EDUCATIONAL ASSESSMENT
2022, VOL. 27, NO. 2, 170–178
https://doi.org/10.1080/10627197.2022.2042682

Routledge
Taylor & Francis Group

Disrupting White Supremacy in Assessment: Toward a Justice-Oriented, Antiracist Validity Framework

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

[a]University of Massachusetts Amherst, Amherst Center, MA, United States; [b]University of Lethbridge, Lethbridge, NWT, Canada; [c]Northeastern University, Boston, Massachusetts, United States; [d]University of Nebraska Lincoln, USA

ABSTRACT

In this article, we propose a justice-oriented, antiracist validity framework designed to disrupt assessment practices that continue to (re)produce racism through the uncritical promotion of white supremist hegemonic practices. Using anti-Blackness as illustration, we highlight the ways in which racism is introduced, or ignored, in current assessment and validation processes and how an antiracist approach can be enacted. To start our description of the framework, we outline the foundational theories and practices (e.g., critical race theory & antiracist assessment) and justice-based framings, which serve as the base for our framework. We then focus on Kane's interpretive use argument and Mislevy's sociocognitive approach and suggest extending them to include an antiracist perspective. To this end, we propose a set of heuristics organized around a validity argument that holds justice-oriented, antiracist theories and practices at its core.

As we witness the disproportionate impact of a global pandemic on Black communities and the anti-Black racist violence in this country manifested through the murders of Black citizens such as Daunte Wright, Ahmaud Arbery, George Floyd, and Breonna Taylor, it reminds us of the anti-Black violence mirrored in American K-16 classrooms daily. In a written statement to the AERA membership, President Shaun Harper (S. Harper, email communication, June 1, 2020) noted "Evidence from multiple sources across numerous academic disciplines and fields consistently highlights systems that cyclically disadvantage Black people." Indeed, schooling is one of those systems. Yet, far too often, these inequities are interpreted through a deficit – racist – lens that blames the students and we neglect to openly acknowledge the systems of oppression and marginalization that create the systemic disadvantage in which Black students live.

Historically, the assessment field (i.e., assessment companies, developers, & educators) has contributed to further perpetuate these injustices. In fact, Dixon-Roman (2020) writes: "even with the sociocultural and postmodern turns in educational assessment and measurement, there remains a haunting logic in the epistemology of psychometrics that maintains colonialist formations … "(p.94). Repeatedly, the use of standardized educational assessments has resulted in harm to Black students by, for example, erroneously attributing their difference as their pathology (Franklin, 2007; Larry, 1979; Williams, 1971); reifying racist stereotypes to substantiate narratives of inferiority and deficiency (Hernstein & Murray, 1994; Hilliard, 1976, 1992); limiting access to a comprehensive, equitable, and quality education by facilitating tracking policies (Hilliard, 1992; Larry, 1979; Slavin, 1987; for examples). Moreover, more recent uses of standardized assessments as an attempt to remedy educational inequities has arguably created even more harm such as increases in high school dropout rates (Madaus & Clarke, 2001) and curricular changes that encourage cultural erasure/avoidance (Au, 2009; Darder & Torres, 2004) in minoritized populations. In fact, Au (2008) has argued that the use of

CONTACT Jennifer Randall jrandall@umass.edu University of Massachusetts Amherst

high-stakes standardized assessments to facilitate educational reform has served, instead, primarily as a tool for reproducing race-based and class-based inequities in education. Add to this, an "increasingly covert nature of racial discourse and practices," continues to infiltrate the measurement profession through the use of expressions such as "lack of rigor" when engaging in conversations about (re) developing justice-oriented assessments, thereby enabling the implicit use of mechanisms that reproduce racial inequality and embedding them into the very fabric of educational assessment design, use, and interpretation (Bonilla-Silva, 2013).

Engagement with questions of equity in assessment have occurred through the lens of fairness and through the inter-relationship of fairness and validity (Kane & Bridgeman, 2017). While valuable, such discussions have not squarely interrogated how the validation process, arguably the most important process in educational measurement, has served as one of measurement's primary weapons of racist oppression and marginalization. Furthermore, as noted by Cushman (2016), "fairness can address content of particular questions, but it does little to adjust the overall ways in which validity measures themselves, from the start are based on the colonial difference that they help to create and maintain" (section 2.0, para 5). Consider that validity theory, like most assessment theory, was developed in a color-blind framework, thus lending credibility to ideas that fairness is comparable to validity across subgroups (Xi, 2010) without asking, for example, how the notion of comparable validity itself may be unfair.

Cushman (2016) goes on to argue that "validity as a tenet is used to claim, gather, and justify results with so many performance and survey tools, it has now more than ever been used to routinize inequities as naturalized parts of systems of educational access . . . " (para 19.) In other words, how do the interpretations drawn from sources of evidence result in the logics that naturalize Black students' failures? Until the assessment community can answer such a question, we argue that traditional approaches to the validation of assessments maintain a white supremist hegemony.

This tendency to uncritically center whiteness exclusively has not gone unaddressed in assessment (see Cushman, 2016; Moss, 1992; Poe & Inoue, 2016; Randall, 2021; Slomp, 2016; Solano-Flores & Nelson-Barber, 2001). Still, the field of measurement, largely, continues to rely on approaches to validation that- instead of seeking justice for marginalized communities- treats the sociocultural identities of test takers as barriers of inferiority to be mitigated outside of the assessment design process. When the ongoing realities of social oppression are not recognized, the use of validity arguments becomes another racist tool, reproducing- rather than disrupting-systems of oppression.

Foundational theories and practices informing our antiracist framework

Our proposed justice-oriented, antiracist validity (JAV) framework is built on foundational theories and practices related to justice-oriented perspectives, critical race theory, and antiracist assessment.

A justice-oriented approach to assessment and measurement rejects a utilitarian theory of "greatest good" (Mill, 1863) in favor of a theory that rests on equal rights, "fair equality of opportunity," and "the greatest benefit of the least advantaged members of society" (Rawls, 2001, §13, pp. 42–43). A justice-oriented, *antiracist* approach to assessment and measurement takes these ideas further, acknowledging "the way racist violence in our contemporary world seems to recapitulate the violence of the past" against Black Americans (Davis, 2018, p. 232). An antiracist approach to assessment, thus, is historical and critical.

Any discussion of antiracist justice in the U.S. would be remiss without acknowledging the long history of Black intellectual thought (see DuBois, 1968; Wells, 1892; Wilkerson, 2020) and the central role, in contemporary theory, of Critical Race Theory (CRT; Bell, 1995; Delgado & Stefancic, 2017). CRT asserts that (1) race is a social construct, which can be shifted and differentially applied based on the needs of the dominant culture; (2) racism is not aberrational; rather it is typical, pervasive, and ingrained in the fabric and system of American society, and (3) it is critical to recognize the relevance of people's everyday lives to scholarship, which includes acknowledging the lived experiences of minoritized peoples and rejecting deficit-informed research that excludes the epistemologies of

marginalized groups. CRT draws attention to the ways that assessment is part of larger social structures in the U.S. in which racism is typical, pervasive, and ingrained. For example, an antiracist theory of justice demands attention to the barriers created through institutional, not just personal, racism:

> Differential access to the goods, services, and opportunities of society by race. Institutionalized racism is normative, sometimes legalized, and often manifests as inherited disadvantage. It is structural, having been codified in our institutions of custom, practice, and law, so there need not be an identifiable perpetrator. Indeed, institutionalized racism is often evident as inaction in the face of need (Jones, 2000, p. 1212)

Finally, because racism is normative, not aberrant, responsibility for justice is not left to a few individuals (for example, to determine whether a test is racially biased or not); rather justice is a shared responsibility to which individuals "contribute by their actions to the processes that produce unjust outcomes" (Young, 2011, p. 105). Responsibility, Young argues, derives "from participating in the diverse institutional processes that produce structural injustice" (p. 105).

More than a decade ago, Inoue called for "racial validity," "an argument that explains the degree to which empirical evidence of racial formations around our assessments and the theoretical frameworks that account for racial formations support the adequacy and appropriateness of inferences and actions made from the assessments" (2009, p. 110). A justice-oriented, antiracist assessment approach builds off such thinking in requiring an explicit confrontation of racism in our assessment practices and works to disrupt these systems of oppression through assessment practices specifically designed to sustain, not eradicate, students' cultures, languages, and ways of knowing/being (Inoue, 2015). Antiracist assessment interrogates the ways in which racism rationalizes and continues to perpetuate injustice and support differential hierarchical power structures (based on race) in society. Furthermore, like antiracist pedagogy, antiracist assessment recognizes that issues associated with race and racism cannot simply be abstracted from the broader historical and sociopolitical context which have maintained inequitable systems of power. Ultimately, a commitment to antiracist assessment processes moves beyond simply ensuring representation (of individuals in the field) and includes a fundamental shift in assessment processes. For assessment professionals, it requires self-reflection and analysis of one's own social identities and relationship to power and how these identities/power relationships are enacted in their scholarship/research and their work in communities (Randall, Poe, & Slomp, 2021). Antiracist assessment is explicit about its politics and its intent to reconstruct hierarchical racial power arrangements that have been historically (re)produced via assessments (Randall, 2021).

Motivation for our framework

Our proposed JAV aims to liberate the field of measurement from validation practices that continue to (re)produce racism through the uncritical promotion of white supremist assessment practices. Our motivation for proposing the JAV framework is that current theories have not gone far enough to explicitly include key considerations relevant to meaningfully informing inferences for all test-takers, especially Black students. Randall (2021) argues that a deliberate departure from so-called/assumed race-neutral guidelines and practices in the assessment design process- especially with respect to construct articulation- is necessary to disrupt these oppressive, white supremacist notions frequently replicated in/through most assessment systems. When defining the construct and evaluating construct representation, she extends the frameworks of Randall et al. (2021) and Inoue (2015) with a suggested series of interrelated questions with respect to (a) purpose, (b) positionality, (c) people/places, (d) power, (e) processes, and (f) products/consequences) that seek to ensure an explicit and unapologetic commitment to a justice-oriented antiracist construct definition process. In this paper, we extend that antiracist construct articulation framework to include the entire validation process. Specifically, we describe how a JAV framing can be used to extend Kane's (2013) Interpretation Use Argument (IUA)

model of validation and Mislevy's (2018) sociocognitive enhancement of that model. We also provide a heuristic for JAV that extends key elements of widely used validity theory arguments from an antiracist perspective.

Proposed Justice-oriented anti-racist validity (JAV) framework

Kane (2013) and Mislevy (2018) bring awareness of the importance of considering populations and their sociocultural, socioeconomic, and other differences in their way of interacting with the assessment situation in their proposal for a conditional sense of validity. Kane observes that validity is conditional with respect to temporality, completeness of an IUA, plausibility of underlying assumptions, and degree of ambition associated with claims being made on the basis of test scores. Applying a sociocognitive lens, Mislevy extends this sense of conditionality to the underlying measurement and construct models validation arguments are built upon. This conditional structure draws attention from the elements and procedures of the assessment to the model-based reasoning that takes us from performance on the assessment itself to performance-based inferences and decisions, as applied to test takers in other non-test conditions.

This conditional view recognizes that the constructs that underpin an assessment design are themselves social constructs (e.g., "writing proficiency" is a social construct in which standard American English is a privileged dialect); interpretations and uses of assessment data are based on these construct models. However, these constructs and the reasoning models based on them are necessarily limited by designers' understandings of the network of knowledge, skills, and dispositions that inform these constructs. Inherent limitations are baked into an assessment program to the degree that both understanding of these constructs, and the relationships between construct elements, are limited. Further complicating the challenge, Mislevy observes that examinees' performances are dependent on the cognitive resources they have accumulated over time (e.g., patterns of discourse in one's home environment and local community), and not, strictly speaking, the construct models that drive interpretation and use of assessment results. In some respects, then, it can be said that what is construct-relevant behavior is dependent on the eye of the beholder. What is construct irrelevant for the assessment designer, may very well be construct relevant for the examinee.

To address this issue, Mislevy argues that tasks, performances, and criterion situations need to be examined through the lens of targeted LCS (linguistic, cultural, and substantive) patterns and practices. To the degree that examiners and examinees share LCS knowledge and experiences, their access to LCS patterns may be similar. However, to the degree that these experiences differ, their access to these patterns may differ. This brings us to the crux of validity from an antiracist perspective. What Mislevy is pointing to, what validity theory historically speaking has been missing, is a set of fundamental questions that ground and initiate assessment design and validation. From a sociocognitive perspective, the essential question is, "Whose LCS patterns are being privileged by an assessment, whose LCS patterns are being devalued, omitted, suppressed, or marginalized?"

Tan, Fan, Braunstein & Lane Holbert (2021) demonstrate an application of this question to writing assessment. They examined Chinese and American undergraduate students' responses to two writing tasks through the lens of LCS patterns, demonstrating how culturally based differences in these patterns shaped how these two groups of students responded to the same tasks. Their analysis prompted the following observation:

> Researchers and practitioners must grapple with whether the cultural and substantive features expressed in L2 essays would/should be seen as less "proficient". Hypothetically, because of potential inter-cultural communication barriers resulting from differences in cultural beliefs, assessment tools designed with limited consideration of cross-cultural difference in cultural and substantive features might have a built-in bias.

Tan et al.'s (2021) study highlights the importance of critically interrogating the constructs that underpin assessment design with respect to their stability across racial contexts. It does not matter if test designers make tests more "culturally relevant" 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.

Extending this focus on sociocognition to include a perspective rooted in CRT, we might rephrase our earlier question as: "What characteristics of the assessment, the assessment design process, and/or the inferences drawn from the assessment provide evidence of antiracism?" These questions are summarized in Table 1 and go beyond our traditional views of validity, which have often been asked of the overall population without examining the ways that racism may be embedded in each source of evidence. Table 1 demonstrates how to surface racist thinking in order to combat it.

As illustration, we refer to one of the best-known K12 assessment systems in the nation – the Massachusetts Comprehensive Assessment System (MCAS). Developed in 1993 by the Department of Elementary and Secondary Education (DESE) as a response to the Massachusetts Education Reform Act, the MCAS is a statewide standards-based assessment program which requires that all students educated with public funds be tested annually (in specific subjects) beginning in the third grade (i.e., grades 3, 4, 5, 6, 7, 8, & 10). Although we argue that the JAV approach should be considered from the first stage (construct articulation) through every stage of collecting validity evidence, we also maintain that even a post-hoc interrogation of assessments and the assessment development process can be informative in disrupting embedded white supremacist and racist notions. When used from the beginning of the test design process, the JAV approach is woven throughout construct articulation as well as the collection, interpretation, and use of validity evidence (Slomp, Corrigan, & Sugimoto, 2014). When used a post-hoc heuristic, the JAV approach demonstrates how racist logics are embedded in the supporting documentation and artifacts of test design.

For the sake of brevity, we focus on two sources of validity evidence: relations to external variables and test content in a post-hoc analysis of the MCAS exam as provided through an example drawn from the 2019 Legacy MCAS Technical Report (Massachusetts Department of Elementary and Secondary Education, 2019a) and separate Appendix U (Massachusetts Department of Elementary and Secondary Education, 2019b).

Relations to External Variables. The MCAS technical manual reports that three sets of validity analyses with concurrent measures were conducted examining the: (1) relationship between MCAS scaled scores in ELA and math and students' course grades in grades 6, 7, 8, and 10 and compared those relationships with other student demographic variables; (2) relationships among MCAS achievement levels and students' course grades in grades 6, 7,8, and 10; and (3) incidences of taking higher-level math courses in grades 8 and 10 by MCAS achievement levels and by MCAS scores on Mathematics exams. The collection of these types of data as evidence of relations to other variables is typical practice (and not at all particular to the MCAS). We argue, however, that this practice fails to disrupt the white supremist hegemony in several ways: (1) Although the analyses involved the examination of the relationships across student demographic variables (i.e., disability status, economic status, and EL status), no examination of the extent to which the test-criterion relations generalize across other historically marginalized groups (mainly Black, Brown, and Indigenous students) was presented. Bonilla-Silva (2013) writes that the tremendous influence of race (even when ignored) is made apparent by both the words we use for race and the ways in which we avoid its reference altogether. Failure to examine the relationship between test-criteria and race does not remove the relationship. The [possible] existing relationship is simply ignored resulting in the further perpetuation of the white supremist hegemony. (2) Moreover, this process- and the criterion selected (i.e., course grades) – fails to consider the extent to which course grades can also reflect the influence of white supremist thinking and practices. Indeed, there is considerable literature that suggests instructors assign lower grades to minoritized students than their

Table 1. Heuristic for a Justice-Oriented, Antiracist Approach to Building a Validity Argument.

Criteria Element	Traditional Validity Arguments Ask:	Antiracist Validity Arguments Ask:
Construct Articulation		• Are marginalized stakeholders involved at every stage of the construct definition and refinement stage? • How well understood is the construct being measured for all including minoritized learners? • Whose values, perspectives, ways of knowing, and experiences does the construct reflect, normalize, or marginalize? • How stable is the construct across social, cultural, and racial contexts? • Is the construct explicitly antiracist? Does it articulate the specific false and oppressive narratives it seeks to disrupt?
Content	Do the test items represent the targeted domain of interest?	• Do the test items reflect/reify negative stereotypes of minoritized populations? • Are there test items that actively disrupt negative stereotypes about minoritized populations? • Has antiracist content been integrated into items explicitly? • Does the content/language of the items privilege a particular linguistic or cultural way of thinking/making sense of the world?
Consequences	Do items function differentially across student groups?	• Do test/assessment results serve to further marginalize already minoritized populations? • What groups may be advantaged and disadvantaged by the administration of this assessment? In the short term? The long term? • How does/can structural racism impact the results of this assessment? • What groups will be privileged by this assessment? In the short term? The long term? • What systems would have to be in place for a student to be successful on this assessment? Are those systems rooted in white supremist values?
Response Processes	Are the test takers interpreting the tasks as intended?	• Have we considered/interrogated whether or not the assessment requires responses that are in alignment with the dominant white discourse? • Have a wide range of interpretations been considered that acknowledge the different ways of knowing, thinking, and experiencing of Black students? • What historical logics of testing and racism are students bringing to the test situation? • Are the items written in such a way that assumes only one "right" way of getting a correct response? Do the items allow for multiple ways of thinking and knowing? • Are Eurocentric ways of knowing and processing information being privileged over other ways of knowing and processing information?
Internal Structure	Does the relationship among test items and test components conform to the construct?	• How have values shown up in the items/tasks? And which social identity groups do these values reflect?
Relations to Other Variables	Does the observed relationship between the test results and external variables match the predicted relationship?	• To what extent do test-criterion relations generalize across historically marginalized populations? • How are criterion variables selected? Does this selection process consider the history/impact/legacy of white supremist hegemonic practices? Does this process of criterion selection seek to disrupt these hegemonic practices? • Have alternative criterion variables – those that center and reflect the values of nonwhite students been considered/examined?

white peers for the same quality of work (Malouff & Thorsteinsson, 2016; Tenenbaum & Ruck, 2007). Additionally, large racial disparities in school-based disciplinary policies/actions disproportionately exclude BIPOC students from instruction (see Riddle & Sinclair, 2019; U.S. Department of Justice Civil Rights Division and U.S. Department of Education Office of Civil Rights) resulting in – among other adverse consequences – lower course grades. The exclusive use of grades as the external criterion fails to disrupt the legacy of racist practices in schools and the ways in which the assessment experiences of historically marginalized students are impacted by these practices.

A justice-oriented approach to the MCAS test would (a) explicitly acknowledge the role/impact of race and ethnicity by examining it in all analyses to gauge the extent these criterion relations generalize across these groups and (b) rely on the use of multiple criterion variables including those that center and reflect the values and priorities of nonwhite communities (and these criteria would be determined in collaboration and consultation with the impacted communities).

Test Content. In detailing the item development process, the Massachusetts Department of Elementary and Secondary Education provides traditional evidence with respect to test content describing how it ensures that all items used on the MCAS can be directly linked to the curriculum frameworks of the state. Items are examined for alignment to the standards, the extent to which the content represents a depth of understanding of the subject, appropriate and realistic use of context, grade-level appropriateness, mechanics (adherence to the conventions of item writing), and the plausibility of distractors (p. 17). We maintain that attention to these criteria – and these criteria alone – misses the opportunity to ensure that MCAS exams serve justice-oriented aims. For example, test items are examined during the development process to ensure that they adhere to style guidelines (pp. 18–19) adhering to correct grammar, punctuation, and usage. This criterion, in effect, mandates that the content/language of all items privilege a particular linguistic or cultural way of thinking. Moreover, no evidence is described that suggests that items are designed, or selected, to actively disrupt negative stereotypes or include antiracist content.

Although here we focused on applying the JAV framework with respect to two sources of validity evidence, we encourage test developers to consider the framework in its entirety during the assessment development process from construct articulation to score reporting. For example, the MCAS reports four achievement levels: *Failing, Needs Improvement, Proficient* and *Advanced* (MA DESE, 2019, p. 47). To be sure assessment developers and the clients they serve should consider the ways in which the language used both within the assessment itself (i.e., test content) and in the reporting of the results can serve to further marginalize already marginalized student populations (i.e., testing consequences). The use of achievement level language such as *failing* sends an unkind, deficit -based narrative to the student, the parents/guardians, and entire communities whose schools are deemed failing. Most importantly, we admonish assessment developers to articulate (in detail) in technical manuals the ways in which issues of equity and justice have been addressed explicitly throughout the validation process. Doing so empowers stakeholders – especially those committed to justice and/or from marginalized communities – to make informed decisions about the value and trust they assign these assessments.

Disclosure statement

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

References

Au, W. (2008). Devising inequality: Bersteinian analysis of high-stakes testing and social reproduction in education. *Journal of Sociology of Education*, *29*(6), 639–651. doi:10.1080/01425690802423312

Au, W. (2009). High-stakes testing and discursive control: The Triple bind for non-standard student identities. *Multicultural Perspectives*, *11*(2), 65–71. doi:10.1080/15210960903028727

Bell, D. A. (1995). Who's Afraid of Critical Race Theory? *University of Illinois Law Review*, *4*, 893–910.
Bonilla-Silva. (2013). *Racism without racists: Color-blind racism and the persistence of racial inequality in the United States*. Lanham, MD/ Boulder,CO/New York/ Oxford, UK: Rowman & Littlefield.
Cushman, E. (2016). Decolonizing Validity. *The Journal of Writing Assessment*, *9*(1). http://journalofwritingassessment.org/article.php?article=92. Retrieved on from October 1, 2020.
Darder, A., & Torres, R. D. (2004). *After race: Racism after multiculturalism*. New York: New York University Press.
Davis, A. Y. (2018). The Past, Present, and Future of Assata's Message. *WSQ: Women's Studies Quarterly*, *46*(3–4), 232–234. doi:10.1353/wsq.2018.0044
Delgado, R., & Stefancic, J. (2017). *Critical Race Theory: An Introduction* (3rd edition ed.). New York: New York University Press.
Dixon-Roman, E. (2020). A Haunting logic of psychometrics: Toward the speculative and indeterminacy of Blackness in measurement. *Educational Measurement: Issues and Practice*, *39*(11), 94–96. doi:10.1111/emip.12375
DuBois, W. E. B. (1968). *The souls of black folk; essays and sketches*. Chicago, IL: A. G. McClurg, 1903. Reprint Corp.
Franklin, V. P. (2007). The tests are written for the dogs: "The Journal of Negro Education," African American children, and the intelligence testing movement in historical perspective. *The Journal of Negro Education*, *76*(3), 216–229.
Hernstein, R., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life*. New York, NY: Free Press.
Hilliard, A. G., III. (1976). *Alternatives to IQ testing: An approach to the assessment of gifted "minority" children* (Final Report to the Special Education Support Unit). Sacramento: California State Dept. of Education. (ERIC Document Reproduction Service No. ED 147 009
Hilliard, A. G., III. (1992). Behavioral style, culture, and teaching and learning. *The Journal of Negro Education*, *61*(3), 370–377. doi:10.2307/2295254
Inoue, A. (2009). The technology of writing assessment and racial validity. In C. Schreiner (Ed.), *Handbook of Research on Assessment, Technologies* (pp. 97-120). Hershey, Pennsylvania: Methods, and Applications in Higher Education. IGI Global.
Inoue, A. (2015). *Antiracist Writing Assessment Ecologies: Teaching and Assessing Writing for a Socially Just Future*. Fort Collins, Colorado: WAC Clearinghouse/University of Colorado Press.
Jones, C. P. (2000). Levels of racism: A theoretic framework and a gardener's tale. *American Journal of Public Health*, *90*, 1212–1215.
Kane, M. T. (2013). Validating the interpretation and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. doi:10.1111/jedm.12000
Kane, M., & Bridgeman, B. (2017). Research on Validity Theory and Practice at ETS. In R. Bennett, and M. von Davier (Eds.), *Advancing Human Assessment. Methodology of Educational Measurement and Assessment* (pp. 489-552). New York City, NY: Springer Open.
Larry, P. V. R.(1979). No. C-71–2270 RFP California
Madaus, G. F., & Clarke, M. (2001). The adverse impact of high-stakes testing on minority students: Evidence from 100 years of data. In G. Orfield, and M. Kornhaber (Eds.), *Raising standards or raising barriers? Inequality and high stakes testing public education* (pp. 1- 49). New York: The Century Foundation.
Malouff, J. M., & Thorsteinsson, E. B. (2016). Bias in grading: A meta-analysis of experimental research findings. *Australian Journal of Education*, *60*(3), 245–256. doi:10.1177/0004944116664618
Massachusetts Department of Elementary and Secondary Education (2019a). Technical Manual Validity Evidence, Appendix U: Retrieved on from October 11, 2021 http://www.mcasservicecenter.com/documents/MA/Technical%20Report/2019/NextGen%20ADA/2019_MCAS_NG_TechReport_Appendix%20U.pdf
Massachusetts Department of Elementary and Secondary Education (2019b). Legacy MCAS Technical Report. Retrieved on October 11, 2021 from http://www.mcasservicecenter.com/documents/MA/Technical%20Report/2019/Legacy%20ADA/2019_MCAS_Legacy_TechReport.pdf
Mill, J. S. (1863). *Utilitarianism*. London, UK: Parker, Son, and Bourn.
Mislevy, R. (2018). *Sociocognitive Foundations of Educational Measurement*. New York, NY: Routledge.
Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, *62*(3), 229–258. doi:10.3102/00346543062003229
Poe, M., & Inoue, A. (2016). Toward writing as social justice: An Idea whose time has come. *College English*, *79*(2), 119–126.
Randall, J. (2021). Color-neutral is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice*, *40*(4), 82–90. doi: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. doi:10.1002/jaal.1142
Rawls, J. (2001). *Justice as Fairness: A Restatement*. Cambridge, MA: Harvard University Press.
Riddle, T., & Sinclair, S. (2019). Racial disparities in school-based disciplinary actions are associated with county-level rates of racial bias. Proceedings of the National Academy of Sciences of the United States of America. Retrieved from October 14, 2021 https://www.pnas.org/content/116/17/8255

Slavin, R. E. (1987). *A review of research on elementary ability grouping*. Baltimore, MD: John Hopkins University Press.
Slomp, D. (2016). An Integrated design and appraisal framework for ethical writing assessment. *The Journal of Writing Assessment*, *9*(1). http://journalofwritingassessment.org/article.php?article=91. Retrieved from October 2.
Slomp, D., Corrigan, J., & Sugimoto, T. (2014). A framework for using consequential validity evidence in evaluating large-scale writing assessments: A Canadian study. *Research in the Teaching of English*, *48*(3), 276–302.
Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching*, *38*(5), 553–573. doi:10.1002/tea.1018
Tan, T., Fan, X., Braunstein, L., & Lane-Holbert, M. (in press). Linguistic, cultural and substantive patterns in L2 writing: A qualitative illustration of MisLevy's sociocognitive perspective on assessment. *Assessing Writing*. doi:10.1016/j.asw.2021.100574
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. doi:10.1037/0022-0663.99.2.253
Wells, I. (1892). *Southern Horrors: Lynch Law in All Its Phases*. New York: The New York Age Print.
Wilkerson, I. (2020). *Caste: The Origins of Our Discontents*. New York, NY: Random House.
Williams, R. L. (1971). Abuses and misuses in testing black children. *The Counseling Psychologist*, *2*(3), 62–73. doi:10.1177/001100007100200314
Xi, X. (2010). How do we go about investigating test fairness. *Language Testing*, *27*(2), 147–170.
Young, I. M. (2011). *Responsibility for Justice*. Oxford: Oxford University Press.