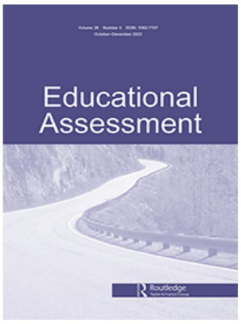

Routledge
Taylor & Francis Group

Educational Assessment

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/heda20

Justice-Oriented, Antiracist Validation: Continuing to Disrupt White Supremacy in Assessment Practices

Jennifer Randall, Mya Poe, Maria Elena Oliveri & David Slomp

To cite this article: Jennifer Randall, Mya Poe, Maria Elena Oliveri & David Slomp (22 Nov 2023): Justice-Oriented, Antiracist Validation: Continuing to Disrupt White Supremacy in Assessment Practices, Educational Assessment, DOI: 10.1080/10627197.2023.2285047

To link to this article: https://doi.org/10.1080/10627197.2023.2285047

Published online: 22 Nov 2023.

EDUCATIONAL ASSESSMENT
https://doi.org/10.1080/10627197.2023.2285047

Routledge
Taylor & Francis Group

Justice-Oriented, Antiracist Validation: Continuing to Disrupt White Supremacy in Assessment Practices

Jennifer Randall[a], Mya Poe[b], Maria Elena Oliveri[c], and David Slomp[d]

[a]University of Michigan; [b]Northeastern University; [c]University of Nebraska; [d]University of Lethbridge

ABSTRACT

Traditional validation approaches fail to account for the ways oppressive systems (e.g. racism, radical nationalism) impact the test design and development process. To disrupt this legacy of white supremacy, we illustrate how justice-oriented, antiracist validation (JAV) framework can be applied to construct articulation and validation, data analysis, and score reporting/interpretation phases of assessment design/development. In this article, we use the JAV framework to describe validation processes that acknowledge the role and impact of race/racism on our assessment processes—specifically construct articulation, analysis, and score reporting—on Black, Brown, Indigenous, and other students from historically marginalized populations. Through a JAV framework, we seek to disrupt inaccurate white supremacist approaches and interpretations that for too long have fuelled measurement practices.

Keywords

Antiracist validity; justice-oriented validity; antiracist assessment

Current practice in assessments seems to treat validity through assumptions, or rather, through assumptive bias. Commonality or homogeneity or difference and heterogeneity are simply assumed these days. However, the question remains: how are these assumptions influenced by racism, sexism, classism, or, in the case of language, nationalism.
-Gordon (1995), p.363

Introduction

Today, educational measurement researchers have largely accepted the need for culturally responsive approaches to assessment (Hood, 1998; Lee, 1998; Montenegro & Jankowski, 2017; Moss, 1992; Qualls, 1998; Shepard, 2021) which attempt to address the socio-cultural influences (e.g., language backgrounds, cultural traditions, and a range of other identity formations) that shape the ways students respond to test items. What the field of measurement, however, has not done yet is explicitly addressed the historical and ongoing discriminatory legacies of measurement on Black, Brown, Indigenous, and other students from historically marginalized populations.[1] These are legacies that cannot simply be ignored or papered over with methods for accommodation. To address the history of discrimination in measurement requires grappling with the very assumptions that lie beneath responsiveness and accommodation. In the U.S. context, that history means reckoning with the design and misuse of tests or what Asa Hilliard III called a "history of abuse of minorities and the poor" (2000, p. 296).

In the case of anti-Black discrimination, measurement researchers have to grapple with the ways test design has long been intertwined with larger social struggles. For example, in challenging

CONTACT Jennifer Randall jennrand@umich.edu University of Michigan

[1]We use the phrase "Black, Brown, Indigenous, and other students from historically marginalized populations" with caution. All terms, including BIPOC, students of color, and minoritized students, that seek to capture the range of historically oppressed groups in the U.S. are always partial, contested, and shift over time. For example, we understand that Asian Pacific Islander Desi American (APIDA) and Middle Eastern/North African (MENA) students are not specifically named in our chosen term in this chapter.

desegregation, Black Americans also had to confront admissions testing and standards (See Nettles, 2019). The stories of Heman Marion Sweatt, Ada Lois Sipuel, George W. McLaurin, Lloyd Gaines, Pollie Myers, Autherine Lucy, Vivian Malone, James Hood, and Edith Mae Irby reveal the lie that meeting admissions criteria is grounds for acceptance. And Lloyd Gaines's story is a chilling reminder that challenging discrimination in admissions standards can be lethal (Palmer, 2018).

Today, we also refer to the effects of widespread and unchecked systems of oppression rooted in racist logics. For example, Black and Hispanic communities have historically been subjected to aggressive, over-policing practices such as coercion, unjustified stops, verbal abuse, and excessive force (Epp et al., 2014; Gottfredson et al., 2020; Kane, 2002; Smith, 1986). This over-policing extends to schools (see Research for Action, 2020; Weiler & Cray, 2011), resulting in exclusionary discipline practices that disproportionately impact Black, Brown, and Indigenous students (Finn & Servoss, 2014; Fisher & Hennessy, 2016, U.S. Department of Education OCR, 2016; Whitaker et al., 2019). These inequities reveal themselves at every stage of the schooling cycle. As de Brey et al. (2019) shows, Black students are less likely (23%) to earn Advanced Placement or International Baccalaureate credits than White (40%) and Asian (72%) students. Moreover, Black students are more likely (2.7%) to be retained than White students (1.7%), and less likely (36%) to enroll in college than White (42%), Hispanic (39%), or Asian (58%) students. Among college enrolling students, studies have shown strong links between developmental course-taking and the racial composition of the high school (Howell, 2011) and students' race (Adelman, 2004; Attewell et al., 2006; Radford & Horn, 2012; Snyder & Dillow, 2012). Often based on placement test scores, Black (66%) and Hispanic (53%) students are placed in remedial courses (see Gilman, 2019; Nastal, 2019; Ngo & Melguizo, 2020) more frequently than White (36%) students (Chen, 2016). While some may argue that students need additional support for math and literacy, it has been shown that too often placement tests misplace students (Scott-Clayton et al., 2014). For too many students, misplacement becomes an academic death sentence (Klausman & Lynch, 2022).

While new methods such as ungrading and multiple measures assessment may open new methods for classroom assessment and placement, historically, these differential rates of completion, enrollment, and course remediation have been explained through deficit narratives, which view minoritized populations as inferior, without examining how social and structural barriers (e.g., requirements for sequential course completion) work to create such outcomes. These systemic inequities (e.g., differential enrollment and graduation rates among undergraduate student groups in the United States, participation in remedial courses, etc.) represent the context in which assessment developers, psychometricians, and policy makers do their work, and should not be ignored as they have implications with respect to how assessments are designed and used (Oliveri et al., 2020).

In the previous work, we have argued that measurement researchers are continuing to produce racist instruments "if the logics that inform test design and the interpretation of test scores draw from underlying logics of racism that devalue, omit, suppress, and marginalize nonwhite students" (Randall et al., 2022, p. 5). For example, writing assessments that penalize students for composing in African-American Vernacular English (AAVE) or math assessments that require students to complete problem sets in ways that reflect the mathematical algorithms of the Greeks (to the exclusion of the approaches propagated in Indigenous populations, S.E. Asia, & Africa) demonstrate a lack of awareness of the wider ways that students compose and solve math problems. By designating only one way to compose or solve math, such instruments presuppose that the ways of communicating, knowing, and understanding produced in Black, Brown, and Indigenous populations are inherently inferior (based on racist, white supremacist logics).

In response to such myopic visions of knowledge production and learning, we proposed a justice-oriented, antiracist validity (JAV) approach (Randall et al., 2022) to acknowledge and remedy the discriminatory legacies against racially minoritized students in U.S. educational assessment.[2] Built on

[2]Our work aligns with other educational scholars who have called for the interruption of injustice in educational research (e.g., Souto-Manning & Winn, 2017).

foundational theories and practices related to justice-oriented perspectives; Critical Race Theory; antiracist assessment; Kane's (2013) Interpretation Use Argument (IUA) validation model; and Mislevy's (2018) sociocognitive extension of that model, JAV offers a set of critical questions about construct articulation and validity evidence in building a validity argument. A JAV approach builds on culturally affirming assessment approaches while also demanding attention to the ways validity claims may conceal racist logics (e.g., defining constructs in a way that assumes standard edited American English is a more legitimate linguistic system than AAVE). For example, resonant with culturally responsive assessment (Hood, 1998; Lee, 1998; Qualls, 1998), JAV recognizes the richness of students' identities while also acknowledging the ways in which racial injustice (fueled by white supremacist logics) encapsulates the lived experiences of all students. This context of injustice cannot be disaggregated from (or ignored by or through) the assessment experience.

We employ the JAV frame across three critical stages of the educational assessment process to address specifically (a) the ways in which white supremacist logics show up in how we define our constructs resulting in inadequate (i.e., flawed, underrepresented) articulations of the construct, (b) how uncritical approaches to data analysis—especially those centered on discrete variables—fail to capture nuances within and across minoritized populations; and (c) how both (a) and (b) contribute to the perpetuation of deficit-framed interpretations of minoritized student performance which fail to appropriately reflect what these students know and are able to do, which is the ultimate purpose of assessment.

Quantitative critical race theory

In this article, our objectives are to build on the work of Randall et al. (2022) to describe and propose specific approaches to building a JAV argument. The authors, using Kane's IUA, Mislevy's sociocognitive extension, antiracist principles, and Critical Race Theory provided a framing that seeks to disrupt white supremist notions of knowledge and knowledge-making, while simultaneously affirming the ways of knowing and understanding of racially minoritized populations during the test development process. In our application of JAV, we rely on the principles of Critical Race Theory broadly (as outlined in Randall et al., 2022) and also specifically with respect to quantitative analyses (quantitative Critical Race Theory, QuantCrit). QuantCrit, which applies Critical Race Theory's primary tenets (Bell, 1995; Delgado & Stefancic, 2017) to quantitative methodologies provides a useful framework for methodological decisions in this antiracist context. QuantCrit (Solórzano & Ornelas, 2002; Solórzano & Villalpando, 1998; Solórzano & Yosso, 2002) approaches maintain, as with Critical Race Theory, that (1) racism is not an aberration, but rather a permanent and typical aspect of everyday life under which we all labor; (2) acknowledging the centrality of oppression through/within our economic, political, and educational systems (including ableism, sexism, religious oppression) is important; (3) numbers are not infallible or neutral and that they can – and historically have been used (and manipulated) to perpetuate racist ideologies; (4) we must critically examine our use of categories constructed around race; (5) the voices and insights of communities of color in the data decisions/determination, collection and interpretation stages (and care should be given to identify the appropriate, critical voices) should be included and not ignored; and (6) all methodological decisions and approaches should be of service to social justice aims.

QuantCrit rejects the ontological perspective that, through the use of rigorous quantitative methodologies, we can come to know some objective truth; a truth that is not at all influenced/impacted by the subjectivity of the researcher (Zuberi & Bonilla-Silva, 2008). 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. QuantCrit approaches acknowledge that numbers do not have more inherent value (in that they are more objective, factual, or real) than qualitative data detailing the stories and experiences of populations of people. Gillborn et al. (2018) write that "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). From a QuantCrit perspective, we are not arguing

that the field of measurement should dispense of quantitative approaches to data analysis and interpretation, but rather "adopt a position of principled ambivalence, neither rejecting numbers out of hand nor falling into the trap of imagining that numeric data have any kind of enhanced status, value, or neutrality" (Gillborn et al., 2018, p. 174). In other words, we argue that QuantCrit principles help advance validation processes for diverse populations because they demand more critical approaches to data use and interpretation.

QuanCrit principles help advance validation processes for diverse populations in each of the three areas we identified at the beginning of this article: construct articulation and validation, data analysis, and data interpretation/score reporting. First, we argue, as have others, that construct articulation and validation are foundational phases of assessment development (Kane, 1992; Messick, 1980; Moss, 1995); Second, because the field of measurement relies heavily on quantitative data analysis during the validation process (e.g., to correlate measures, identify score differences, calculate p-values, etc.), these processes must reflect a critical, justice-oriented lens. Finally, there is a long and extensive literature on the importance of accurate and interpretable score reporting (Goodman & Hambleton, 2010). In fact, it is arguably the most public-facing phase of the assessment design and development process. It impacts policy decisions (both local, state, and federal) and personal decisions (including home buying decisions, Brasington & Haurin, 2006). Indeed, regardless of the validation framework a researcher/practitioner chooses to operate within, these three phases must occur—construct articulation (the beginning), score interpretation (the end), and analysis (the connective tissue between the two)—and all require a justice-oriented, antiracist critical lens.

Consequently, our discussion of procedural transformations using QuantCrit focuses on these three stages of assessment design and development: (1) construct articulation and validation (i.e., a re-envisioning of construct definition and representation to include ways of knowing and understanding that extend beyond privileging or centering whiteness, and instead, acknowledge, affirm, and seek to sustain the sociocultural identities of the most marginalized populations); (2) critical analysis (i.e., a shift away from an over-reliance solely on so-called objective variable-centered statistical methods to more person-centered approaches that consider the wider sociopolitical and historical context in which the numbers exist); and (3) interpretation (i.e., a new commitment to an antiracist approach to score/performance interpretation that is rooted in asset-based, as opposed to deficit-based, perspectives of marginalized populations).

I: construct articulation and Validation[3]

As the *Standards for Educational and Psychological Testing* have evolved, the role of construct validation has changed with it. In 1955, *The Technical Recommendations for Achievement Tests* (AERA & NCMUE, 1955) treated construct validation as an afterthought, a fall-back option, recommending that "construct validation should be invoked when the preceding three methods [content validation, concurrent validation, and predictive validation] are insufficient to indicate the degree to which the test measures what it intends to measure" (p 16). By 1989, Messick's unitary view of validity placed the construct at the center of the validation process. Yet, even then, the *Standards for Educational and Psychological Testing* (AERA & NCME, 1999) permitted significant latitude in how robust construct development and articulation needed to be, stating, "[t]he construct of interest should be embedded in a conceptual framework, no matter how imperfect that framework may be" (p 9). Post-hoc construct development and articulation has always been problematic, though often it is the most prevalent form of construct development work in assessment design and use (Zumbo & Chan, 2014). Zumbo and Chan (2014) describe validation practices reported in published validity

[3]We see construct articulation as defining, describing, and critically interrogating a construct. It requires one to determine prior to developing the assessment itself, what constellation of knowledge and skills define what we are trying to measure (and teach to, and learn). Construct validation includes critical appraisal of our construct of interest, analysis of construction representation, and analysis of cognitive data (including process data) to determine if the test items are in fact measuring the construct elements they were blueprinted to measure.

research across the fields of social, behavioral, and health sciences as being haphazard, piecemeal, and unfocused. Applying this critique to the issue of construct validity, they state:

> [T]here is no universal understanding, or claim, of what constitutes "construct validity" evidence. The most common form of construct validity evidence is a statistical investigation of the internal structure of the measure via some form of dimensional analysis such as factor analysis. This operationalization of construct validity has persisted since the 1960s. (p. 322)

Validation arguments that depend on convergent and/or discriminant measures do not require *a priori*, clearly articulated construct models; rather, through an analysis of associated relationships between variables, researchers and assessment developers define (post hoc) what constructs are being measured—for example, "writing proficiency" ends up being characterized through clusters of linguistic features of Standard Edited English and text length, rather than modeled as a set of socio-cognitive features and elaborated through the response processes of students. In this approach, Standard Edited English becomes the unspoken norm for determining which linguistic features are valued, e.g., in the number and placement of metadiscourse markers, verb endings, and personal pronouns – while other Englishes (e.g., African American English, Appalachian English, Singlish, and Indian English) are labeled as "non-standard" or "errors." Standard Edited English, thus, becomes the stand-in for "correct" English. Given the historical association between Standard Edited English and whiteness, the unwritten elevation of Standard Edited English as "the" marker of writing proficiency becomes yet another tool of racism (Alim et al., 2016; Lippi-Green, 2012).

Bernal and Villalpando (2002) refer to this as an apartheid of knowledge in which only certain types of knowledge and knowledge production are valued. Inoue (2021) argues that such the habits of white language and judgment (HOWL) become invisible and then used against Black, Brown, Indigenous, and other students from historically marginalized populations:

> 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 [e.g., Individualized, rational, controlled self; Rule-governed, contractual relationships; Clarity, order, and control]. 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. (p. 22–23)

In response, Baker-Bell and other scholars advocating for linguistic justice and, specifically, "anti-black linguistic racism" (2020a, p. 8) suggests that "if language scholars and educators are truly interested in linguistic justice for linguistically and racially diverse students, we have to question whose linguistic and cultural norms are privileged by labels like "academic language" (2020b, p. 2).

But how are such features embedded in construct models for writing? And how do they get reproduced in assessment? A study by Dryer (2013) offers some insights (See also Beck & Jeffery, 2007; Jeffery, 2009; Mo & Troia, 2017). Dryer's (2013) analysis of 83 end-of-semester scoring rubrics from first-year writing courses at 157 U.S. public research universities provides a useful study of common constructs for assessing argumentative writing proficiency in college-level first-year writing. As Dryer argues, rubrics "are material artifacts of theoretical constructs–beliefs about writing and what it should look like" (p. 6) and that scoring rubrics "present their criteria and performance categories as uncomplicated means to an ideologically neutral end" (p. 27).

Based on his coded analysis of word usage within the rubrics, Dryer identified 10 "canonical traits" that commonly describe the construct of writing proficiency: (1) Grammar (mechanics, conventions, usage); (2) Evidence (support, development); (3) Thesis (focus, purpose, argument); (4) Style (voice, tone, variety, paragraphing); (5) Organization, structure; (6) Critical thinking, analysis; (7) Audience, rhetorical awareness; (8) Assignment, Engagement; (9) Creativity, originality; and (10) Writing process, revision (p. 12).

Using NVivo, Dryer then conducted further analysis of specific words associated with each trait along the performance levels. He found that 81 of the 83 rubrics assumed that the writer was

Table 1. Typical performance categories for the Trait of *Organization* in scoring a first-year college writing argumentative essay.

5	4	3	2	1
Overall structure, organization, and paragraph construction are appropriate to the assignment and an academic audience. All ideas in the paper flow logically. Transitions show originality and sophistication.	Overall structure, organization, and paragraph construction are appropriate to the assignment and an academic audience, though may be less evident and understandable in some places. Most ideas in the paper flow logically. Transitions are adequate but may be unclear or missing at times.	Overall structure, organization, and paragraph construction are readable, somewhat appropriate to the assignment and an academic audience, though may be a bit awkward in some places. The paper does not always flow logically and make sense. Transitions are formulaic and may be few or weak.	The paper lacks coherence, providing no discernable argument; ideas do not flow logically and do not make sense. Overall structure, organization, and paragraph construction are difficult to read, or inappropriate for audience. Transitions are confusing or nonexistent.	The writing is very difficult to understand owing to major problems with organization and structure.

a native speaker of English: there clearly was an embodied writer assumed in the construct model–i.e., a writer who was not multilingual. Furthermore, Dryer found that "in all traits, at all performance levels, readers' experiences of the texts are presented as intrinsic qualities of those texts" (p. 26), but mentions of "reader"' most likely appear "in those performance categories that register one's "difficulty" or where one's reading is "impeded" or "distracted" by the writer's lack of "control" (p. 23). In other words, the naturalized orientation to the world of writing proficiency in these models was one in which the assessor was meant to be pleased and that the weakest writers were those that most burdened the reader. Conversely, Dryer found that "apart from two instances of 'ignore' (as in 'ignores evidence that contradicts the claim')," the rubric corpus contained "no agentive verbs of failure [such as] 'refuse,' 'decline,' 'object,' 'resist,' 'oppose,' 'subvert,' 'parody,' 'mock,' or 'satirize'" (p. 27). A third finding from Dryer's study—"the heavy weighting of the entire corpus toward higher performance categories, both in simple word count and in richness of description"—points to the uneven articulation of the construct model in which the theoretical construct is constrained "to an announcement of its most favored conventions" (p. 27). As Table 1 demonstrates, the weakest writing is so evidently bad that it deserves little articulation; those students who are unable to produce those HOWL canonical traits are labeled not only the most cognitively inferior but also the most burdensome on their whitely-reading raters.

Untangling the legacies of racism flow through the scoring of writing using canonical constructs will not be solved through traditional means. Instead, if we are to really emancipate measurement from those legacies then we need a construct articulation process that disrupts this white-centric approach to defining knowledge.

Construct Articulation

In the case of writing, we offer one construct articulation possibility following Randall (2021, see Table 2) and Randall et al. (2021). Randall's justice-oriented, antiracist approach requires assessment developers to consider a series of interrelated questions related to when articulating an antiracist construct, beginning with articulating the *purpose* (an unapologetic commitment to justice) at every stage of the assessment development process. This is followed by considering and interrogating one's social *positionality* and *power* and the ways in which position and power have the potential to bias/influence one's assumptions about the communities the assessment is intended to serve resulting in marginalization and/or erasure of racially minoritized persons. Finally, Randall (2021) underscores the importance of adhering to an inclusive *process* for construct articulation that calls on the

Table 2. JAV six questions for the design of argumentative writing construct definition: example case study from University of Michigan (table from Randall, 2021).

Purpose: Articulate explicitly to stakeholders a commitment to justice in assessment practices (Organizational)	As we have discussed with stakeholders, the purpose of this writing assessment is to gauge what a student who is entering college knows about academic essay writing. Based on a review of assessment responses, this process will help students select a first-year writing (FYW) course or a Pre-FYW course. (See Gere et al., 2013)
Positionality: Reflect upon assumptions (Individual)	The design team includes three U.S. citizens—a Black woman from the South, a Latina who was born in Peru, and a white woman who identifies as Appalachian. The design team also includes a Canadian citizen—a white man. We have a shared commitment to a JAV approach.
People/Places: Consider who is being assessed and what sociocultural and racial identities they bring. (Organizational)	The enrolled student population at University of Michigan-Ann Arbor is 65% White, 15% Asian, 6% Hispanic or Latino, 5% Black or African American, 1% American Indian or Alaska Native, and 10% unknown. Students come from 50 states and 122 countries (University of Michigan, 2023). University of Michigan students are high-achieving students, meaning they have a familiarity with traditional academic writing, but it is also known that entering students do not access assessment prompts in the same way (Aull, 2021).
Power: Name those groups who hold the power, or are being privileged or harmed, by the way in which the construct is defined (Organizational)	Recent research has shown that the assessment has resulted in a "process at this institution [that] results in the unintended disproportionate placement and enrollment of domestic under-represented minority (URM) students and women in a pre-FYW course, which means that they take two courses rather than one to fulfill the FYW requirement" (Tinkle et al., 2022, p. 2).
Processes: Evaluate the procedures used for defining the construct. (Organizational)	Given the previous validation research related to this writing assessment, we have begun a process to better understand how admitted students understand the construct of academic writing and the associated consequences of the assessment. That process includes interviews with a purposeful sample of students as well as a critical review of the research base that informs our current understandings of the writing construct with respect to what populations this research has been conducted with.
Products & Consequences: Carefully consider the consequences of the construct as defined (Organizational)	Before recommending any curricular or policy changes, we question how the construct of argumentative writing should change, given the ongoing changes in academic writing related to AI and technological advances.

representation/influence of the most marginalized *people* to ensure that the resulting *products/consequences* does not, in fact, simply privilege whiteness.

For the sake of illustration, we draw on the example above from Dryer to argue that a JAV construct would begin by interrogating what test designers and stakeholders mean by "argumentative" writing. A JAV model might completely displace the canonical traits with a model of writing that includes metacognition (Negretti, 2012), transfer across contexts (Beaufort, 2007), or social learning (Vygotsky, 1997) (See also Corrigan & Slomp, 2021). A JAV construct model might eliminate the notion of a thesis while retaining the idea of purpose or stance, thus opening the possibility of valuing a deferred thesis text or offering a personal standpoint from which to speak. Instead of the singular notion of "voice," a JAV construct model for argumentative writing might value collaboration and poly-vocality. The forensic allusion to "evidence" might give way to a broader range of support, one that invokes a community of individuals, not a sole individual marshaling evidence to defend their case. Ideas such as translingual resources and linguistic repertoires (Canagarajah, 2015; Gumperz, 1964), community cultural wealth (Yosso, 2005), and rhetorical attunement (Leonard, 2014) would inform the construct model. And, critically, "grammar" approached through a sociocognitive lens would apply a critical discourse perspective that recognizes that linguistic forms are culturally and contextually mediated, that no dialect is inferior to another, so that effectiveness of linguistic choices is shaped by the

interaction between authorial intention and expectations and values of various audiences and/or communities of practice. In short, writing is a complex construct embedded in social systems; it is not canonical or fossilized, ahistorical, or acontextual. "Errors"' do not "violate some existential category of 'correctness'" (Dryer, p. 26).

Importantly, a JAV construct would include the perspectives of stakeholders. Within a JAV model, this could be completed following a dynamic criteria mapping (DCM) process (see Broad, 2003; Broad et al., 2009) that draws together the values and perspectives of diverse stakeholders as means of defining these criteria. A DCM approach to the design of scoring criteria begins with the assemblage of a corpus of texts that represent a range of performances in response to a specific task. A diverse set of stakeholders are then brought together to analyze and reflect on that corpus, interrogating each text by identifying features and patterns within the text as well as strengths and weaknesses from various standpoints. Once the corpus has been analyzed, the patterns, features, and lists of strengths and weaknesses are organized and coded thematically. The themes are used to derive scoring criteria. The strength of this process is that it allows for diverse perspectives, it derives criteria out of those perspectives, and enables the training of raters in a manner that is not normative or reductive as tends to be the case in psychometric models of rater training. Within this design process, terms like "appropriate" and "logically" (which are limited—who decides what is "appropriate"? Whose "logic" is deemed most valuable?) are replaced with phrases such as "unity in motion" or "focus over time" or "logical movement from one part of the journey to another" (Broad, 2003, p. 60). Implicit as well as explicit transitions are recognized as meaningful in academic writing (which they are). In the end, the DCM process enables the design of criteria that are grounded in the corpus and in the experiences of the diverse raters and helps researchers see what they cannot see and better understand the emotional, disciplinary, and cultural richness of a specific trait.

Construct Validation

With a purposeful construct model of argumentative writing articulated, we draw on Mislevy's (2018) argument that the construct validation processes should explore the interplay between model-based reasoning and the linguistic, cultural, and substantive (LCS) patterns that shape tasks, individual performances, and interpretations of both. For example, a workplace e-mail writing task designed to measure the construct of "writing ability" requires a clear definition of the knowledge, skills and dispositions that comprise that construct, a concrete mapping of task features and scoring criteria to that construct, and a thoughtful consideration of how LCS patterns influence designer, examinee, and scorer approaches to that task. Responses to this task will be stylistically and rhetorically different in many ways when written by examinees based in India, compared to examinees working in a corporate setting in North America, or compared to examinees working in a small-scale start-up founded by a team of Caribbean women. Such differences are a product of variations in LCS patterns each of these populations draw upon to inform their response. Tan et al. (2021), for example, found differences in LCS patterns find expression in the written responses of Chinese and American students responding to the same writing prompt. If the task, the scoring criteria, and interpretations of performance are centered in an American corporate context— thereby privileging the cultural patterns at play within that context—then, there will be misinterpretations of the performances of the other two populations of examinees. A JAV approach assumes that culture, race, gender, neurodiversity, and maturation (among many other factors) shape the ways in which writers approach writing tasks. This assumption calls for a program of research that utilizes purposeful sampling plans to enable exploration of how writers at different stages of maturation, and across cultural, racial, gender, and neurological groupings approach the act of writing.

The final requirement for a JAV construct validation process would be analysis of the implementation of the construct model to answer questions, such as: Have we considered the anticipated intended, unintended, immediate and long-term consequences stemming from decisions and choices being made when defining the construct? How does the embodiment of justice-oriented construct

definition/representation practices align with the actual construct definition practices that result in justice for stakeholders? (Randall et al., 2021, p. 598). Inspired by QuantCrit, JAV rejects the notion that there is some objective truth that cannot be/is not influenced by the subjectivity that every individual brings with them to the assessment enterprise; and this includes the articulation of the construct. Indeed, with respect to writing specifically in this illustration, what Inoue (2021) describes as habits of White language (HOWL) and judgment are endowed with the same degree of assumed truthfulness and objectivity when articulating the construct as numbers are when analyzing and interpreting data. The construct articulation and validation process we describe here is critically important for an anti-racist approach to test design and validation: it both breaks down the idea that constructs are monolithic-one truth for everyone – and, it opens up the realization that constructs themselves are cultural constructions. A commitment to model-based reasoning within a sociocognitive framework challenges us to rigorously interrogate what sits beneath the construct model: what research base, with what populations, based on what perspectives or assumptions? Traditional post hoc construct articulation and validation neither requires this level of critical reflection, nor does it lend itself to it. Instead, we acknowledge – and attempt to disrupt – the systems of oppression perpetuated through the white supremacist hegemony that seek to muffle, or de-legitimize, the ways of knowing and understanding of minoritized students in our construct articulation processes.

II: Analytic considerations

Suzuki et al. (2021) note that when studying young people of color the wider historical and socio-political contexts are rarely considered. Instead, we argue, explanations for/about their performance (to include assessment performance) are decontextualized, or essentialized, to include only the most immediate (uncritically obvious) contexts (e.g., free/reduced lunch status). For example, it is common practice to incorporate discrete variables such as race, free/reduced lunch status, and mother's education into statistical models of student performance. Indeed, these variables represent obvious (and easily quantifiable) indicators of performance. Far less often are contextual variables such as disparate exclusionary discipline policies and rates, the extent of police surveillance in schools and neighboring communities, and access to clean drinking water, fresh fruits/vegetables taken into consideration. A QuantCrit framing would require researchers to interrogate the data far more deeply than what is allowed by the gross categorization of students as White or Black, high or low income, etc. In fact, the inclusion of such variables alone serves to place the "blame" on the students and their families; and fails to acknowledge the oppressive systems in which these populations must operate.

Moreover, as Zuberi (2001) notes, despite claims of neutrality and objective truth, statistics play a role in the social construction of both race and racism. They are, in fact, used to demonstrate assumptions about between-group differences across racial groups; for example, through statistical tests of the mean which seek to highlight between-groups differences, while minimizing intra-group differences. Indeed, statistical analyses (including the interpretations) have been used as tools to perpetuate and/or sustain white supremacy (Gillborn, 2010; Zuberi & Bonilla-Silva, 2008). These blunt categorizations fail to take into account the diversity and heterogeneity of, for example, Black Americans (Blackwell, 1975; Valentine, 1971). As Boykin (1986) wrote: "Black people in their 'deficient' state are needed to provide living testimony of the cultural sanctity of being Euro-American: 'Things may be tough for me, but thank God at least I'm not black'" (p.63).

Research in predictive validity studies (commonly employed when establishing validity evidence with respect to relations to other variables) provide useful examples of psychometric limitations when analyzing data from minoritized learners. For example, Zwick and Sklar (2005) discussed the degree to which SAT scores and high school grade-point average (GPA) predicted first-year college GPA and college graduation and examined four groups: Hispanic students whose first language was Spanish, and Hispanic, Black, and White students whose first language was English. The authors conducted their analyses using two methods: regression and survival analyses. Study findings revealed that there

were differences in achievement patterns between the Hispanic/Spanish and Hispanic/English groups. These findings highlight two issues with typical, non-critical quantitative approaches to data analysis when establishing evidence of validity (especially predictive validity). First, most analyses fail to account for the intersecting sociocultural identities of test takers (e.g., a Hispanic student can be a native-English speaker, English speaking only, multilingual, Black, and/or a 3rd generation college student, etc.). In this case, the authors examined the intersection between first-language and ethnicity (indeed, finding a relationship), but still failed to consider other identities that could explain differential prediction—e.g., first-generation college student, whether the student was working, etc. Second, and most importantly, the authors failed to acknowledge the inherent limitations of the data especially with respect to its use and application among racially marginalized populations.

To be sure, the role/impact of systemic racism (as well as other systems of oppression) and the ways in which it manifests in student performance on large-scale admissions tests (Au, 2016, 2020; Feagin & Barnett, 2004; Tempel et al., 2012) and even in the ways in which instructors rate student performance (i.e., grades, Malouff & Thorsteinsson, 2016; Quinn, 2020; Tenenbaum & Ruck, 2007) has been well-documented (readers are encouraged to read Sireci & Randall, 2022 for a summary of the history of the use of large-scale assessment as a tool for oppression). Such analytic methods simply employ a network of deficit-framed data that ultimately serves to further marginalize entire populations of students. For example, Randall et al. (2022) point out that when building validity evidence with respect to relations to external variables for large-scale assessments, relying primarily (and uncritically) on student grades as the external criteria can be problematic if the goal is to suss out white supremacist logics in the assessment. Indeed, these criteria have been found to hold similar logics with teachers routinely assigning higher grades to White students than their minoritized counterparts for the same quality work (see for example Malouff & Thorsteinsson). Moreover, minoritized students are more likely to be subjected to restrictive school-based disciplinary policies (U.S. Department of Education, Office of Civil Rights, 2021) limiting their access to instruction (which would likely be reflected in lower classroom grades). In other words, white supremacist logics can inform the design and development of the assessment, while (independently and simultaneously) those white supremacist and racist logics have been found to work to inform teachers' perceptions of minoritized students' performance as well as teachers/administrators' perceptions of their behavior – all resulting in a cyclical series of score interpretations that both perpetuate deficit narratives and fail to recognize the systems of oppression that create them.

Moreover, Garcia et al. (2010) raised questions regarding the predictive validity of using only one test to exit students from English Language Learner (ELL) programs and whether such assessments can predict success in the mainstream classroom. Study findings suggest that through time English tests might become less predictive, leading at times to over prediction of student achievement and, as a consequence, potentially under-serving students and denying equal educational opportunities. Similar questions arise in the use of phonological processing assessments to predict reading readiness in students who speak African American Vernacular English (AAVE). Gopaul McNicol et al. (1999) found that early reading assessments (e.g., Peabody Individual Achievement Test) frequently provide an underestimation of capacity to read for speakers of AAVE. Sligh and Conners (2003) have also suggested that early reading assessments can both under- and over-predict reading readiness in speakers of AAVE.

Current operational procedures often rely on approaches in which test-takers groups are clustered together in uncritical and conventional ways that can limit, or reduce, the value of the information it provides to students, parents, or teachers. A justice-oriented approach to data analysis, especially within the context of building a validity argument, would require analysts to think critically about the important questions to ask, always holding in mind the ways in which their sociocultural identities influence the identification of these questions (and for the record, this influence should be acknowledged in every technical manual). In the following section, we highlight critical analytic approaches that seek to disrupt current practices that rely on false assumptions of homogeneity, white supremacy, and objective neutrality.

Critical Race Transformative Convergent Mixed Methods (CRTCMM)

Garcia and Mayorga (2018) argue that through the use of CRTCMM, one is able to capture what is happening at the macro level quantitatively while funneling down to a micro understanding qualitatively. The authors describe CRTCMM—which incorporates a CRT framework into a mixed method approach- as a "social justice or transformative advanced design addressing the oppression that marginalized communities experience on individual and systemic levels" (p.241). The process includes (a) applying the theoretical frameworks to the research questions, data interpretations, as well as the actual dataset used; (b) identifying and naming the limitations (e.g., the ways in which race and/or ethnicity are defined/reported) of the dataset, specifically attending to the ways in which white supremacy is made invisible through its normalization; (c) carefully examining the sampling approach and sampling decisions made (always acknowledging that statements about generalizability are rooted in a positivist paradigm); and (d) acknowledging the limitations of referring to racial groups as homogeneous populations (especially gross categories such as non-White, people of color).

Person-Centered Analysis

Person-centered analysis (PCA) can be employed as an alternative analytic strategy to identify specific factors when attempting to account for complex constructs. Unlike variable-centered analyses that focus on how characteristics (e.g., race, gender, income) relate to each other, person-centered analyses focus on how these variables group within individuals. Primarily conducted through cluster analysis, latent class analysis, or mixture modeling, PCA allows scholars to investigate, or identify, variables/categories that are not readily apparent (e.g., a PCA model might help reveal something about types of learning differences (ADHD) or socio-emotional attributes). *K*-means cluster analysis groups individuals together based on their scores/results on two or more variables. The ultimate goal is to reduce within-group and maximize between-group variability. A model-based extension, latent class analysis seeks to form clusters (i.e., subgroups) based on observed indicators (including categorical variables). Ultimately, LCA output provides a hypothesized grouping based on a combination of indicators. The considerable benefit to such an approach is that we are not presupposing where the homogeneity is (based on discrete variables such as race or gender, that we pre-identified) when conducting the analyses. Because the model first searches for similar patterns in the data (based on person responses) and creates homogenous groups on those responses, it lends itself more effectively to the interpretation of those groups with a critical lens.

Indeed, Suzuki et al. (2021) point out that mixture modeling techniques are particularly well suited within a QuantCrit framework as they, unlike methods that rely on group comparisons, allow the researcher to explore sample heterogeneity in the construct of interest. This refocus is particularly important from an antiracist perspective because "the interpretations from comparative work can and have been used to draw unidimensional and essentializing conclusions about people of color while upholding white people as the norm against which others are compared" (p.542).

Neblett et al. (2016) provide an overview of several studies that relied on the use of PCA to investigate the relationship between racial and ethnic protective processes that mitigate the negative impacts of racism for Black children, adolescents, and young adults. The authors reported that the use of PCA provided a fuller picture of racial identity than more traditional unidimensional approaches exposing the extensive heterogeneity within explicit groups (i.e., Black students). In fact, the authors found that, in some instances, different conclusions were drawn based on whether a variable-centered or person-centered approach was employed. Still, even with PCA, the clusters are identified by the observed variables that the researcher has selected to be included. Consequently, if the observed variables are problematic, so too will be the results of using PCA. In addition, naming the clusters is not always an easy task; and this process should always be conducted through a critical lens (For example, as Zhang-Wu, 2021 shows, it is too often assumed that Chinese international students'

educational performances in U.S. colleges are attributed to linguistic competence rather than whether they attended a school that taught a Western curriculum or not).

A QuantCrit framing with respect to data analysis discourages the gross inclusion of the variable race (and by extension other marginalized sociocultural identities) in favor of a more nuanced and critical approach to data analysis (and interpretation). PCA allows for that greater nuance. JAV, specifically, calls for an analysis of which groups are privileged by the assessment; and variable-centered approaches may not, necessarily, help in uncovering these groups (especially when these groups are less obvious at the onset). PCA also better enables the consideration of a wider range of interpretations that can support identifying different ways of knowing, thinking, and experiencing, especially for historically marginalized students.

Intersectionality Approaches

CRT maintains that individuals have multiple identities that intersect (Crenshaw, 1991) and influence their lived experiences. Covarrubias (2011) called for analytic procedures to account for the intersections of multiple identities when engaging in quantitative analyses. Specifically, the author disaggregated data among Chicano and Chicana students by race, class, gender, and citizenship status to gain an understanding of their educational trajectories which exposed the differentiation/heterogeneity within this group. Irizarry (2015), when examining the ratings that teachers give their students, considered the intersection of race, gender, and citizenship status noting that although East and South Asian immigrant students tended to receive more positive remarks than White (non-Hispanic) students, the same did not hold for East and South Asian students born in the United States. More recently, López et al. (2018), using a fully saturated mixed effects logistic model, examined the intersection of race, gender, and income with respect to graduation rates and developmental course-taking/placement at a large public university in New Mexico. Their findings revealed nuances beyond those typically reported through gross groupings (e.g., men v. women; white v. Black). For example, the authors point out that the graduation gap for high-income Black men mirrors that of low-income white men suggesting a unique set of experiences and exposure to structural racism for those racialized as Black which results in negative impacts on educational outcomes even with high levels of income. Still, there remains a dearth of intersectional quantitative studies in the field of education; and fewer still in the assessment/measurement literature.

The *Standards on Educational and Psychological Testing* (2014) note that DIF is "said to occur when equally able test takers differ in their probabilities of answering a test item correctly as a function of group membership" (p. 15). To be sure, DIF is the most common approach employed in the field of measurement to investigate potential item-level bias. As Russell and Kaplan (2021) point out, however, typical approaches to examining DIF fail to acknowledge intersecting marginalized identities and the compounding effects of oppression (p.3). Indeed, Zhang et al. (2005) suggested that a more comprehensive approach to creating more homogeneous groups within the overall test-taker population was to create more finely defined groups in DIF analyses by dissecting DIF wherein groups are crossed to create finer groupings (e.g., gender is crossed by ethnicity), which may lead to more accurate DIF results when investigating item response patterns from gender and ethnic groups. Russell and Kaplan (2021) proposed a similar application of an intersection approach to DIF in which the reference group is a select intersectional group (e.g., white, economically advantaged male) and each remaining intersectional group (e.g., Black, economically advantaged, women) forms a focal group. Such approaches allow scholars to uncover differences that can/do go undetected when examining sub-groups separately (men only v. women only).

Using data from 67,000 5th grade test takers on a 25-item English Language Arts (ELA) state examine, Russell and Kaplan found considerable differences in the proportion of items flagged for investigation (DIF) with the traditional single-group approach to DIF (20% of items flagged as suspicious) for one or more focal groups and the intersectional approach (68% of items flagged as suspicious and 16% flagged as likely). Despite practical challenges (i.e. sample size requirements,

increased multiple comparisons, and additional time required for item review), Russell and Kaplan's (2021) suggested the field of measurement embrace a process for detecting/addressing that DIF relies on a justice as fairness (Rawls, 1971/Rawls, 1999) approach (all benefit) rather than the utilitarian notion of the "greatest good."

Similarly to PCA approaches to data analysis, intersectionality approaches allow researchers to more carefully consider the consequences of assessment results across a wider range, and far more well-defined, of marginalized groups (e.g., Black x low-income, x access to highly qualified teachers x access to honors courses) to determine the extent to which these groups may, or may not, be further marginalized by the assessment results. Moreover, as part of a larger JAV argument, it likewise allows researchers to determine which intersecting sociocultural identities (especially those that have historically marginalized) are the most disadvantaged by the assessment results. Still, the limitations of intersectional analyses can mirror the limitations of PCA in that we are bound by the data (in both quantity and quality) that is collected. For example, the extent to which brown-skinned Hispanic students differ in their experiences from light-skinned (white) Hispanic students cannot be captured through the use of gross categories like "White," "Black," "Hispanic," and "Asian" alone. A justice-oriented approach to analyses calls out these limitations and addresses them through a critical interpretation of the findings. We discuss these justice-oriented approaches to score interpretation and use in the final section.

III: Score interpretation and use

Murnane et al. (2005) note that educators engage with assessment results in three ways: (1) an *instrumental approach* in which they rely directly on student scores to make decisions (e.g., who receives additional supports) with no consideration for the causes of the test scores; (2) a *symbolic approach* is used to justify a prior decision (e.g., additional funding for an existing program); and (3) a *conceptual approach* which calls on educators to interrogate the assessments results for explanations and patterns, followed by a systematic investigation of these possible causes (p.271). The authors argued that a conceptual approach to score interpretations was both the most valuable and under-utilized approach employed by educators.

Similarly, in 2016, Moss encouraged the field of educational measurement to shift its approach to validity theory to support conceptual approaches that "help educators connect the test-based data to their own practice and to consider explanations and explore solutions" (Moss, 2016, p.2). We argue, further still, that a *JAV conceptual* approach to score interpretation would include a close examination of the ways in which systems of oppression are manifest in assessment results as well as a discussion of ways in which stakeholders can disrupt those systems. 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 (Steele, 1997; Steele & Aronson, 1995) serves as a useful example here of the pernicious impact of the perpetuation of persistent, negative, racist narratives on the performance of minoritized students. Although some experts have argued that racial differences in scores found on so-called culture-free cognitive ability tests such as Raven's Advanced Progressive Matrices (Raven, 1936) represent evidence of genuine biological differences in ability (famously argued by Hernstein & Murray, 1994), such claims fail to consider the broader sociocultural contexts that minoritized students must navigate/endure daily. For example, Brown and Day (2006) found a meaningful difference in the scores of African American participants when they accounted for stereotype threat. African American students in the no-threat group scored three-fourths of a standard deviation higher than African Americans in the high or standard threat groups; and about as well as White students in the no threat group. The frequent failure of scholars to even acknowledge the possibility of the role of stereotype threat in differences revealed in test scores represents merely one example of the ways in which the broader sociocultural context is often ignored by researchers (yet this context is ever present for the examinee).

De-centering whiteness as part of the score interpretation process requires us to do more than simply disaggregate and report data across racial, ethnic, and socioeconomic groups. In fact, it is important when interpreting scores, not to use demographic data (e.g., Black, multilingual) as a proxy for constructs such as systemic or institutional racism. Harper (2012) examined the instances of the use of the terms race or racism in the discussion sections of 255 articles across seven academic journals. The author found that racism was rarely discussed as a source of disparate negative outcomes in racially minoritized groups. Although scholars were willing to acknowledge that minoritized groups endure considerable hardship and that this hardship contributes to negative outcomes, they failed – almost always – to acknowledge that these hardships could be a direct result of racism. He describes one example in which the researchers found racial differences in undergraduate students' grade point averages and wrote: "The lower GPA for Blacks may result from disadvantaged backgrounds" (p.16). The authors, however, failed to consider the cumulative effects of racist stereotypes Black students report and endure in their undergraduate courses (a toxic environment their White counterparts do not have to endure), which can make it more difficult for them to achieve the same outcomes. He wrote "my argument is that ongoing attempts to study race without racism are unlikely to lead to racial equity and more complete understandings of minoritized populations in postsecondary contexts" (p.15). As Bonilla-Silva and Baiocchi (2001) have argued elsewhere because scholars "fail to explain that the "race effect" presented in their findings is the outcome of "racism" or "racial stratification" [this] leads their audiences to interpret "race effect" findings as embodying truly racial effects ("There must be something wrong with Blacks if they are three times more likely than Whites to participate in crime!")" (p.125). We posit that uneven learning profiles, differences in opportunities to learn, limited curricular exposure are (a) simply consequences of a racist system designed (from the beginning) to produce these outcomes and/or (b) the result of a system that has persistently sought to gaslight both racially minoritized and dominant communities into ignoring/erasing/de-valuing the ways of knowing and understanding present and valued in these racially minoritized communities.

To support more meaningful score interpretation, prior research in the field of educational assessment and measurement has encouraged the inclusion of a wide range of explanatory variables (see Ercikan et al., 2005, for example) suggesting a strong relationship between scores from large-scale assessments and student-level variables (e.g., student beliefs, confidence, or motivation levels), which may impact achievement across curricular areas such as mathematics and science. Measurement scholars argue that because test takers from diverse demographic backgrounds may differ greatly with respect to previous schooling experiences, school resources, computer literacy, and teacher instructional practices (Areepattamannil et al., 2010; Klinger et al., 2006), including such background characteristics can help increase accuracy in score-based inferences and more appropriate interventions. Scholars rarely, however, propose the inclusion of system-level measures of oppression that (most assuredly) contribute to the manifestation of these student-level and school-level characteristics and the resulting score performance. We argue that accurate score interpretation requires a comprehensive interrogation of the full historical and sociopolitical context in which student scores are assigned. For example, when examining and interpreting student performance and relations to other variables to describe concurrent and/or predictive relationships, researchers should examine and report the relationship between student scores and indicators of systemic oppression (e.g., over-policing, housing costs, proximity to/access to healthful foods, exclusionary discipline rates, proportion of teachers who speak students' home language) without respect to the racial/ethnic identities of the students.

Moreover, care should be taken not to use scores/score reports to problematize racially minoritized groups as this aids in their further marginalization. We caution against the use of deficit-framed language (e.g., attributing score differences to the racial group instead of, for example, the educational intervention) in reporting the results; or failing to be explicit about the instrument/scale used (e.g., Regents exam, GRE, (fill in the blank)'s Scale of Self-Regulation) and instead referring to the broad construct (e.g., reading achievement, college readiness, or self-regulatory skills), which all contribute to racist notions about what minoritized students know and are able to do. These critical re-

orientations in the framing of how data are interpreted encourages a focus of resources not in "fixing" racially minoritized students (i.e., making them more white); but rather in addressing the systems of oppression that create and perpetuate the education debt. For example, according to the Department of Education's Office for Civil Rights, Black students are disproportionately denied access to education through exclusionary discipline practices. Black boys are suspended and expelled at proportions three times their enrollment; and Black girls represent 7.4% of the population and 11.2% of the suspensions and expulsions (U.S. Department of Education, Office for U.S. Department of Education, Office of Civil Rights, 2022 And these discriminatory disciplinary practices begin in preschool with Black preschoolers receiving one or more suspensions at 2.5 times greater than their share of the total preschool population (U.S. Department of Education, Office of Civil Rights, 2021). Moreover, Black, Brown, and Indigenous students are more likely to be taught by teachers who are uncertified and/or less experienced than their white counterparts (U.S. Department of Education, Office for S & Department of Education, Office of Civil Rights, 2014. In other words, a critical interpretation of the data might reveal that racially minoritized students are, too often, restricted from receiving even the substandard educational experience (e.g., less qualified teachers) made available to them; and that these racist systems (not students' race) are the source of differential performances.

Finally, we also encourage test developers to employ asset-based language choices when describing test-taker performance (and this recommendation applies across all racial and ethnic groups). Scholars have long argued that the words we use to describe students and their work processes/products matter (Becker, 1963; Rist, 1997; Roduta Roberts et al., 2018; Triplett & Upton, 2015) with both classroom and large-scale assessments. O'Donnell and Sireci (2022) write "In addition to influencing teachers, labels from these assessments may have an impact on students' self-perception and educational plans as well as their parents' expectations and actions" (p.2). These labels – often the only consistent form of communication a parent or student receives from local or state education agencies—authoritatively signal a students' perceived value and standing in the educational space. Consequently, these reports have the capacity to elevate (and provide information about a productive way forward) or demoralize (and predict inevitable doom). Still, the score report design process in many states, seemingly, ignores this scholarship and potential impact. O'Donnell and Sireci (2022) reviewed the achievement labels (middle school) of statewide assessments from all 50 states. In this review, they found that some states use deficit-based labels such as *Fail-below basic*, *Far below proficient*, *unsatisfactory* or *inadequate* to describe student performance. Instead the authors encourage the use of more encouraging terminology—such as not yet or approaching-when describing students with the lowest levels of performance (e.g., not yet meeting expectations, or approaching expectations). As the authors explain, such a practice would reduce the likelihood of harmful overgeneralizations about what students know and are capable of doing.

Conclusions

Stage (2007) defined the quantitative criticalist researcher as one who "adapts a proactive stance by consciously choosing questions that seek to challenge . . . to illuminate conflict, and develop critique through quantitative methods to move theory, knowledge, and policy to a higher plane"(p.8). In practice, a justice-oriented, antiracist approach to the validation process (from construct articulation to score interpretation) requires us all to engage as quantitative criticalist researchers. In their 2022 article, Randall et al., when referring to the process of building a validity argument for any assessment, suggest that we ask "What characteristics of the assessment, the assessment design process, and/or the inferences drawn from the assessment provide evidence of antiracism?" (p.5). Here, using QuantCrit as our framework, we focused on considerations during three important stages in the assessment development process: (1) construct articulation – what ways of knowing and understanding are privileged- what data do we value and collect?; (2) analysis – what sociocultural identities are erased/unacknowledged in the data analysis process- how do investigate/organize these data? and (3) score interpretation – which sociocultural identities are marginalized/problematized in our

interpretation of these data – what narratives do we perpetuate with the data?. These processes serve to uncover racist logics that may be deeply embedded – and seemingly invisible- within our procedures for assessment design.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Adelman, C. (2004). *Principal indicators of student academic histories in postsecondary education, 1972–2000*. U.S. Department of Education.

AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. AERA.

AERA & NCMUE. (1955). *Technical recommendations for achievement tests AERA*. Washington DC: National Education Association.

Alim, H. S., Rickford, J. R., & Ball, A. A. (Eds.). (2016). *Raciolingistics: How language shapes our ideas about race*. Oxford University Press.

Areepattamannil, S., Freeman, J., & Klinger, D. (2010). Influence of motivation, self-beliefs, and instructional practices on science achievement of adolescents in Canada. *Social Psychology of Education*, *14*(2), 233–259. https://doi.org/10.1007/s11218-010-9144-9

Attewell, P. A., Lavin, D. E., Domina, T., & Levey, T. (2006). New evidence on college remediation. *The Journal of Higher Education*, *77*(5), 886–924. https://doi.org/10.1353/jhe.2006.0037

Au, W. (2016). Meritocracy 2.0: High stakes, standardized testing as a racial project of neoliberal multiculturalism. *Educational Policy*, *30*(1), 39–62. https://doi.org/10.1177/0895904815614916

Au, W. (2020). *Testing for whiteness? How high-stakes standardized tests promote racism, undercut diversity, and undermine multicultural education in visioning multicultural education: Past, present, future* (Eds ed.). Prentice Baptiste & Jeannette Haynes Writer.

Aull, L.(2021). Directed self-placement: Subconstructs and group differences at a US university. *Assessing Writing*, *48*. Article 100522. https://doi.org/10.1016/j.asw.2021.100522

Baker-Bell, A. (2020a). Dismantling anti-black linguistic racism in English language arts classrooms: Toward an anti-racist black language pedagogy. *Theory into Practice*, *59*(1), 8–21. https://doi.org/10.1080/00405841.2019.1665415

Baker-Bell, A. (2020b). *Linguistic justice: Black language, literacy, identity, and pedagogy*. NCTE Routledge Research Series.

Beaufort, A. (2007). *College writing and beyond: A new framework for university writing instruction*. Utah State University Press.

Becker, H. S. (1963). *Outsiders: Studies in the sociology of deviance*. Free Press.

Beck, S. W., & Jeffery, J. V. (2007). Genres of high-stakes writing assessments and the construct of writing competence. *Assessing Writing*, *12*(1), 60–79. https://doi.org/10.1016/j.asw.2007.05.001

Bell, D. (1995). Who's afraid of critical race theory? *University of Illinois Law Review*, *1995*(4), 893–910.

Bernal, D., & Villalpando, O. (2002). An apartheid of knowledge in academia: The struggle over the "legitimate" knowledge of faculty of color. *Equity & Excellence in Education*, *35*(2), 169–171. https://doi.org/10.1080/713845282

Blackwell, J. (1975). *The black community: Diversity and unity*. Dodd Mead.

Bonilla-Silva, E., & Baiocchi, G. (2001). Anything but racism: How sociologists limit the significance of racism. *Race and Society*, *4*(2), 117–131. https://doi.org/10.1016/S1090-9524(03)00004-4

Boykin, A. W. (1986). The triple quandary and the schooling of African American children. In U. Neisser (Ed.), *T*he school achievement of minority children: New perspectives (pp. 57–92). Lawrence Erlbaum Associates.

Brasington, D., & Haurin, D. (2006). Educational outcomes and house values: A test of the value added approach. *Journal of Regional Science*, *46*(2), 245–268. https://doi.org/10.1111/j.0022-4146.2006.00440.x

Broad, B. (2003). *What we really value: Beyond rubrics in teaching and assessing writing*. University Press of Colorado.

Broad, B., Adler-Kassner, L., Alford, B., Detweiler, J., Estrem, H., Harrington, S., McBride, M., Stalions, E., & Weeden, S. (2009). *Organic writing assessment: Dynamic criteria mapping in action*. University Press of Colorado.

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

Canagarajah, S. (2015). *Translingual practice: Global englishes and cosmopolitan relations*. Routledge.

Chen, X. (2016). Remedial coursetaking at U.S. Public 2- and 4-year institutions: Scope, experiences, and outcomes (NCES 2016-405). U.S. Department of Education. National Center for Education Statistics. http://nces.ed.gov/pubsearch.

Corrigan, J. A., & Slomp, D. H. (2021). Articulating a sociocognitive construct of writing expertise for the digital age. *The Journal of Writing Analytics*, *5*(1), 142–195. https://doi.org/10.37514/JWA-J.2021.5.1.05

Covarrubias, A. (2011). Quantitative intersectionality: A critical race analysis of the chicana/o educational pipeline. *Journal of Latinos and Education*, *10*(2), 86–105. https://doi.org/10.1080/15348431.2011.556519

Crenshaw, K. (1991). Mapping the margins: Identity politics, intersectionality, and violence against women. *Stanford Law Review*, *43*(6), 1241–1299. https://doi.org/10.2307/1229039

de Brey, C., Musu, L., McFarland, J., Wilkinson-Flicker, S., Diliberti, M., Zhang, A., Branstetter, C., & Wang, X. (2019). Status and trends in the education of racial and ethnic groups 2018 (NCES 2019-038). U.S. Department of Education. National Center for Education Statistics. https://nces.ed.gov/pubs2019/2019038.pdf

Delgado, R., & Stefancic, J. (2017). *Critical race theory: An introduction*. New York University Press.

Dryer, D. (2013). Scaling writing ability: A corpus-driven inquiry. *Written Communication*, *30*(1), 3–35. https://doi.org/10.1177/0741088312466992

Epp, C., Maynard-Moody, S., & Haider-Markel, D. (2014). *Pulled over: How police stops define race and citizenship*. University of Chicago Press.

Ercikan, K., McCreith, T., & Lapointe, V. (2005). Factors associated with mathematics achievement and participation in advanced mathematics courses: An examination of gender differences from an international perspective. *School Science and Mathematics*, *105*(1), 301–321. https://doi.org/10.1111/j.1949-8594.2005.tb18031.x

Feagin, J., & Barnett, B. (2004). *Success and failure: How systemic racism trumped the Brown v. Board of education decision*. University of Illinois Law Review.

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), Article 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*(3), 217–233. https://doi.org/10.1007/s40894-015-0006-8

Garcia, E., Lawton, K., & Diniz de Figueiredo, E. (2010). Assessment of young english language learners in arizona: Questioning the validity of the state measure of English proficiency. The Civil Rights Project. Retrieved September 15, 2022 from https://civilrightsproject.ucla.edu/research/k-12-education/language-minority-students/assessment-of-young-english-language-learners-in-arizona-questioning-the-validity-of-the-state-measure-of-english-proficiency

Garcia, N., & Mayorga, O. (2018). The threat of unexamined secondary data: A critical race transformative convergent mixed methods. *Race Ethnicity and Education*, *21*(2), 231–252. https://doi.org/10.1080/13613324.2017.1377415

Gere, A. R. Aull, L. Perales, M. D. Escudero, Z. L. & Vander Lei, E.(2013). Local assessment: Using genre analysis to validate directed self-placement. *College Composition and Communication*, *64*(4), 605–633.

Gillborn, D. (2010). The colour of numbers: Surveys, statistics and deficit-thinking about race and class. *Journal of Education Policy*, *25*(2), 253–276. https://doi.org/10.1080/02680930903460740

Gillborn, 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

Gilman, H. (2019). Are we whom we claim to be? A case study of language policy in community college writing placement practices. *Journal of Writing Assessment*, *12*(1). http://journalofwritingassessment.org/article.php?article=137

Goodman, D., & Hambleton, R. (2010). Student test score reports an interpretive guide: Review of current practices and suggestions for future research. *Applied Measurement in Education*, *17*(2), 145–220. https://doi.org/10.1207/s15324818ame1702_3

Gopaul McNicol, S., Reid, G., & Wisdom, C. (1999). The psychoeducational assessment of ebonics speakers: Issues and challenges. *The Journal of Negro Education*, *67*(1), 16–24. https://doi.org/10.2307/2668236

Gordon, E. (1995). Toward an equitable system of educational assessment. *The Journal of Negro Education*, *64*(3), 360–372. https://doi.org/10.2307/2967215

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 & Public Policy*, *19*(3), 905–940. https://doi.org/10.1111/1745-9133.12512

Gumperz, J. J. (1964). Linguistic and social interaction in two communities. *American Anthropologist*, *66*(6), 137–153. https://doi.org/10.1525/aa.1964.66.suppl_3.02a00100

Harper, S. R. (2012). Race without racism: How higher education researchers minimize racist institutional norms. *The Review of Higher Education*, *36*(1), 9–29. https://doi.org/10.1353/rhe.2012.0047

Hernstein, R. J., & Murray, C. A. (1994). *The bell curve: Intelligence and class structure in American life*. Free Press.

Hilliard, A. G. (2000). Excellence in education versus high-stakes standardized testing. *Journal of Teacher Education*, *51*(4), 293–304. https://doi.org/10.1177/002248710005100400 5

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

Howell, J. S. (2011). What influences students' need for remediation in college? Evidence from California. *The Journal of Higher Education*, *82*(3), 292–318. https://doi.org/10.1353/jhe.2011.0014

Inoue, A. B. (2021). *Above the well: An antiracist literacy argument from a boy of color*. Colorado State University Press.
Irizarry, Y. (2015). *Utilizing multidimensional measures of race in education research the case of teacher perceptions. Sociology of Race and Ethnicity*: 2332649215580350. https://doi.org/10.1177/2332649215580350
Jeffery, J. V. (2009). Constructs of writing proficiency in US state and national writing assessments: Exploring variability. *Assessing Writing*, *14*(1), 3–24. https://doi.org/10.1016/j.asw.2008.12.002
Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527
Kane, M. T. (2013). Validating the interpretation and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000
Kane, R. (2002). The social ecology of police misconduct. *Criminology*, *40*(4), 867–896. https://doi.org/10.1111/j.1745-9125.2002.tb00976.x
Klausman, J., & Lynch, S. (2022). From ACCUPLACER to informed self-placement at Whatcom Community College: Equitable placement as evolving practice. In J. Nastal, M. Poe, & C. Toth (Eds.), *Writing placement in two-year colleges: The pursuit of equity in postsecondary education* (pp. 59–83). The WAC Clearinghouse, University Press of Colorado.
Klinger, D., Rogers, W., Anderson, J., Poth, C., & Calman, R. (2006). Contextual and school factors associated with achievement on a high-stakes examination. *Canadian Journal of Education / Revue canadienne de l'éducation*, *1*(3), 771–797. https://doi.org/10.2307/20054195
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
Leonard, R. L. (2014). Multilingual Writing as Rhetorical Attunement. *College English*, *76*(3), 227–247. http://www.jstor.org/stable/24238241
Lippi-Green, R. (2012). *English with an accent: Language, ideology, and discrimination in the United State*. Routledge.
López, N., Erwin, C., Binder, M., & Chavez, M. J. (2018). Making the invisible visible: Advancing quantitative methods in higher education using critical race theory and intersectionality. *Race Ethnicity and Education*, *21*(2), 180–207. https://doi.org/10.1080/13613324.2017.1375185
Malouff, J., & Thorsteinsson, E. (2016). Bias in grading: A meta-analysis of experimental research findings. *Australian Journal of Education*, *60*(3), 425–256. https://doi.org/10.1177/0004944116664618
Messick, S. (1980). Test validity and ethics of assessment. *The American Psychologist*, *35*(11), 1012–1027. https://doi.org/10.1037/0003-066X.35.11.1012
Mislevy, R. (2018). *Sociocognitive foundations of educational measurement*. Routledge.
Montenegro, E., & Jankowski, N. (2017). Equity and assessment: Moving towards culturally responsive assessment. Occasional Paper #29. National Institute for Learning Outcomes Assessment, 23 pages.
Moss, P. (2016). Shifting the focus of validity for test use. *Assessment in Education Principles, Policy & Practice*, *23*(2), 236–251. https://doi.org/10.1080/0969594X.2015.1072085
Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, *62*(3), 229–258. https://doi.org/10.3102/00346543062003229
Moss, P. A. (1995). Themes and variations in validity theory. *Educational Measurement: Issues and Practice*, *14*(2), 5–12. https://doi.org/10.1111/j.1745-3992.1995.tb00854.x
Mo, Y., & Troia, G. A. (2017). Similarities and differences in constructs represented by U.S. States' middle school writing tests and the 2007 national assessment of educational progress writing assessment. *Assessing Writing*, *33*, 48–67. https://doi.org/10.1016/j.asw.2017.06.001
Murnane, R. J., Sharkey, N. S., & Boudett, K. P. (2005). Using student-assessment results to improve instruction: Lessons from a workshop. *Journal of Education for Students Placed at Risk (JESPAR)*, *10*(3), 269–280. https://doi.org/10.1207/s15327671espr1003_3
Nastal, J. (2019). Beyond tradition: Writing placement, fairness, and success at a two-year college. *Journal of Writing Assessment*, *12*(1). http://journalofwritingassessment.org/article.php?article=136
Neblett, E. W., Jr., Sosoo, E. E., Willis, H. A., Bernard, D. L., Bae, J., & Billingsley, J. T. (2016). Racism, racial resilience, and African American youth development: Person-centered analysis as a tool to promote equity and justice. *Advances in Child Development and Behavior*, *51*, 43–79. https://doi.org/10.1016/bs.acdb.2016.05.004
Negretti, R. (2012). Metacognition in student academic writing: A longitudinal study of metacognitive awareness and its relation to task perception, self-regulation, and evaluation of performance. *Written Communication*, *29*(2), 142–179. https://doi.org/10.1177/0741088312438529
Nettles, M. (2019). History of testing in the United States: Higher education. *The Annals of the American Academy of Political and Social Science*, *683*(1), 38–55. https://doi.org/10.1177/0002716219847139
Ngo, F., & Melguizo, T. (2020). The equity cost of inter-sector math misalignment: Racial and gender disparities in community college Student outcomes. *The Journal of Higher Education*, *92*(3), 410–434. https://doi.org/10.1080/00221546.2020.1811570
O'Donnell, F., & Sireci, S. (2022). Language matters: Teacher and parent perceptions of achievement labels from educational tests. *Educational Assessment*, *27*(1), 1–26. https://doi.org/10.1080/10627197.2021.2016388

Oliveri, M. E., Mislevy, R., & Elliot, N. (2020). New horizons for postsecondary placement and admission practices in the United States. In M. E. Oliveri & C. Wendler (Eds.), *Higher education admission practices: An international perspective* (pp. 347–375). Cambridge University Press.

Palmer, K. (2018). *Legacy of missing lloyd gaines, 1938 supreme court plaintiff, still haunts higher education*. KBIA. https://www.kbia.org/education/2018-12-12/legacy-of-missing-lloyd-gaines-1938-supreme-court-plaintiff-still-haunts-higher-education

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

Quinn, D. (2020). Experimental evidence on teachers' racial bias in student evaluation: The role of grading scales. *Educational Evaluation and Policy Analysis*, *42*(3), 375–392. https://doi.org/10.3102/0162373720932188

Radford, A. W., & Horn, L. (2012). Web tables—an overview of classes taken and credits earned by beginning postsecondary students (NCES 2013-151REV). National Center for Education Statistics, Institute of Education Sciences, U.S Department of Education.

Randall, J. (2021). "Color-neutral" is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement Issues & Practice*, *40*(4), 82–90. https://doi.org/10.1111/emip.12429

Randall, J., Poe, M., & Slomp, D. (2021). Ain't oughta be in the dictionary: Getting to justice by dismantling anti-black literacy assessment practices. *Journal of Adolescent & Adult Literacy*, *64*(5), 594–599. https://doi.org/10.1002/jaal.1142

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

Raven, J. C. (1936). Mental tests used in genetic studies: The performance of related individuals on tests mainly educative and mainly reproductive. MSc Thesis, University of London.

Rawls, J. (1971). *A theory of justice*. Harvard University Press.

Rawls, J. (1999). *A theory of justice* (Revised ed.). Harvard University Press.

Research for Action. (2020). School policing in pennsylvania: Prevalence and disparities. *Police in Pennsylvania Schools Series*. https://files.eric.ed.gov/fulltext/ED608042.pdf

Rist, R. C. (1997). On understanding the processes of schooling: The contributions of labeling theory. In J. Karabel & P. W. Halsey (Eds.), *Power and ideology in education* (pp. 292–305). Oxford University Press.

Roduta Roberts, R., Gotch, C., & Lester, J. (2018). Examining score report language in an accountability system. *Frontiers in Education*, *3*(42), 1–17. https://doi.org/10.3389/feduc.2018.00042

Russell, M., & Kaplan, L. (2021). An intersectional approach to differential item functioning: Reflecting configurations of inequality. *Practical Assessment, Research & Evaluation*, *26*(21). https://doi.org/10.7275/20614854.

S, U., & Department of Education, Office of Civil Rights. (2014). Data snapshot: teacher equity. https://www2.ed.gov/about/offices/list/ocr/docs/crdc-teacher-equity-snapshot.pdf

Scott-Clayton, J., Crosta, P. M., & Belfield, C. R. (2014). Improving the targeting of treatment: Evidence from college remediation. *Educational Evaluation and Policy Analysis*, *36*(3), 371–393. https://doi.org/10.3102/0162373713517935

Shepard, L. A. (2021). Ambitious teaching and equitable assessment: A vision for prioritizing learning, not testing. *American Educator*, *45*(3), 28–48.

Sireci, S., & Randall, J. (2022). Evolving notions of fairness in testing in the United States. In M. Bunch & B. Clauser (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 111–1350). Routledge.

Sligh, A. C., & Conners, F. A. (2003). Relation of dialect to phonological processing: African American vernacular English vs. *Standard American English Contemporary Educational Psychology*, *28*(2), 205–228. https://doi.org/10.1016/S0361-476X(02)00013-9

Smith, D. (1986). The neighborhood context of police behavior. In A. Reiss & M. Tonry (Eds.), *Crime and justice* (Vol. 8, pp. 313–341). University of Chicago Press. https://doi.org/10.1086/449126

Snyder, T. D., & Dillow, S. A. (2012). Digest of education statistics, 2011 (NCES 2012-001). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Solórzano, D., & Ornelas, A. (2002). A critical race analysis of advance placement classes: A case of educational inequalities. *Journal of Latinos and Education*, *1*(4), 215–229. https://doi.org/10.1207/S1532771XJLE0104_2

Solórzano, D., & Villalpando, O. (1998). Critical race theory, marginality, and the experience of students of color in higher education. In C. A. Torres & T. R. Mitchell (Eds.), *Sociology of education: Emerging perspectives* (pp. 211–224). State University of New York Press.

Solórzano, D., & Yosso, T. (2002). Critical race methodology: Counter-storytelling as an analytical framework for education research. *Qualitative Inquiry*, *8*(1), 23–44. https://doi.org/10.1177/107780040200800103

Souto-Manning, M., & Winn, M. T. (2017). Introduction: Foundational understandings as "show ways" for interrupting injustice and fostering justice in and through education research. *Review of Research in Education*, *41*(1), ix–xix. https://doi.org/10.3102/0091732X17703981

Stage, F. (2007). Answering critical questions using quantitative data. *New Directions for Institutional Research*, *133*(133), 5–16. https://doi.org/10.1002/ir.200

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *The American Psychologist*, *52*(6), 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 & Social Psychology*, *69*(5), 797–811. https://doi.org/10.1037/0022-3514.69.5.797

Suzuki, S., Morris, S., & Johnson, S. (2021). Using QuantCrit to advance an anti-racist development science: Applications to mixture modeling. *Journal of Adolescent Research*, *36*(5), 535–560. https://doi.org/10.1177/07435584211028229

Tan, T., Fan, X., Braunstein, L., & Lane-Holbert, M. (2021). Linguistic, cultural and substantive patterns in L2 writing: A qualitative illustration of Mislevy's sociocognitive perspective on assessment. *Assessing Writing*, *51*, 51. https://doi.org/10.1016/j.asw.2021.100574

Tempel, M., Au, W., & Bollow-Tempel M. (2012). *Pencils down: Rethinking high stakes testing and accountability in public schools*. Rethinking Schools.

Tenenbaum, H. R., & Ruck, M. D. (2007). Are teachers' expectations different for racial minority than for European American students? A meta-analysis. *Journal of Educational Psychology*, *99*(2), 253–273. https://doi.org/10.1037/0022-0663.99.2.253

Tinkle, T. Godfrey, J. Menon, A. R. Moos, A. Romaine, L. & Sprouse, M.(2022). (In)equities in directed self-placement. *Assessing Writing*, *54*, Article 100671.

Triplett, R., & Upton, L. (2015). Labeling theory: Past, present, and future. In A. R. Piquero (Ed.), *The handbook of criminological theory* (pp. 271–289). Wiley.

University of Michigan. (2023). Data & reports. Diversity, equity, and inclusion. https://diversity.umich.edu/data-reports/

U.S. Department of Education, Office of Civil Rights. (2016). *Key data highlights on equity and opportunity gaps in our nation's public schools*. U.S. Department of Education.

U.S. Department of Education, Office of Civil Rights. (2021). Discipline practices in preschool. https://www2.ed.gov/about/offices/list/ocr/docs/crdc-DOE-Discipline-Practices-in-Preschool-part1.pdf

U.S. Department of Education, Office of Civil Rights. (2022). Suspensions and expulsions in public schools. https://www2.ed.gov/about/offices/list/ocr/docs/suspensions-and-expulsion-part-2.pdf

Valentine, C. (1971). Deficit, difference, and bi-cultural models of Afro-American behavior. *Harvard Educational Review*, *41*(2), 137–157. https://doi.org/10.17763/haer.41.2.n141n3g84145506m

Vygotsky, L. (1997). *Educational psychology*. St. Lucie Press.

Weiler, S., & Cray, M. (2011). Police at school: A brief history and Current status of school resource officers. *The Clearing House: A Journal of Educational Strategies, Issues, and Practices*, *84*(4), 160–163. https://doi.org/10.1080/00098655.2011.564986

Whitaker, A., Torres-Guillen, S., Morton, M., Jordan, H., Coyle, S., Mann, A., & Sun, W. (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

Yosso, T. J. (2005). Whose culture has capital? A critical race theory discussion of community cultural wealth. *Race Ethnicity and Education*, *8*(1), 69–91. https://doi.org/10.1080/1361332052000341006

Zhang, Y., Dorans, N., & Matthews-Lopez, J. (2005). Using DIF detection method to assess effects of item deletion. College Board Research Report No. 2005-10. https://files.eric.ed.gov/fulltext/ED563094.pdf.

Zhang-Wu, Q. (2021). *Languaging myths and realities: Journeys of Chinese international students*. Multilingual Matters.

Zuberi T. (2001). *Thicker than blood: How racial statistics lie*. University of Minnesota.

Zuberi T, & Bonilla-Silva E. (2008). *White logic, white methods: Racism and methodology*. Rowman & Littlefield.

Zumbo, B. D., & Chan, E. K. (2014). Validity and validation in social, behavioral, and health sciences. New York City, NY: Springer International Publishing.

Zwick, R., & Sklar, J. (2005). Predicting college grades and degree completion using high school grades and SAT scores: The role of student ethnicity and first language. *American Educational Research Journal*, *42*(3), 439–464. https://doi.org/10.3102/00028312042003439