioneers in this area. These pioneers include, but are not limited to, Robert Ebel, Samuel Messick, Sylvia Johnson, and Janet Helms. In this section, we provide brief descriptions of some of their writings in this area.

Robert Ebel

In 1963, the Educational Testing Service hosted an *Invitational Conference on Testing Problems* with the closing keynote featuring Robert Ebel, who titled his address "The Social Consequences of Educational Testing" (Ebel, 1963). In this address, Ebel pointed out the testing community was being criticized for "having shown lack of proper concern for the social consequences of our educational testing," and he argued "what testing needs most is a large program of research on its social consequences" (p.131). We agree with his suggestion from 58 years ago and point out it has yet to be acted upon!

Ebel (1963) specified four harmful potential consequences of educational tests: (a) predetermining the status of children based on how they are labeled by test performance, (b) supporting a narrow definition of ability that will "reduce the diversity of talent available to society," (c) placing too much control over education in the hands of the testing industry, and (d) promoting "mechanistic decision

making" (pp. 132–133). Similar to APA et al. (1954), he forcefully argued against using test scores as measures of innate intelligence and claimed "One of the important things test specialists can do to improve the social consequences of educational testing is to discredit the popular conception of the [intelligence quotient]" (p. 135). He also encouraged the use of tests for positive consequences for individuals and society by suggesting, "We should judge the value of the tests we use not in terms of how accurately they enable us to *predict* later achievement, but rather in terms of how much help they give us to *increase* achievement by motivating the efforts of students and teachers" (p. 136). Clearly, Ebel was a precursor to many of the same arguments being made today that educational tests should do more to educate our children (Gordon, 2020; Sireci, 2021).

Ebel (1963) argued if we ignored the criticisms of educational tests, the costs to society would be dire. As he described, "If we ignore them and undertake to manage the lives of others so that those others will qualify as worthy citizens in our own particular vision of utopia, we do justify the concern that one harmful social consequence of educational testing may be mechanistic decision making and the loss of essential human freedoms" (p. 141). In other words, requiring students to answer assessment items in a particular way may overlook the creativity, individuality, and consciousness students possess and bring to the assessment. Ebel's point is that by continuing testing practices in a mechanistic, unchecked manner, we may foster mechanistic education in a way that stifles the diversity of ideas across educators and students. Recognizing that ideas regarding educational constructs differ across race and culture (Malda, van der Vijver, & Tamane, 2010; Randall, 2021), Ebel's call for addressing the criticisms of testing was a call for expanding the worldview of the insular testing profession. For this reason, we consider him one of the pioneers of social justice concerns in educational assessment, and his message endures in contemporary writings in this area (e.g., Dixon-Roman, 2020).

Samuel Messick

In his Presidential Address to the quantitative psychology division of APA, Messick (1975) invited measurement professionals to consider the meaning and values in educational measurement. In addition to focusing on the "construct" measured by a test, he also focused on consequences, pointing out "the social consequences of test use should be weighed against the social consequences of *not* testing"

(p. 962). In his seminal chapter on validity theory (Messick, 1989), he further elevated the importance of considering the consequences of testing by specifying two "interconnected facets" of validity that comprised the "consequential basis" of test validity (p. 20). The first facet he termed the *consequential basis of test interpretation*, which he defined as the "appraisal of value implications of construct label, theory underlying test interpretation, and ideologies in which theory is embedded" (p. 20). The second facet he called the *consequential basis of test use*, which he described as the "appraisal of both potential and actual social consequences of applied testing" (p. 20).

In calling attention to the consequences associated with how test scores are both interpreted and used, Messick underscored how value systems underlie the determination of what is tested, the meaning ascribed to test scores, and how scores are used. His requirement that testing consequences be fully examined was comprehensive, and he provided several examples of how negative consequences or unvalidated test use could lead to injustice at both individual and societal levels. These examples included value-laden labels attached to test scores, adverse impact, and effects of tests on instruction.

Validity evidence in support of test use was not to be taken as an excuse to avoid a comprehensive study of the potential consequences of tests. As Messick (1989) put it, "Even if adverse testing consequences derive from valid test interpretation and use, the appraisal of the functional worth of the testing in pursuit of the intended ends should take into account all of the ends, both intended and unintended, …including… individual, institutional, societal, and systemic effects" (p. 85). In short, Messick's introduction of the consequential basis of test validity compelled the measurement community to discuss whether testing programs foster (or reduce) societal inequities. In essence, he, like Ebel, brought conversations of social justice to psychometrics, albeit indirectly.

Messick's consequential basis of test validity sparked much debate in the testing literature, including two special issues in the NCME flagship journal *Educational Measurement: Issues in Practice* (EM: IP) in 1997 and 1998. The authors in those special issues generally agreed on the importance of evaluating testing consequences, but disagreed whether such evaluation should be considered part of the concept of validity. Nevertheless, as the AERA et al. (1999) *Standards* illustrated (and AERA et al., 2014, too), Messick's call to evaluate consequences—

both intended and unintended—became embodied in best practice guidelines for educational assessment. However, promoting guidelines and standards does not guarantee they will be followed (Johnson, Trantham, & Usher-Tate, 2019). Thus, enforcement of the *Standards* remains a problem (Gitomer et al., 2021).

Sylvia Johnson

Johnson (2000) epitomized the argument that testing consequences are social justice issues in educational assessment. She cautioned against the unintended negative consequences of test-based reform efforts on marginalized students, particularly the use of tests that result in students receiving less rigorous instruction. She claimed such test-based practices had not only negative effects on the individual children, but on society as well because using test scores to make such decisions may result in "a serious loss to society…through failure to identify and develop the real talents of all its members" (p. 151).

Johnson pointed out test-based education reform efforts may lose sight of what is important in education (supporting children to reach their potential) in pursuit of institutional goals. She stated,

> *tests often are advertised as being designed to assess rigorous curriculum standards, but far more attention is typically given to the match between standards and tests than to the essential prerequisite—that is, the match between high standards and instruction for all students being assessed. As a result, the available test products serve mainly institutional needs, offering little benefit to test-takers, teachers, or even to schools in terms of prescriptive information or instructional value.* (p. 155)

She also noted test-based reform efforts were not serving students well due to the harsh and negative messages and actions being sent by policymakers. Her description of this concern falls squarely in the realm of social justice in educational assessment:

> *The language of high standards and testing is often conveyed to the recipients of today's testing products, usually students and their teachers, in a punitive, blame-filled, and even threatening rhetoric which asserts that both have left undone what should have been done and have done what they should not have done. This language further asserts that both groups will reap dire consequences*

> *if test scores do not evidence achievement of high standards. Researchers have shown what tends to happen in many urban schools serving predominantly Black and Hispanic communities under such conditions: Teachers teach to the test in routinized style and emphasize lower order skills in hopes of getting some level of minimal test performance out of their students at the expense of more motivational and interest-eliciting activities that might engage students in a constructive process of learning but which are also risky, may be noisy, take time, and may be more difficult to justify to a supervisor.* (p. 155)

Johnson's description of the effects of tests on historically marginalized students illustrates why studying unintended negative consequences is so important. Studying such consequences is the hallmark of social justice research in educational assessment.

Janet Helms

Like Johnson, Helms (2006) called attention to social justice issues in educational assessment, specifically the issue of adverse impact (the consequences of test score use affecting some groups of test takers more than others). She rejected AERA et al.'s (1999) position that test score differences across groups did not signify bias if they could not be traced to a source of construct-irrelevant variance. In her view, "When test scores that differ by racial groups are used for assessment purposes, resulting decisions regarding members of the lower scoring group are potentially unfair" (p. 845). She pointed out that for historically minoritized students, test scores contained construct-irrelevant variance due to factors that today we would describe as effects of systemic racism. She claimed fair tests involve removing "systematic variance attributable to experiences of racial or cultural socialization" from the scores of minoritized test takers.

Helms (2006) argued the traditional models of test fairness that focus on differential predictive validity miss the mark, because "the focus of these models is on the adverse consequences of using potentially unfair test scores rather than the consequences of such scores" and these models "erroneously treat racial groups as meaningful constructs" (p. 848). Instead of studies of differential predictive validity and measurement invariance, she called for "Replacement of racial and ethnic categories with cultural constructs derived from conceptual frameworks [as] a necessary condition for fair assessment" (p. 848).

Ford and Helms (2012) further pointed out the effects of test-based adverse impact on African Americans claiming,

> *such racial-group test-score disparities mean that typically more than half of African Americans are excluded from a variety of academic and vocational experiences and domains in society beginning when tests first enter their lives. Such exclusion contributes to not only low achievement, but also underachievement; and denied opportunities fuel the greatest educational problem—the achievement gap.* (p. 187)

They further described the lack of voice, and the lack of power, African Americans have in the education system as a clear social justice issue. As they put it,

> *unlike their White counterparts, African Americans…exist in and face racially and economically discriminatory contexts and practices that are not aspects of the socialization experiences of White people. Essentially and undeniably, African Americans exert little influence over the structure and content of the tests on which they are expected to perform as if they are White middle-to-upper class, monolingual Americans.*

Of course, test score differences and adverse impact are not the only manifestations of social injustice in education, or in educational assessment. Ford and Helms helped connect these manifestations to the larger picture or injustices in education and how they interact with the traditional educational assessment practices.

Summary of social justice pioneers in educational assessment

We could elaborate further on the calls for socially just assessment policies made by Ebel, Messick, Johnson, and Helms; as well as by other strong voices that focused on addressing adverse consequences (e.g., Lane, 2014; Linn, 1984; Mislevy, 2018; Shepard, 1993). There have also been professional organizations outside of AERA, APA, and NCME that have promoted more just assessment practices (e.g., the International Test Commission's *Guidelines for the Large-Scale Assessment of Linguistically and Culturally Diverse Populations*, 2018). However, our point in reviewing some of the pioneers in this area is to call attention to the fact that social justice issues are not new. In fact, for decades, prominent measurement specialists have warned of the injustices being caused to both individuals and societies due to the use of educational assessments.

We also acknowledge that the previous calls for more attention to issues of fairness, testing consequences, and validity did not represent a comprehensive discussion of social justice issues in educational assessment. A more comprehensive discussion requires considering consequences at both the individual and societal levels. For example, in describing notions of justice at these levels, Rawls (1999) remarked,

> *justice denies that the loss of freedom for some is made right by a greater good shared by others. It does not allow that the sacrifices imposed on a few are outweighed by the larger sum of advantages enjoyed by many.* (p. 3)

In the case of educational assessment, this conflict becomes evident when the use of a test denies access to something that can be considered an individual right, such as the pursuit of a college degree. The social consequences of such test use can benefit colleges and some individuals. However, a focus on social justice requires recognition of consequences that cannot be accepted in any scenario (e.g., denial of fundamental rights). Currently, it appears the measurement profession is stuck on delineating and forming consensus on what are unacceptable consequences associated with educational tests. Forming that consensus, if possible, will require formal dialogues on defining social justice in educational assessment. Understanding and studying the consequences of testing is a prerequisite for a serious conversation in this area. We hope the dialogue can occur soon in the academic and other professional spaces in which measurement specialists operate. Otherwise, it will unfold in the courtroom (e.g., IntegrateNYC vs. State of NY, 2021). In the spirit of promoting this dialogue, in the next section, we offer some suggestions on how we can move forward toward more socially just educational assessment practices.

Promoting Social Justice in Educational Assessment: A Call to Action

As mentioned at the opening of this chapter, social justice has been a difficult concept to define in education. It is likely to prove challenging to define within educational assessment, as well. However, in our view, social justice in educational assessment means developing and providing assessments that value all students; embrace cultural diversity; allow students to access their various funds of knowledge when interacting with the assessment; provide proper support systems for students to successfully access and navigate the assessment; and support the

positive development of students' self-esteem, and their acquisition of academic, occupational, and other goals. These actions acknowledge and honor test taker rights (e.g., APA, 2020) and increase assessment transparency by helping test-takers and educators better grasp and trust the assessment process. Although other measurement specialists may add or delete from our definition, we offer this definition as a starting point, or rather a guidepost, for assessment equity (i.e., supporting the principle that assessments should be fair and considerate of all students' backgrounds and opportunities to learn).

In our previous review of pioneers in social justice in educational assessment, we highlighted some of the social justice issues to be addressed. In this section, we turn to what we can do to become more socially just measurement researchers and practitioners. We believe a social justice perspective in educational assessment must begin with the best interests of each student in mind. These interests include the student's personal and communal values, and so the interests of the communities in which students live must also be front-of-mind. Consideration of these interests involves five key components: equity, access, participation, rights, and diversity. In the next section, we briefly describe actions we in the measurement community can take with respect to each of these components of social justice in educational assessment.

Equity

Like many terms in education, "equity" can mean different things to different people, and is often confounded with "equality." Gordon (1995) distinguished between the two by stating,

> *care must be taken to make clear the difference between equity and equality. Equity speaks to and references fairness and social justice; it requires that the distribution of social resources be sufficient to the condition that is being treated. Equality, on the other hand, connotes sameness and the absence of discrimination.* (p. 363)

Gordon points out equality is not always just, if individuals have different needs, but are given the same resources.

One suggestion we have for making assessments more socially just with respect to equity begins with a request for AERA, APA, and NCME to revise their guidance

on observations of group differences in test scores and adverse impact. The current guidance gives the appearance of "don't shoot the messenger," which is a theme with which many can identify. However, the messenger can also return new information to the original sender, and for far too long, the message returned—that the adverse impact observed based on test scores is unacceptable—has not been adequately responded to. Our suggestion to these organizations is to revise statements in the *Standards* to promote not only the acceptance of responsibility but also to encourage the psychometric community to take collaborative action. Thus, our recommendation is to change a statement such as the following,

> …the *Standards'* measurement perspective explicitly excludes one common view of fairness in public discourse: fairness as the equality of testing outcomes for relevant test-taker subgroups. Certainly, most testing professionals agree that group differences in testing outcomes should trigger heightened scrutiny for possible sources of test bias…However, group differences in outcomes do not in themselves indicate that a testing application is biased or unfair. (AERA et al., 2014, p. 54),

to,

> …the *Standards'* measurement perspective acknowledges equality of testing outcomes for subgroups may not always be realized. Thus, when unequal outcomes occur, they should trigger heightened scrutiny for possible sources of test bias, and any discovered sources of bias should be eliminated. Moreover, testing agencies should go beyond the tests themselves and work with educators, employers, and other stakeholders to promote equitable systems of education and employment that do not differ in quality or outcomes for student groups defined by race, ethnicity, culture, and other sociopolitical or sociological characteristics.

This change will acknowledge the point that group differences are not directly caused by tests, which was the point of the wording in the current version of the *Standards*, but it simultaneously acknowledges (a) the problem is not acceptable, and (b) there is an expectation that we help fix it. Our suggestion to add these acknowledgements is consistent with Gordon (1995) who argued,

> *it can be argued that the problems of equity in educational assessment are largely secondary to the failure to achieve equity through educational treatments. However, the fact that these problems of equitable educational assessment are only secondarily problems for assessment does not mean that they should not be engaged by the assessment community, even if they cannot be solved through assessment alone.* (p. 360).

In our view, the AERA, APA, and NCME *Standards* provide an opportunity to describe what testing professionals can do to promote social justice in education. Thus far, the *Standards* have not fully embraced that opportunity. Requiring that testing consequences be studied, without consequences for the testing agencies who ignore them, is a blatant disregard of social justice in educational assessment.

Access and Participation:

Our suggested recommendations to promote greater access and participation in educational assessments are based on the concept of *UNDERSTANDardization* (Sireci, 2020), which extends the concept of promoting access to individuals with disabilities to educational assessments, to promoting access based on all types of test-taker characteristics. UNDERSTANDardization involves keeping the fundamental principles of standardization (i.e., keep everything the same for everyone), but "loosens" what is required to be the same. As Sireci described, testing professionals "must understand the numerous dimensions of heterogeneity that exist within the populations of people we test, and embed that understanding in our standardization processes" (p. 101). Essentially, the goal of this more flexible approach to test development, administration, and scoring is to understand (a) what each student brings to the testing situation in addition to the proficiency measured, (b) how these personal characteristics may interact with testing conditions, and (c) how testing conditions can be sufficiently flexible to accommodate and account for these potential interactions. This understanding will lead to more valid assessment of each individual student's proficiencies.

UNDERSTANDardization uses the principles of standardization to create a more socially just assessment environment by allowing for flexibility in the content and tasks presented to students (i.e., allowing student choice with respect to assessment content and/or context), how they respond (e.g., code-switching between languages, responding orally rather than by writing), and what is

scored (English grammar versus grammar that acknowledges linguistically diverse cultures). By incorporating the principles of UNDERSTANDardization into assessment, students are more likely to "see themselves" in the assessment and feel empowered by the choices they are given.

Participation and Diversity

An important part of UNDERSTANDardization is doing the research to understand the diversity of the population of test takers. This understanding is also key to the larger concept of an antiracist framework for assessment, which Randall (2021) described as an approach to assessment that,

> *requires an explicit confrontation of racism in our assessment practices; and works to disrupt these systems of oppression…An antiracist framework for assessment critically questions the structures and assumptions that make up the judgment of all assessment developers. It is explicit about its politics and its intent to reconstruct hierarchical racial power arrangements that have been historically (re)produced via assessments.* (p. 1)

Thus, the antiracist approach has diversity at its core from the earliest stages of test development. Following these practices will lead to culturally sustaining assessments that de-center whiteness and explicitly acknowledge the beauty and relevance of all cultures, particularly those that have been historically minoritized. A key element of this approach is, rather than screening out test material specific to a particular culture, include it to value that culture. As Randall (2021) argued,

> *the problem is by removing race, or pretending that it does not exist/matter, one is not removing racism. Indeed, such a practice only perpetuates racism as it is simply a proxy for elevating whiteness.* (p. 4)

Another suggestion we have related to diversity is increasing the numbers of African American, Hispanic/Latino, Indigenous, and people from other minoritized groups into the measurement profession. Randall, Rios, and Jung (2021) reported that of 3,124 degrees in the measurement field conferred from 1997 to 2016, only 6.8% were Black and 3.7% were Hispanic/Latino. Clearly, more needs to be done, and the collaboration among the Chan-Zuckerberg Initiative, the Center for Educational Assessment at the University of Massachusetts Amherst, and NCME from 2020–2023 was one important step to increase these numbers. Through this

collaboration, two cohorts of 30 Black and Brown graduate students in educational measurement were awarded travel and mentoring scholarships to attend NCME's annual conference and participate in a program where they were mentored by a Black or Brown Ph.D. scholar. Our suggestion is to improve these efforts and extend them to funding stipends and tuition for Black and Brown students interested in pursuing masters and doctorates in educational measurement.

Call to Action Summary

We have made four modest recommendations for specific actions we can take to promote more socially just assessment practices. These actions are,

(a) Revise the AERA et al. (2014) *Standards'* text related to group differences and adverse impact to encourage testing agencies to work with other stakeholders to reduce inequities and comprehensively investigate validity evidence based on testing consequences;

(b) use the principles of UNDERSTANDardization to provide more flexible test administration practices;

(c) use an antiracist framework for construct definition and test development;

(d) initiate serious recruitment and support efforts to bring Black/Brown colleagues into the measurement field.

It is important for us to add one more action to this list, because this action reflects the good work that has already been done in pursuit of fairness and justice in educational assessment:

(e) continue and expand research on test fairness.

This last action is important because we must acknowledge that the work done in our field for decades to evaluate item and test bias and make tests more accessible is crucial for accomplishing the goals of socially just educational assessment. We must continue such research, but we should not limit ourselves to what has become routine. We must continue to push the boundaries of what we consider to be bias, fairness, and equity; to listen to what the communities within our tested populations consider as fair in *their* contexts, and develop the statistical machinery to evaluate fairness in a situated way. Only then can we claim psychometrics

has done what it can to help all individuals in society attain their academic, occupational, and other personal goals. Furthermore, these actions will lead to assessments of higher quality that support more valid interpretations of students' proficiencies.

Our list of five actions we can take is a brief list, and it is certainly not exhaustive of all that can be done. However, it represents a reasonable starting point for others to join in. We note others are joining in (e.g., Lyons, Johnson, & Hinds, 2021), and we applaud those efforts.

Concluding Remarks

In an earlier section of this chapter, we focused on the work done by APA, AERA, and NCME in promoting guidelines for the testing industry that promote sound and fair testing practices. We also pointed out discussions of social justice and fairness were hard to find in the first two versions of the *Standards* and we recommend improvements to the latest version, which is eleven years old at the time of this writing. A more recent development is relevant and important for us to mention before closing. On October 29, 2021 APA released a formal apology for its longstanding contributions to systemic racism in the profession. Included in the apology was an acknowledgement of the damage done via educational and psychological tests. This acknowledgement is a first step toward reparations, which are also being considered and enacted by APA (See https://www.apa.org/about/policy/racism-apology). Other institutions, such as NCME, should follow APA's lead and acknowledge the misuse of educational tests in a similar manner (Sireci, 2021).

We opened this chapter with a lyric from Peter Tosh. His lyrics are provocative in that he rejects peace, but his point is, by achieving equal rights and justice, peace is a given. We believe by pursuing social justice in educational assessment, we will do our part to move us forward on the path to peace, on which we all should be traveling. By incorporating a social justice perspective to assessment in the service of learning, we not only move closer to equal rights and justice in assessment; we also facilitate the goal of making our assessment systems more transparent, more equitable, and consequentially more valid, for every learner.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

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

American Psychological Association, Committee on Test Standards. (1952). Technical recommendations for psychological tests and diagnostic techniques: A preliminary proposal. *American Psychologist, 7,* 461–465.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*, (2, supplement).

American Psychological Association. (1966). *Standards for educational and psychological tests and manuals*. Author.

American Psychological Association (2020). *The rights and responsibilities of test takers.* Author. https://www.apa.org/science/programs/testing/rights.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological tests*. American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. American Psychological Association.

Bell, L. A. (1997). *Theoretical foundations for social justice education*. In M. Adams, L. Bell, & P. Griffin (Eds.), Teaching for diversity and social justice: A sourcebook (pp. 3–15). Routledge.

Bertrand, M., & Marsh, J. (2021a). How data-driven reform can drive deficit thinking. *Phi Delta Kappan, 102*(8). https://kappanonline.org/how-data-driven-reform-can-drive-deficit-thinking-bertrand-marsh/

Bertrand, M., & Marsh, J. A. (2021b). Opting out of standardized tests: The role of parents and their social networks. *American Journal of Education, 127*(2), 231–259. https://doi.org/10.1086/711828

Boyles, D., Carusi, T., & Attick, D. (2009). Historical and critical interpretations of social justice. In W. Ayers, T. M. Quinn, & D. Stovall (Eds.), *Handbook of social justice in education* (pp. 30–42). Routledge.

Dixon-Román, E. (2020). A haunting logic of psychometrics: Toward the speculative and indeterminacy of blackness in measurement. *Educational Measurement: Issues and Practice, 39*(3), 94–96.

Ebel, R. L. (1963). The social consequences of educational testing. Proceedings of the 1963 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service.

Ford, D. Y., & Helms, J. (2012). Overview and Introduction: Testing and assessing African Americans: "Unbiased" tests are still unfair *Journal of Negro Education, 81*, 186–189.

Gitomer, D. H., Martinez, J. F., Battey, D., & Hyland, N. E. (2021). Assessing the assessment: evidence of reliability and validity in the edTPA. *American Educational Research Journal, 58*, 3–31.

Gordon, E. (1995). Toward an equitable system of educational assessment. *Journal of Negro Education, 64*, 360–372.

Gordon, E. W. (2020). Toward assessment in the service of learning. *Educational Measurement: Issues and Practice, 39*(3), 72–78.

Guthrie, R. V. (1998). *Even the rat was white: A historical view of psychology* (2nd ed.). Boston, MA: Allyn & Bacon.

Helms, J. (2006). Fairness is not validity or cultural bias in racial-group assessment: A quantitative perspective. *American Psychologist, 61*(8), 845–859.

Hobson, C. J., Szostek, J., & Griffin, A. (2021). Adverse impact in Black student 6-year college graduation rates. *Research in Higher Education Journal*, *39*, 1–15.

IntegrateNYC vs. State of NY (2021). *Index No. 152743/2021.*

International Test Commission (2018). *Guidelines for the large-scale assessment of linguistically and culturally diverse populations*. https://www.intestcom.org/files/guideline_diverse_populations.pdf

Jean-Marie, G., Normore, A. H., & Brooks, J. S. (2009). Leadership for social justice: Preparing 21st century school leaders for a new social order. *Journal of Research on Leadership Education, 4*(1), 1–31.

Johnson, J. L., Trantham, P., & Usher-Tate, B. J. (2019). An evaluative framework for reviewing fairness standards and practices in educational tests. *Educational Measurement: Issues and Practice, 38*(3), 6–19.

Johnson, S. T. (2000). The live creature and its expectations for the future. *Journal of Negro Education, 69,* 150–158.

Koljatic, M., Silva, M., & Sireci, S. G. (2021a). College admission tests and social responsibility. *Educational Measurement: Issues and Practice*. https://doi.org/10.1111/emip.12425

Koljatic, M., Silva, M., & Sireci, S. G. (2021b). College admission tests and social responsibility: A response to comments. *Educational Measurement: Issues and Practice*.

Koljatic, M., & Silva, M. (2021). College entrance exams: International perspectives. *Pensamiento Educativo: Revista de Investigación Educacional Latinoamericana, 58*(1), 1–19. https://doi.org/10.7764/PEL.58.1.2021.2

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema, 26*(1), 127–135. https://doi.org/10.7334/psicothema2013.258

Linn, R. L. (1984). Selection bias: Multiple meanings. *Journal of Educational Measurement, 21,* 33–47.

Lyons, S., Johnson, M., & Hinds, B. F. (2021). *A Call to Action: Confronting inequity in assessment*. https://www.lyonsassessmentconsulting.com/assets/files/Lyons-JohnsonHinds_CalltoAction.pdf.

Malda, M., van de Vijver, F. J. R., & Tamane, Q. (2010). Rugby versus soccer in South Africa: Content familiarity contributes to cross-cultural differences in cognitive test scores. *Intelligence, 38,* 582–595.

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

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement*, (3rd ed., pp. 13–100). American Council on Education.

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.

Porfilio, B. J., Strom, K., & Lupinacci, J. (2019). Getting explicit about social justice in educational doctoral programs in the US: Operationalizing an elusive construct in neoliberal times. *Journal of Educational Foundations, 32*, 104–123.

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*.

Randall, J., Rios, J. A., & Jung, H. J. (2021). A longitudinal analysis of doctoral graduate supply in the educational measurement field. *Educational Measurement: Issues and Practice*, *40*(1), 59–68. https://doi.org/10.1111/emip.12395

Rawls, J., (1999). *A theory of justice: revised edition*. Cambridge, MA: Belknap Press of Harvard University Press.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405–450.

Sireci, S. G. (2016a). Comments on valid (and invalid?) commentaries. *Assessment in Education: Principles, Policy & Practice, 23,* 319–321.

Sireci, S. G. (2016b). On the validity of useless tests. *Assessment in Education: Principles, Policy & Practice, 23,* 226–235.

Sireci, S. G. (2020). Standardization and UNDERSTANDardization in educational assessment, *Educational Measurement: Issues and Practice, 39*(3), 100–105. https://doi.org/10.1111/emip.12377

Sireci, S. G. (2021). Valuing educational measurement. *Educational Measurement: Issues and Practice, 40*(1), 7–16. https://doi.org/10.1111/emip.12415.