

Our validity looks like justice. Does yours?

Jennifer Randall 
University of Michigan, USA

Mya Poe 
Northeastern University, USA

David Slomp
University of Lethbridge, Canada

Maria Elena Oliveri
University of Nebraska, USA

Language Testing
2024, Vol. 41(1) 203–219
© The Author(s) 2023



Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/02655322231202947
journals.sagepub.com/home/ltj



Abstract

Educational assessments, from kindergarden to 12th grade (K-12) to licensure, have a long, well-documented history of oppression and marginalization. In this paper, we (the authors) ask the field of educational assessment/measurement to actively disrupt the White supremacist and racist logics that fuel this marginalization and re-orient itself toward assessment justice. We describe how a justice-oriented, antiracist validity (JAV) approach to validation processes can support assessment justice efforts, specifically with respect to language assessment. Relying on antiracist principles and critical quantitative methodologies, a JAV approach proposes a set of critical questions to consider when gathering validity evidence, with potential utility for language testers.

Keywords

Antiracist validity, justice-oriented assessment, justice-oriented measurement, justice-oriented validity, language assessment, validity

Introduction

Junot Diaz, a Pulitzer Prize-winner author once said, “You know how vampires have no reflections in the mirror? If you want to make a human being a monster, deny them, at

Corresponding author:

Jennifer Randall, University of Michigan, Ann Arbor, MI 48109-1382, USA.
Email: jennrand@umich.edu

the cultural level, any reflection of themselves. And growing up, I felt like a monster in some ways. I didn't see myself reflected at all. I was like, 'Yo, is something wrong with me?' That the whole society seems to think that people like me don't exist? And part of what inspired me was this deep desire, that before I died, I would make a couple of mirrors. That I would make some mirrors, so that kids like me might see themselves reflected back and might not feel so monstrous for it" (Stetler, 2009, para. 2). Despite efforts by a few educational measurement scholars to promote the use of, and demonstrate the need for, culturally responsive approaches to assessment (Hood, 1998; Lee, 1998; Montenegro & Jankowski, 2017; Moss, 1992; Qualls, 1998; Shepard, 2021), the field of assessment has offered little. Rather, we (the assessment community including test developers and researchers) have opted to deny Black, Brown, and Indigenous test takers access to any real mirrors so that they might see their true selves reflected back at them; choosing to offer up reading passages with Juan in place of John and tacos instead of hamburgers as consolation. We then double-down on our White supremacist practices by maintaining the status quo constructs of White supremacy and describing these populations using deficit language such as achievement gap and lack of parental engagement, ostensibly contributing to the racist narrative that Black, Brown, and Indigenous persons are inferior, uneducable, and (it follows) monsters.

More still, the field of assessment has been even less willing to consider the impact of the larger White supremacist, racist ecosystem in which all students operate/navigate daily; and how these oppressive systems can, and do, influence student response processes and performance outcomes. For example, in the United States, Black communities have a long history of being the target of aggressive, and too often lethal, over-policing practices (Epp et al., 2014; Gottfredson et al., 2020; Kane, 2002; Smith, 1986); and this state-sanctioned violence extends to schools (see Finn & Servoss, 2014; Fisher & Hennessy, 2016; Research for Action, 2020; Weiler & Cray, 2011; Whitaker et al., 2019). In states where corporal punishment is allowed in schools, Black students are 2.3 times more likely than White students to receive it. Specifically, Black boys are twice as likely as White boys to be victimized by corporal punishment, and Black girls are four times as likely as White girls to be subjected to this (U.S. Department of Education, Office of Civil Rights, 2023). On 24 March 2023, the US Secretary of Education, Miguel Cardona, wrote in a public letter to governors, chief state school officers, and school district and school leaders "Despite years of research linking corporal punishment to poorer psychological, behavioral, and academic outcomes, tens of thousands of children and youth are subjected to beating and hitting or other forms of physical harm in school every academic year, with students of color and students with disabilities disproportionately affected."

To be clear, we are not suggesting that educational assessments, or language assessments specifically, should be considered the sole or primary instigator of the physical and emotional violence against Black peoples. In fact, almost a 100 years ago, American historian, Carter G. Woodson (1933/1999), outlined the many ways that Black students at American schools were denied education of African contributions to science, history, literature, and fine arts and the implications of such erasure: "Unlike other people, then, the Negro, according to this point of view, was an exception to the natural plan of things,

and he had no such mission as that of an outstanding contribution to culture. The status of the Negro, then, was justly fixed as that of an inferior. Teachers of Negroes in their first schools after Emancipation did not proclaim any such doctrine, but the content of their curricula justified these inferences” (p. 34). Similarly, W.E.B. DuBois (1935/2017), co-founder of the National Association for the Advancement of Colored People (NAACP), argued that education, broadly speaking, has provided a consistent damaging and racist narrative of Black Americans, in particular, as incapable of citizenship. He wrote: “In propaganda against the Negro since emancipation in this land, we face one of the most stupendous efforts the world ever saw to discredit human beings, an effort involving universities, history, science, social life and religion” (p. 595). Nonetheless, the long history of deficit-framed assessment-based narratives, including and especially with respect to literacy, against Black peoples have also supported their marginalization and dehumanization and should be added to DuBois’ list. And this dehumanization happens from the beginning of one’s educational journey to the end—from reading readiness exams to college writing placement exams. For example, when referring to literacy in young children, American Pulitzer Prize-winning novelist, Toni Morrison said:

It is terrible to think that a child with five different present tenses comes to school to be faced with books that are less than his own language. And then to be told things about this language, which is him, that are sometimes permanently damaging. He may never know the etymology of Africanisms in his language, not even know that “hip” is a real word or that “the dozens” meant something. This is a really cruel fallout of racism. . . (LeClair, 1981).

Inoue (2015) chronicles the extension of these racist ideologies in discussing writing assessment in higher education writing:

What troubles me are people who look at racial inequalities, look at racism in writing classrooms and programs, and say, “how do we know that is racism?” My mind often whirls at such questions. Forget for a moment how it happened, inequalities are here. No African-Americans in your classes, few in your school. Where are the Native Americans? Most who are there, do not do well. They fail. Why? Isn’t it enough to see such patterns? . . . Here’s what matters to me. White students uniformly and historically do better on most if not all writing assessments, large-scale or classroom. It may not be intentional, but it is racism, and it is a product of the writing assessment ecologies we create. Do not get me wrong. I do not blame White students or teachers. I blame writing assessments (p. 22).

And one need only look to recent headlines (re: Students for Fair Admissions, Inc. v President and Fellows of Harvard College) in which an anonymous group of Asian American applicants claimed that African-American and Hispanic students (with lower test scores, but higher ratings on other measures) receive an unfair advantage in Harvard admissions; and that Asian students’ higher test scores (e.g., ACT and SAT), on average, serve as evidence that they are more qualified for admission and being discriminated against. This case serves as a prominent example of how admissions test scores have been used successfully as evidence of Black inferiority and preferential treatment in higher education admissions. In response, Haynes (2023) wrote,

Scores on these tests are numbers, on these constructed measures, which contain sources of human error, and that cannot, and do not reflect the human variations of ability and creative capacities of individuals. The main arguments among those who have opposed affirmative action, have been focused on test score disparities, and GPA differences, among, and between college applicants, while neglecting the fact that these measures are flawed, imprecise, contain bias (p. 3).

Indeed, the results of these so-called, or so-assumed, objective large scale language assessments are too often used to perpetuate White supremacist and anti-Black racist logics of Black inferiority, thereby blaming the student and failing to consider the systems of oppression and marginalization that the results reflect. These interpretations represent a new form of racism (Bonilla-Silva, 2013), which is more nuanced and subtle, making it more difficult to detect and easier to dismiss with a nonracial explanation.

Goal of Viewpoint contribution

In this Viewpoint piece, we discuss our ideas for getting us closer to something that looks like, feels like, and operates like justice, specifically assessment justice,¹ which we interpret here 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. Justice-oriented approaches to assessment do not seek to serve the *greater good* of the many and powerful to the exclusion of the few and minoritized, but rather assertively prioritize the most marginalized populations. A commitment to assessment justice goes further than a commitment to equity, as equity-driven approaches to assessment seek merely to provide scaffolds that compensate for historical and contemporary barriers. Justice-oriented approaches, on the other hand, actively seek to remove those barriers and also make amends for the damage those barriers have already created (Randall, 2023).

This kind of commitment to justice will require an extraordinary shift in our typical approaches to assessment design and development. This shift will mandate an elevation of the measurement field's critical consciousness (Lyons et al., 2021; Randall, 2021; Randall, 2023) and a willingness to examine and reexamine/interrogate all of our assumptions. Assessments could serve as a tool for liberation if we want them to. We can design them to be that way from the beginning (Oliveri et al., 2021; Slomp, 2016). But what so many of our exams—including oral language, literacy, reading, and writing assessments—seemingly fail to do repeatedly is consider the larger sociopolitical context in which our work happens and the consequences of our design and development decisions for the most marginalized communities. Doing so (i.e., pretending not to know or see) allows assessment developers to nurture the public illusion of tests as neutral tools and that their results represent some objective measure of capacity, achievement, and [most absurdly] merit (see also Shohamy, 2001).

Validity has long been acknowledged as the most important concept in measurement; and if we are to move forward with a justice orientation, then we must begin there. Although the current validity paradigm was not constructed with racially minoritized persons in mind, we can still work within it and completely re-orient these processes with a justice framing/lens. We (Randall et al., 2022) proposed a justice-oriented, antiracist validity (JAV) approach intended to acknowledge and disrupt the discriminatory legacies against racially minoritized students in US educational assessment. Using Quantitative Critical Race Theory (QuantCrit) and antiracist educational principles as the frame, we build on Kane's (2013) interpretation/use argument (IUA) validation model and Mislevy's (2018) sociocognitive extension of that model to critically interrogate the validation process for evidence of White supremacist/racist logics and offer suggestions for disrupting those logics. In other words, we are going to take currently accepted justice-agnostic approaches to validation and talk about how we can make them justice-seeking.

The foundation

We begin with Kane's IUA, which has been well-established and widely accepted as a rigorous approach to validation for over a decade (see also Kane, 2010 article, "Validity and fairness," in *Language Testing* published shortly after his Messick lecture at the Language Testing Research Colloquium (LTRC) in Cambridge, UK, in 2010). Kane, through his IUA, maintains that the kinds of evidence required for validation are determined by the claims being made, and more-ambitious claims require more evidence than less-ambitious claims. Theory of action (ToA) research (Oliveri et al., 2021) extends those claims into immediate and long-term intended and unintended consequences. Josh Lederman (2023) takes the idea of ToA one step further in addressing racial injustice, writing that "Any assessment created with the intention of a desired social impact would need to explicate these claims in the IUA—and could not validate the IUA without evidence addressing these claims." He goes on to write succinctly, "If the test developers did not want these matters to be included in the validity/validation of the exam, they would have to exclude the claims from their IUA, no longer claiming that this test was trying to impact racial justice in educational achievement. But once the States claim to have equity goals as an articulated purpose, that is where its validity is to be found" (p. 250).

Now, we argue from a justice perspective that the consequences of an assessment begin with the moment a test taker is made aware that the assessment will take place; and that, without question, descriptions and interpretations in and of themselves (even those independent of any explicit action or test score associated with them) have a very real impact on test takers (see references to stereotype threat, racial trauma, and so on in the following sections). Nonetheless, the vast majority of language assessments, both classroom and large-scale, have scores and a wide-range of actions associated with them that can result in racial injustice: from being labeled literate, or not (even if only in the mind of the teacher), to assignment, or not, to developmental (formerly referred to as remedial) classes to admissions decisions. Moreover, in our review of the mission and purpose statements of several organizations that offer language assessments (e.g., ETS, Duolingo,

International English Testing System), we found that most do make some claim about how their tests ensure fairness, but do not offer the inferences that support those claims or offer any guidance on test misuse.

While we hoped 10 years ago that moving Fairness into the Foundational Concepts section of the *Standards for Psychological and Educational Testing* (American Educational Research Association [AERA], et al., 2014) would draw attention to questions of injustice, it has not. Scholars and practitioners continue to go back and forth about what fairness means, but we are going to interpret fairness as justice (as did John Rawls, 1971, 1999 and Charles Mills, 2017). And if fairness is justice, then injustice must also be addressed in every foundational concept in measurement. In other words, if justice is the stated goal, then assessment developers must show us (meaning all rights holders) transparently what you have done in service to that goal and how you have, or have not, succeeded.

Applying a sociocognitive lens, Mislevy (2018) defined validity in this way: “A test is valid for a given interpretation or use to the degree to which empirical evidence and theoretical rationales support reasoning as if ‘it measures what it is purported to measure’” (p. 202). We maintain that from a sociocognitive perspective, the essential question should be, Whose Linguistic, Cultural, and Substantive (LCS) patterns are being privileged by an assessment, whose LCS patterns are being devalued, omitted, suppressed, or marginalized? (Randall et al., 2022).

The critical re-orientation

A justice-oriented antiracist perspective to validity situates our critical interpretation of Kane’s IUA and Mislevy’s sociocognitive extension within the wider frame of antiracist education and Quantitative Critical Race Theory. An antiracist approach to education requires that we question all assumptions with respect to knowledge and knowledge production (particularly those that privilege, or elevate, whiteness and subjugate all else). It does not focus on the racist behaviors, biases, and/or discriminatory leanings of individuals, but rather focuses on critically examining the ways in which institutions, organizations, and systems support racist practices and policies and then disrupting, or rupturing, those systems (Dunn et al., 2021; Vincent, 1992). Antiracist education is political (hooks, 1994). There is no such thing as knowledge for the sake of knowledge in an antiracist education framework— knowledge exists for the purpose of eliminating racism. An antiracist approach for assessment, consequently, must also be “explicit about its politics and its intent to reconstruct hierarchical racial power arrangements that have historically been produced and reproduced by assessments” (Randall, 2021, p.1).

Moreover, we find Quantitative Critical Race Theory (QuantCrit) to be a useful observational and methodological lens within this justice-oriented, antiracist critical re-orientation. QuantCrit, which applies Critical Race Theory’s primary tenets (Bell, 1995; Delgado & Stefancic, 2017) to quantitative methodologies, maintains that (1) economic, political, and educational spaces are defined by oppressive systems to include racism, sexism, ableism, and religious oppression; (2) because numbers (i.e., statistics) have historically been viewed as objective and neutral (they are not), they have historically been used and manipulated to perpetuate racist ideologies; (3) racial categories (typically

Table 1. Justice-oriented antiracist validity.

IUA	What evidence is there that the assessment supports/facilitates justice-based outcomes?
Sociocognitive model	Whose linguistic, cultural, and substantive (LCS) patterns are being privileged by an assessment? Whose LCS patterns are being devalued, omitted, suppressed, or marginalized?
Antiracist education principles	What characteristics of the assessment, the assessment design process, and/or the inferences drawn from the assessment provide evidence of antiracism?
Quantitative critical race theory	In what ways have systemic oppression been investigated in all stages of assessment development (including the ways in which the assessment itself may perpetuate these systems)?

IUA: interpretation/use argument.

employed inappropriately) must be critically examined; and (4) social justice aims should be the foundation for all methodological decisions and approaches. In summary, QuantCrit pushes back against the very notion that we can come to know some objective truth if our quantitative methodologies are, in fact, rigorous enough. Randall et al. (2023) write that “Data do not speak for themselves. Researchers speak on top of them with all of the values, biases, and assumptions that we carry with us.” According to Gilborn et al. (2018), “Numbers’ authoritative facade often hides a series of assumptions and practices which mean, more often than not, that statistics will embody the dominant assumptions that shape inequity in society” (p. 175). A justice-oriented approach to the validation process neither assigns greater value to nor dismisses out of hand numbers (Gilborn et al., 2018); rather it encourages us to critically interrogate (as with all data) their creation, use, and interpretation.

So, what does this justice-oriented, antiracist approach to validation look like from construct articulation to score Interpretation? Table 1 includes the questions that we posit should be considered broadly speaking. We propose a proactive, intentional approach to validation—one that centers the most marginalized from the beginning—as opposed to a retroactive approach to equity in validation that builds itself on whiteness and then attempts to retrofit for racially and ethnically minoritized students (often time force fitting to the point of breaking).

Construct articulation

Relying on Mislevy’s (2018) sociocognitive perspective that requires a thoroughly developed construct that is understood from the perspective of every stakeholder, or rights holder, we submit that constructs are not, in fact, monolithic with one truth. This means, within a critical framework, we must acknowledge the systems of oppression that seek to convince us otherwise. White supremacist logics would have us believe that the ways of knowing and understanding of racially and ethnically minoritized peoples are somehow less true, less accurate, and/or less legitimate. For example, with respect to writing, Inoue (2021) has argued that so-called race neutral writing standards demonstrate the habits of White language (HOWL) and judgment:

These are the language habits usually assumed or promoted as universally appropriate, correct, or best in writing and speaking by those with power to do so. Historically, these habits of language have come out of elite White racial groups in Western, monolingual, English speaking societies . . . There is nothing inherently racist about these habits of language. However, when they are used as universal standards for communication, used to bestow opportunities and privileges to people, then they become racist and produce White language supremacy (pp. 22-23)

In a traditional test design approach, a designer might say that a persuasive essay or statement of interest for a graduate program intended to measure one's ability to write requires a well-defined description of the knowledge, skill, and dispositions that make up the construct of writing ability. A justice-oriented approach would go on to ask "Why an essay?" "What is meant by persuasive?" and "What linguistic, cultural, and substantive patterns influence every stakeholder's perception (i.e., examinees, raters, task developers) of the writing task?" What it means to be persuasive is highly variable across cultures and communities. And how examinees take up the very idea of any task will be deeply shaped by their LCS backgrounds. For example, some examinees in East Asia, who have not attended Western-style high schools, will likely respond differently than examinees from South Africa or [more differently still] the United States. And stylistic differences will also differ within these populations depending on a wide range of cultural and social norms and expectations (e.g., rules/expectations with respect to humility), all mediated by who the respondent perceives to be the intended audience (Tan et al., 2022). To articulate the writing ability construct in such a way that privileges responses, which reflect the cultural norms and rules of engagement of White elite Americans will inevitably result in inaccurate (and typically deficit-oriented) interpretations about what other populations know and are able to do.

Constructs are, by their very nature, culturally and socially constructed. Constructs change, and how both examinees and other stakeholders understand the same constructs is highly situated and contingent. Indeed, within a QuantCrit framework, we recognize that we must set aside any notion that constructs—however articulated—represent some kind of objective truth that is not influenced by the experiences, knowledges, and values individuals bring with them to the enterprise of assessment. And then, if justice is the goal, we must begin the process of centering the experiences, knowledges, and values of these most marginalized populations. The goal is not to "accommodate" difference, but to make difference the very seed that animates assessment design, score interpretation, and consequence.

This is how your validity begins to look like justice.

Sources of validity evidence

In the following section, we refer to the five sources of validity evidence as outlined in the latest version of the Standards for Educational and Psychological Testing available at the time of writing (AERA et al., 2014). We articulate the high-level questions typically addressed when gathering validity evidence with respect to each source and describe how these questions could, and should, be re-oriented to reflect a justice-orientation.

Table 2. Systemic rules of English language varieties.

White Mainstream English	Black English	Chicano English
Feature pattern: Present perfect tense is uncommon	Feature pattern: Remote past marker	Feature pattern: Double and multiple negation
Example: I just went to the library. (Versus British English, "I've just gone to the library.")	Example: I been went to the library. (Versus: "I went to the library.")	Example: I don't want to go to no library. (Versus: "I don't want to go to the library.")
Purpose: Makes the difference between past events and past events that are still in-process	Purpose: To emphasize and to indicate something was done a while ago or has been going on for a while	Purpose: Emphasizes the intention of the speaker. In addition, double-negatives are grammatically correct in Spanish. Because Chicano English draws on features of Englishes and Spanishes, this feature comes from Spanish

Credit: Pine, A. & Hartwell, K. (2023).

Content validity

When thinking about validity evidence with respect to content, we must re-envision the entirety of our thinking—especially as it relates to language, reading, and writing. For example, when discussing validity evidence with respect to content, justice does more than determine if the test items represent the targeted domain of interest, but goes further and asks, for example, “Are there test items that actively disrupt negative stereotypes about minoritized populations?” Does the content/language of the items privilege a particular linguistic or cultural way of thinking/making sense of the world? It is not enough to avoid the inclusion of items, or illustrations in your passages, that might imply that Black families are less stable, you need to bombard test takers with scenarios and images of Black stability. It is not enough to avoid images of Central and South Americans engaging in questionable behavior. You must develop reading passages that invoke their good citizenship and dignity. And, most importantly, it will never be enough for item writers to craft items that do NOT imply that the language spoken by Black Americans is broken, common, informal, or unintelligible. You need to craft items/scenarios in which the linguistic system employed by Black Americans is described as systematic, sophisticated, and nuanced with rules like the ones you see in Table 2. And, as importantly, you must revise scoring descriptions that reflect this understanding. In other words, it is not enough to support the ideals of plurilingualism (e.g., writing scenarios employing marginalized Englishes) while simultaneously relying on old scoring models (e.g., rubrics) that enforce the same White supremacist logics.

Because of the United States’ long history of perpetuating negative stereotypes about racial inferiority (from the ways in which these populations raise their children to the

ways they have improved the so-called standard English language), everyone involved in the assessment enterprise, in the United States and also everyone around the world, needs to go about the business of crafting a future that actively disrupts these stereotypes. It is not enough to avoid repeating them. They must be called out saying boldly, “These deficit assumptions are a lie”; and we can use the content of our assessments to say it louder.

This is how your validity begins to look like justice.

Internal processes

When gathering validity evidence with respect to internal structure of an assessment, traditional approaches ask: Does the relationship among test items and test components conform to the construct? A justice-oriented, antiracist approach asks: How have values shown up in the item/tasks? And which social identity groups do these values reflect? When we develop tasks intended to get at a particular construct such as self-regulation, executive functioning, writing, and so on—have we considered that these constructs can (and do) look different across identities and have we privileged, for example, the values associated with White middle class identity when examining that structure?

Have we failed to consider certain sociocultural identities and are we, in turn, seeing that items intended to measure self-regulation—only do so in White kids—and actually get at the extent to which schools are policed, or surveilled, in other kids? For example, for Black students, have we considered spirituality, harmony, movement, verve, affect, and communalism (Boykin, 1986); and have those values shown up in the tasks intended to measure the construct? Can a thing be unidimensional within a certain set of cultural values and multidimensional within another set? Possibly. We do not know. Our point is that we need to be asking these questions—whose values are showing up as we investigate the internal structure—with whatever methods we employ—and whose values are being ignored? And then adjust accordingly.

This is how your validity begins to look like justice.

Response processes

Evidence of response processes would go beyond simply determining if the test takers are interpreting the tasks as intended. A justice-oriented, antiracist approach would ask, for example, “What historical logics of testing and racism are students bringing to the test situation?” To be sure, failure to comprehensively understand the various historical and sociocultural factors (to include racist systems of oppression) influencing examinee response processes and performance can lead to erroneous, deficit-oriented, score interpretations. Indeed, stereotype threat (Brown & Day, 2006; Steele, 1997; Steele & Aronson, 1995) and race-based traumatic stress (Mental Health America, n.d.) serve as useful examples here of the pernicious impact of the persistent, negative, racist narratives on the internal processes of minoritized students. The frequent failure of scholars to even acknowledge the possibility of the impact of racial discrimination and violence on human cognition reflects the field’s long history of ignoring the broader sociocultural context (a context that the examinee cannot simply disentangle themselves from) when investigating response processes.

A justice-oriented, antiracist approach would also ask, “What linguistic, cultural, and substantive patterns are students bringing to each assessment task? How do those patterns, either independent of or through interaction with the construct, shape the ways in which students interpret and respond to each task? How are differences in patterns across diverse populations of students in task design, scoring criteria, and inferences being made from performance data?” If justice is at the center of our validation processes, we have to ask, what are marginalized students bringing into the testing space with them and how does it impact their response processes?

This is how your validity begins to look like justice.

Relations to other variables

When thinking about the relationship between the assessment and other variables, an antiracist approach to validation asks questions like “How are criterion variables selected? Does this selection process consider the history/impact/legacy of white supremacist hegemonic practices? Does this process of criterion selection seek to disrupt these hegemonic practices?” We have often used the phrase, garbage in, garbage out; the same applies here: White supremacist criterion, White supremacist validity argument. If every criterion we are using is cloaked in White supremacist logics, then the relationships will be strong. Indeed, most analytic methods simply employ a network of deficit-framed data that ultimately serve to further marginalize entire populations of students. For example, when building validity evidence with respect to relations to external variables for large-scale assessments, relying primarily (and uncritically) on student scores on other large-scale standardized assessments as the external criteria can be problematic if the goal is to suss out White supremacist logics in the assessment: correlations between data generated by two flawed instruments do not constitute a meaningful, compelling, or indeed, just validity argument. The same can be said for criteria such as students’ grades (see Malouff & Thorsteinsson, 2016, for an example of how minoritized students receive lower grades than White students for similar work) and letters of recommendation. The point is that White supremacist and racist logics can simultaneously and independently inform the development of multiple measures/criteria, so it is important to interrogate with intention all of our criteria for those logics before uncritically employing them to support evidence of relations to external variables. Again, we are not suggesting that we throw out all analyses with respect to concurrent or predictive validity. Rather, we are suggesting that we think about the ways systemic oppression can impact all of these relationships. That is, if the goal is for

your validity to start to look like justice.

Consequences

Finally, from a justice perspective, we argue that validity evidence related to consequences must be examined and reexamined continuously. Any assessment system that does not seek out evidence of and then disrupt negative consequences in marginalized

communities is inherently, and without question, unjust. When addressing validity evidence related to consequences, justice framing requires one to ask, for example, do test/assessment results serve to further marginalize minoritized populations? What groups will be privileged by this assessment?—in the short term? And in the long term? And what system would have to be in place for a student to be successful on the assessment? How can we meaningfully address disparate impact (Poe & Cogan, 2016)? Are those systems rooted in White supremacist values? As the final link in any IUA validity argument, consideration of consequences—including the examination of disparate impact—must be predicated on a critical examination of every aspect of an assessment's design and use. When this does not happen, racist logics can lead to brutally racist interpretations of that data. One example is the situation when an anonymous group emailed all Black student groups (as well as individual Black students) at the University of Massachusetts citing Black students' Scholastic Aptitude Test (SAT) scores as evidence that those students were not fit to be enrolled in the predominately White institution:

Herein lays the problem with your presence at our college, you simply did not get here on merit. Believe it or not students are not the only who think this and know this (it is a common fact that a Black person can score hundreds of points lower on the SAT) but also professors and TAs (Press-Reynolds, 2021).

This letter represents an excellent example of the further marginalization of an already marginalized group—and it is the kind of outcome/consequence that could be predicted, because it keeps happening. We are referring to short-term marginalization (being denied admission to higher education or ridiculed when admitted) and long-term marginalization (limited career prospects). And all of this translates to short-term and long-term privilege for a very particular type of student.

Add to this the consideration of what systems have to be in place for a student to be successful and who has access to those systems. Most assessment professionals are aware of the numerous test preparation opportunities (none of which are free; at least not the good ones) that increase the odds of success for some students (e.g., typically middle, upper middle class and White). Randall (2021) writes: “It is easy to declare that the problem is in social and economic inequities and not the test itself and walk away, while simultaneously creating conditions that allow a multi-billion-dollar test preparatory industry targeted at/for White and Asian students to flourish” (p.34). And even more egregious is the systematic creation of conditions that allow, and, in many cases, encourage the off-label uses of large-scale standardized assessments (e.g., the American College Testing [ACT] as a graduation requirement or SAT scores used for course placement), with absolutely no attention given to the negative consequences of those uses. The point is any approach to validation that does not consider the consequences of the assessment and adjust accordingly when evidence of harm becomes apparent is inherently and indisputably unjust.

Conclusion

We will end by saying that Botha (2021) wrote that “dehumanization is defined as the denial of full humanness to others, the denial of a group's community or identity,

exclusion of a group from moral boundaries, the denial of a group's ability to experience complex emotions, or the denial of specific traits which are said to unite all humans, or separate non-human animals from humans . . . Dehumanization and exclusion from moral boundaries serve to facilitate the permissibility of violence against a group . . .” (p. 4). We have described the ways in which assessments have served as a tool of dehumanization and violence in racially and ethnically minoritized communities. And, in articulating a justice-oriented, antiracist approach to validation, we have attempted to present a path forward that re-humanizes these populations in the assessment process.

Indeed, in this Viewpoint contribution, we have offered as many questions as we have solutions, but the most important solution is simply to ask the questions. Our main point, we hope, is clear, which is: if justice is the goal, then it has to be placed at the center (and not an afterthought) in any validation process. This means we have to stop relying on checklists looking for evidence of bias and really start interrogating our processes and content for evidence of justice. We do not wish to imply that we have solved the problem that is assessment injustice. Rather, we are saying as emphatically as we can with mere words on a page that many, if not all, language assessment systems represented in the United States and abroad contribute to injustice—and that must change.

Authors' note

This Viewpoint piece (position paper) is based on Jennifer Randall's Samuel J. Messick Memorial Lecture plenary session, presented on 7 June 2023 at the 44th *Language Testing Research Colloquium* in New York, NY. Her collaborators have been included as co-authors, having also contributed to work on which this manuscript is based.

Author contributions

Jennifer Randall: Writing—original draft; Writing—review & editing.

Mya Poe: Writing—original draft; Writing—review & editing.

David Slomp: Writing—original draft; Writing—review & editing.

Maria Elena Oliveri: Writing—original draft; Writing—review & editing.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Jennifer Randall  <https://orcid.org/0000-0003-4728-8710>

Mya Poe  <https://orcid.org/0000-0002-3349-9093>

Note

1. Readers should refer to Poe et al. (2023), Randall et al. (2021), and Randall et al. (2022) for more comprehensive descriptions and examples of assessment justice.

References

- American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (2014). *Standards for educational and psychological testing*.
- Bell, D. (1995). Who's afraid of critical race theory? *University of Illinois Law Review*, 4, 893–910.
- Bonilla-Silva, E. (2013). *Racism without racists: Color-blind racism and the persistence of racial inequality in America*. Rowman & Littlefield. <https://doi.org/10.2307/40264552>
- Botha, M. (2021). Academic, activist, or advocate? Angry, entangled, and emerging: A Critical reflection on autism knowledge production. *Frontiers on Psychology*, 12, 727542. <https://doi.org/10.3389/fpsyg.2021.727542>
- Boykin, A. W. (1986). The triple quandary and the schooling of African American children. In U. Neisser (Ed.), *The school achievement of minority children: New perspectives* (pp. 57–92). Lawrence Erlbaum Associates.
- Brown, R., & Day, E. (2006). The difference isn't black and white: Stereotype threat and the race gap on Raven's Advanced Progressive Matrices. *Journal of Applied Psychology*, 91(4), 979–958. <https://doi.org/10.1037/0021-9010.91.4.979>
- Delgado, R., & Stefancic, J. (2017). *Critical race theory: An introduction*. New York University Press. <https://doi.org/10.2307/j.ctt1ggijn3>
- DuBois, W. E. B. (2017). The Propaganda of history. In *Black reconstruction: An essay toward a history of the part which Black folk played in the attempt to reconstruct democracy in America, 1860–1880* (pp. 635–651). Routledge. (Original work published 1935). <https://doi.org/10.4324/9781315147413>
- Dunn, D. C., Chisholm, A., Spaulding, E., & Love, B. L. (2021). A Radical doctrine: Abolitionist education in hard times. *Educational Studies*, 57(3), 211–223. <https://doi.org/10.1080/00131946.2021.1892684>
- Epp, C., Maynard-Moody, S., & Haider-Markel, D. (2014). *Pulled over: How police stops define race and citizenship*. University of Chicago Press. <https://doi.org/10.1080/10967494.2018.1432193>
- Finn, J., & Servoss, T. (2014). Misbehavior, suspensions, and security measures in high school: Racial/ethnic and gender differences. *Journal of Applied Research on Children*, 5(2), 11. <https://doi.org/10.58464/2155-5834.1211>
- Fisher, B. W., & Hennessy, E. A. (2016). School resource officers and exclusionary discipline in U.S. high schools: A systematic review and meta-analysis. *Adolescent Research Review*, 1, 217–233. <https://doi.org/10.1007/s40894-015-0006-8>
- Gilborn, D., Warmington, P., & Demack, S. (2018). QuantCrit: Education, policy, “Big Data” and principles for a Critical Race Theory of statistics. *Race, Ethnicity, and Education*, 21(2), 158–179. <https://doi.org/10.1080/13613324.2017.1377417>
- Gottfredson, D. C., Crosse, S., Tang, Z., Bauer, E., Harmon, M., Hagen, C., & Greene, A. (2020). Effects of school resource officers on school crime and responses to school crime. *Criminology and Public Policy*, 19(3), 905–940. <https://doi.org/10.1111/1745-9133.12512>
- Haynes, N. (2023). *Affirmative action myths and the misuse of test scores in college admissions*. Preprint. <https://www.researchgate.net/publication/372079182>
- 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>
- hooks, b. (1994). *Teaching to transgress: Education as the practice of freedom*. Routledge. <https://doi.org/10.4324/9780203700280>
- Inoue, A. B. (2015). *Antiracist writing assessment ecologies: Teaching and assessing writing for a socially just future*. Parlor Press. <https://doi.org/10.37514/PER-B.2015.0698>

- Inoue, A. B. (2021). *Above the well: An antiracist literacy argument from a boy of color*. Utah State University Press. <https://doi.org/10.37514/PER-B.2021.1244>
- Kane, M. (2010). Validity and fairness. *Language Testing*, 27(2), 177–182. <https://doi.org/10.1177/0265532209349467>
- Kane, M. T. (2013). Validating the interpretation and uses of test scores. *Journal of Educational Measurement*, 50, 1–73. <https://doi.org/10.1111/jedm.12000>
- Kane, R. (2002). The Social ecology of police misconduct. *Criminology*, 40, 867–896. <https://doi.org/10.1111/j.1745-9125.2002.tb00976.x>
- LeClair, T. (1981, March 21). 'The language must not sweat.' *The New Republic*. <https://newrepublic.com/article/95923/the-language-must-not-sweat>
- Lederman, J. (2023). Validity and racial justice in educational assessment. *Applied Measurement in Education*, 36(3), 244–254. <https://doi.org/10.1080/08957347.2023.2214654>
- Lee, C. (1998). Culturally responsive pedagogy and performance-based assessment. *The Journal of Negro Education*, 67(3), 268–279. <https://doi.org/10.2307/2668195>
- Lyons, S., Hinds, F., & Poggio, J. (2021). Commentary: Evolution of equity perspectives on higher education admissions testing: A call for increased critical consciousness. *Educational Measurement: Issues and Practices*, 40(4), 41–43. <https://doi.org/10.1111/emip.12458>
- Malouff, J., & Thorsteinsson, E. (2016). Bias in grading: A meta-analysis of experimental research findings. *Australian Journal of Education*, 60, 425–256. <https://doi.org/10.1177/0004944116664618>
- Mental Health America (n.d.) Racial trauma. <https://www.mhanational.org/racial-trauma>
- Mills, C. W. (2017). *Black rights/white wrongs: The critique of racial liberalism*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780190245412.001.0001>
- Mislevy, R. (2018). *Sociocognitive foundations of educational measurement*. Routledge. <https://doi.org/10.4324/9781315871691>
- Montenegro, E., & Jankowski, N. (2017). *Equity and assessment: Moving towards culturally responsive assessment*. Occasional Paper #29. National Institute for Learning Outcomes Assessment. <https://files.eric.ed.gov/fulltext/ED574461.pdf>
- Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, 62, 229–258. <https://doi.org/10.2307/1170738>
- 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. *Journal of Writing Analytics*, 5, 71–108. <https://doi.org/10.37514/JWA-J.2021.5.1.03>
- Pine, A., & Hartwell, K. (2023, April 18). *Language, discrimination, and the English case: A cross-curricular assessment for 9th grade* [Class presentation]. Contemporary Issues in Assessment, University of Michigan, Ann Arbor, MI, United States.
- 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), 08f1c307. <https://escholarship.org/uc/item/08f1c307>
- Poe, M., Oliveri, M., & Elliot, N. (2023). The standards will never be enough: A racial justice extension. *Applied Measurement in Education*, 36(3), 193–215. <https://doi.org/10.1080/08957347.2023.2214656>
- Press-Reynolds, K. (2021, October 6). UMass Amherst has hired a cybersecurity firm to investigate emails reportedly sent to Black students. *Insider*. <https://www.insider.com/umass-amherst-racist-emails-black-student-groups-cybersecurity-university-massachusetts-2021-10>
- Qualls, A. (1998). Culturally responsive assessment: Development strategies and validity issues. *The Journal of Negro Education*, 67(3), 296–301. <https://doi.org/10.2307/2668197>

- 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. (2023, 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, United States.
- Randall, J., Oliveri, M., Poe, M., & Slomp, D. (2023). Justice-oriented validation: Continuing to disrupt white supremacy in assessment practices [Manuscript Submitted for Publication].
- Randall, J., Poe, M., & Slomp (2021). Ain't ought to be in the dictionary: Getting to justice by dismantling anti-Black literacy assessment practices. *Journal of Adolescent & Adult Literacy*, 64(5), 594–599.
- Randall, J., Slomp, D., Poe, M., & Oliveri, M. (2022). Disrupting white supremacy in assessment: Toward a justice-oriented, antiracist validity framework. *Educational Assessment*, 27(2), 170–178. <https://doi.org/10.1080/10627197.2022.2042682>
- Rawls, J. (1971). *A theory of justice*. Harvard University Press. <https://doi.org/10.4159/9780674042605>
- Rawls, J. (1999). *A theory of justice* (Revised ed.). Harvard University Press. <https://doi.org/10.4159/9780674042582>
- Research for Action. (2020). *School policing in Pennsylvania: Prevalence and disparities*. Police in Pennsylvania Schools Series. <https://files.eric.ed.gov/fulltext/ED608042.pdf>
- Shepard, L. A. (2021). Ambitious teaching and equitable assessment: A vision for prioritizing learning, not testing. *American Educator*, 45(3), 28–48. <https://www.aft.org/ae/fall2021/shepard>
- Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Pearson. <https://doi.org/10.4324/9781003062318>
- Slomp, D. (2016). An integrated design and appraisal framework for ethical writing assessment. *Journal of Writing Assessment*, 9(1), 4bg9003k. <https://escholarship.org/uc/item/4bg9003k>
- Smith, D. (1986). The neighborhood context of police behavior. In A. Reiss & M. Tonry (Eds.), *Crime and Justice* (Vol. 8, pp. 313–331). University of Chicago Press. <https://doi.org/10.1086/449126>
- Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, 52, 613–629. <https://doi.org/10.1037//0003-066x.52.6.613>
- Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, 69, 797–811. <https://doi.org/10.1037//0022-3514.69.5.797>
- Stetler, C. (2009, October 26). Junot Diaz: Man in the mirror. *True Jersey*. https://www.nj.com/entertainment/arts/2009/10/junot_diaz_man_in_the_mirror.html
- Tan, T. X., Fan, X., Braunstein, L. B., & Lane-Holbert, M. (2022). Linguistic, cultural and substantive patterns in L2 writing: A qualitative illustration of Mislevy's sociocognitive perspective on assessment. *Assessing Writing*, 51, 100574. <https://doi.org/10.1016/j.asw.2021.100574>
- U.S. Department of Education, Office of Civil Rights. (2023, March). *Corporal Punishment In Public Schools*. <https://www2.ed.gov/about/offices/list/ocr/docs/corporal-punishment-part-4.pdf>
- Vincent, N. (1992, February). *The philosophy and politics of multicultural education and anti-racist education: An analysis of current literature*. Paper Presented at the Annual Meeting of the National Association for Multicultural Education, Orlando, FL. <https://files.eric.ed.gov/fulltext/ED346022.pdf>

- Weiler, S. C., & Cray, M. (2011). Police at school: A brief history and current status of school resource officers. *The Clearing House*, 84(4), 160–163. <https://doi.org/10.1080/00098655.2011.564986>
- Whitaker, A., Torres-Guillén, S., Morton, M., Jordan, H., Coyle, S., Mann, A., & Sun, W.-L. (2019). *Cops and no counselors: How the lack of school mental health staff is harming students*. ACLU. https://www.aclu.org/sites/default/files/field_document/030419-acluschooldisciplinereport.pdf
- Woodson, C. G. (1999). *The mis-education of the Negro*. 12th Media Services. (Original work published 1933)