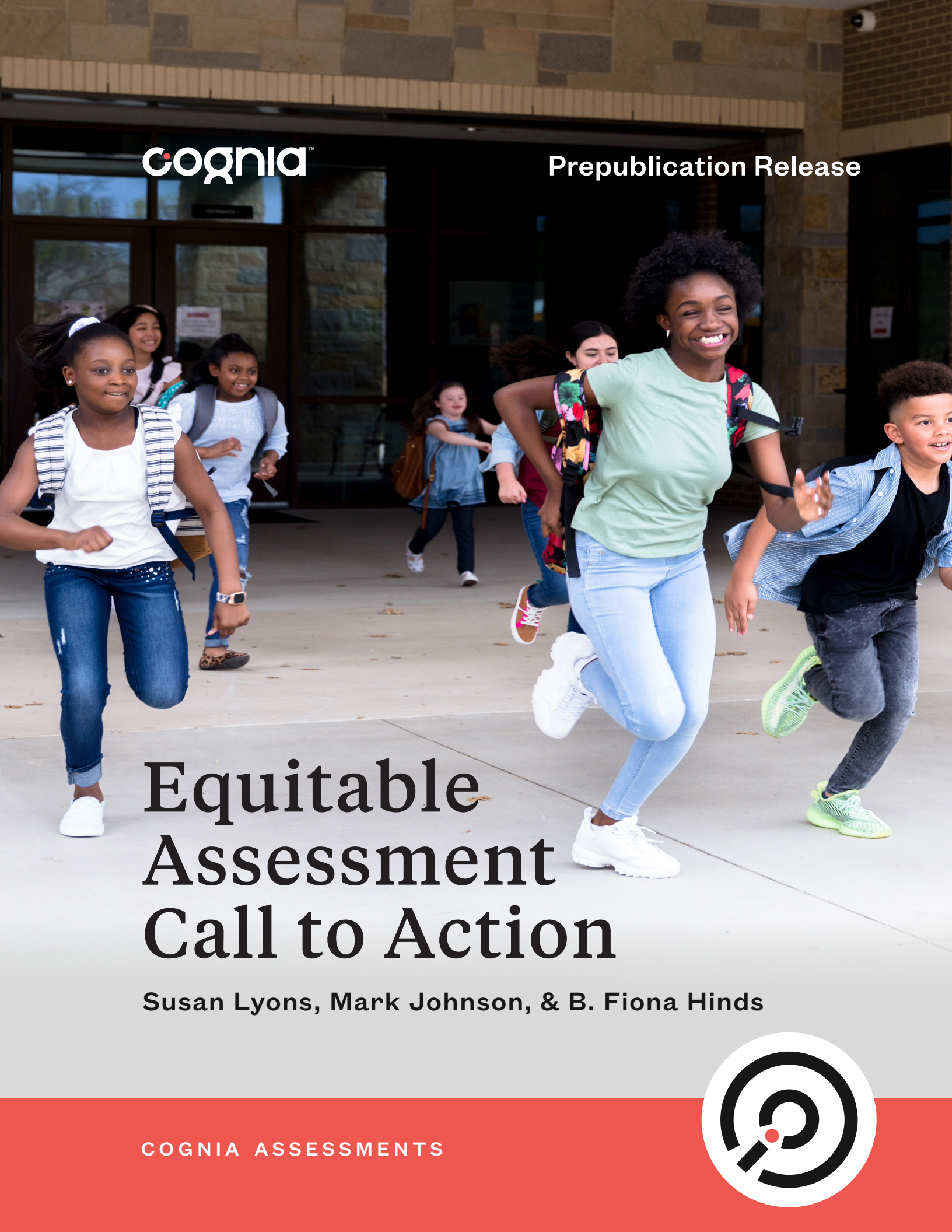


cognia[™]

Prepublication Release



Equitable Assessment Call to Action

Susan Lyons, Mark Johnson, & B. Fiona Hinds

COGNIA ASSESSMENTS





Susan Lyons

Susan Lyons, Ph.D. is the Principal Consultant for Lyons Assessment Consulting. Dr. Lyons works at the intersection of educational measurement and social justice, partnering with clients to provide thought leadership, design systems, lead research, and offer technical advice that leverages the power of assessment to create a more equitable future. Dr. Lyons teaches at Boston College and is the Executive Director of Women in Measurement, Inc., a nonprofit organization aimed at advancing gender and racial equity in the field of educational measurement.



Mark Johnson

Mark Johnson is the Senior Advisor for Content at Cognia. After 17 years as a middle-school teacher, he held several roles in state-level student assessment before joining Measured Progress, now Cognia. For the past 10 years, he has been engaged in the work of educational improvement with Cognia. Mr. Johnson is inspired by the opportunity to bring his previous career and life experiences to this work, to closely examine and act upon the importance of diversity, equity, accessibility, and inclusion in educational assessment.



B. Fiona Hinds

Fiona Hinds, Ed.D., is the Senior Advisor for Equity and Transformation at Cognia. Dr. Hinds is a passionate equity advocate with extensive experience in educational leadership, improvement systems, and quality schools. Dr. Hinds served as an adjunct professor and program designer for a university STEM program initiative for urban middle school Black girls. Dr. Hinds is committed to dismantling gender and racial inequities in educational measurement and serves as the Chief Strategy Officer for the nonprofit Women in Measurement, Inc.

To contact the authors, email equitable-assessment@cognia.org

Equitable Assessment Call to Action

Susan Lyons, Mark Johnson, & B. Fiona Hinds

COGNIA ASSESSMENTS



Table of Contents

| | |
|--------------------------------------------------------------------------------|-----------|
| Introduction | 1 |
| The moment is now | 1 |
| Confronting the history and legacy of racism in standardized assessment | 1 |
| Cognia's approach | 3 |
| A call to action | 3 |
| References | 4 |
| MODULE 1 Moving Toward Culturally Sustaining Classroom Assessment | 5 |
| Understanding the sociocultural embeddedness of learning | 5 |
| Defining the features of culturally sustaining assessment | 6 |
| Discussion Questions | 8 |
| References | 9 |
| MODULE 2 Innovating in Large-Scale Test Design | 11 |
| Embracing multiculturalism within the assessed constructs | 11 |
| Recontextualizing item writing | 12 |
| Diversifying the workforce at testing organizations | 13 |
| Refining item and test bias analyses | 14 |
| Returning to matrix sampling | 14 |
| Discussion Questions | 16 |
| References | 17 |

| | |
|-------------------------------------------------------------------------------------|-----------|
| MODULE 3 Reconceptualizing Psychometrics | 19 |
| Interrogating our quantification methodologies | 19 |
| Changing our conceptions of comparability | 20 |
| Pursuing promising psychometric advances | 21 |
| Shifting to pragmatic evaluation. | 23 |
| Discussion Questions | 24 |
| References | 25 |
| MODULE 4 Reframing Reporting | 27 |
| Promoting a Growth Mindset | 27 |
| Changing the discourse on group test score differences | 28 |
| Discussion Questions | 30 |
| References | 31 |
| MODULE 5 Addressing Inequities in Test Use | 33 |
| Taking a stance against test uses that perpetuate inequity | 33 |
| Collecting evidence related to the consequences of test use for racial equity | 34 |
| Centering racial justice in accountability system redesign | 35 |
| Discussion Questions | 37 |
| References | 38 |

Introduction

The moment is now

The effects of structural racism run through all American institutions, including schools. The global pandemic and racial protests of 2020 ushered in a renewed commitment to examine systemic racism and root out its underlying causes and perpetuating forces in all areas of society.

We believe educational assessment is a powerful tool that can help advance racial equity, but we must first reckon with the damaging role it has played in reinforcing patterns of privilege and oppression. In this paper we grapple with the history and legacy of racism in educational assessment and introduce Cognia's Equitable Assessment Call to Action. The Equitable Assessment Call to Action deeply probes five opportunities to place racial justice at the center of the design and use of educational assessments.

We use racial justice for Black and indigenous people and other marginalized people of color as the lens through which we critique current practices and offer ways to move forward. Of course, problems related to equity are not limited to those of racial injustice, but this Call to Action specifically explores race-related issues, in the hope that dismantling the structures in assessment that oppress racial minorities will provide pathways for addressing other forms of oppression in our society. The Call to Action is a starting point for meaningful conversation and innovative ideas to advance practice in educational measurement toward a more equitable future.

Confronting the history and legacy of racism in standardized assessment

"We are a country founded on the genocide of one people and the enslavement of another" (Ortiz, 2016). The structural oppression of non-Whiteness¹ was an organizing feature in the design of our country, and we're still grappling with those effects today. The overt racism in the formation of our American institutions persists in ways that continue to perpetuate structures of power and social control. This includes American schooling and the design and use of tests, which have become central fixtures in our schools.

Origins

In the United States, standardized testing first emerged in the early 20th century to measure intelligence. At the time, intelligence was understood to be an inherited trait that could be measured and reported with a single number, identifying those who possessed more and those who possessed less. The founders of the first intelligence tests believed that their instruments provided scientific evidence of a racial hierarchy in human intelligence, an idea that has been widely criticized but still finds supporters today (Herrnstein & Murray, 1996; Reynolds & Suzuki, 2012). The tests were used to advance notions of social Darwinism: that those with the most social and economic

¹ For the purposes of this Call to Action we will be capitalizing references to Whites and Whiteness to acknowledge Whites as a race in the same way we acknowledge Blacks. Treating "White" as a proper noun highlights the role of White people in conversations about race and removes the White-centered neutrality of a lowercase designation (see Appiah, 2020).

power had rightfully assumed these positions due to their superior intellect (Goldman, 1952).

The rise in use of standardized assessments coincided with the social efficiency movement, where tests were used to measure, rank, and sort individuals for schooling and training that aligned with their inborn ability, and therefore their optimal function in society (Silverberg, 2008). A central assumption in the use of tests in this manner is that intelligence is fixed, unchangeable by environment or education, and therefore educating those with lower IQs is futile and a waste of resources.

Continued misuse

The legacy of scientific racism in the design and use of our psychoeducational instruments persists today. While our theories about intelligence and learning have advanced, our current psychological and educational assessment tools have not, and rest on many of the same notions of mental measurement posited by the eugenicists who first developed them. Test designers and psychometricians have made theoretical and statistical advances in conceptualizing and detecting test bias, but our notions of quality and fairness in assessment have not kept pace with our understanding of the central role that culture plays in learning.

Similarly, while many of the early uses of psychological tests—such as their role in the eugenics movement—are now considered grotesque, psychoeducational instruments continue to be used and misused in ways that perpetuate structures of racial inequity. The role of assessment in our educational system today serves to “freeze” the social order by codifying the values, priorities, and ways of being of those in power, and enabling policies that continue to stratify society in those terms (Dixon-Román & Gergen, 2013). Ironically, due to their presumable objectivity, educational assessments simultaneously perpetuate the public perception of schooling as a meritocracy rather than as a form of systemic oppression that provides unequal access to educational opportunities (Mehan, 2008).

Given the high stakes of the current uses of educational assessment for our students, schools, and society, we must more fully examine the notion of why racial group test-score differences continue to persist. Many measurement experts will rightly point out that some of the group-score differences can be explained by economic inequities that systematically distribute to Whites but withhold to

non-Whites access to resources such as highly qualified teachers, quality healthcare, and nutritious food. However, socioeconomic status does not fully account for the score differences between Black and White students, nor does level of parental education (Hughes, 2003; Lubienski, 2002; Nettles, 2000; NAEP, 2009). Instead, cultural and racial factors bear directly on the performance of students of color, leading to serious questions about the validity of score interpretations in a multicultural society. Research has shown that culture has a significant effect in how test takers approach standardized tests and on test-taking performance (Arbuthnot, 2020). Additionally, the cumulative effects of systemic racism deliver toxic messages to students of color about their abilities and cause psychological and biological stress, also leading to disparities in performance on tests (Mendoza-Denton, 2014; Levy, Heissel, Richeson & Adam, 2016).

As measurement professionals, it is our responsibility to examine all the possible explanations for score variance that might contribute to the persistent racial group-score differences. While some factors are beyond our control, we must boldly address those factors that we can control, such as the opportunities presented in this Call to Action. As Christine Ortiz of Equity Meets Design puts it, “racism and inequity are products of design. They can be redesigned” (2016, para. 3).

Advancing practice toward more equitable assessment

This Call to Action offers a deep look at five areas of opportunity for advancing racial equity in educational measurement, described below. Each module outlines the major ideas for advancing practice and includes a discussion guide for fostering conversation and a list of references.

Module 1. Moving toward culturally sustaining classroom assessment

We provide an overview of sociocultural learning perspective and how it directly relates to transforming classroom assessment practice. We argue that culturally sustaining classroom assessment has the power to advance racial equity, based on three defining characteristics: students are valued, engaged, and empowered.

Module 2. Innovating in large-scale test design

We discuss advances needed in large-scale test design to put racial justice at the center. These advances include:

- Embracing multiculturalism within the assessed constructs
- Recontextualizing item writing
- Diversifying item writing and test development teams
- Refining our bias detection methods to acknowledge intersectionality
- Advocating for a return to matrix sampling

Module 3. Reconceptualizing psychometrics

We explore how our psychometric methods and values contribute to perpetuating racial power structures; specifically, we question the legacies of racism in our most foundational quantification methods, challenge our beliefs about the optimal conditions for achieving score comparability, and highlight promising areas of innovation in psychometric practice.

Module 4. Improving reporting

We challenge the language used to describe individual student achievement and group-level performance. We show how score reporting choices can undermine racial equity by perpetuating notions of fixed intelligence and reinforcing negative racial stereotypes.

Module 5. Addressing inequities in test use

We suggest that the measurement community take a vocal stance against test uses that perpetuate inequity, broaden and routinize our collection of evidence related to the individual and societal consequences of test use, and advocate for redesigning our school accountability systems to expressly empower communities that have been systemically oppressed.

Cognia's approach

Cognia™ employs a principled approach to assessment design, development, and implementation (PADDI; Ferrara, Lai, Reilly, & Nichols, 2016). At the time of this publication, we are integrating common elements of principled approaches into our procedures and infrastructure, approaching our work on equitable assessment following PADDI. We are implementing processes to identify the specific targets of our work and address the intended score interpretations and uses of assessments. All decisions

regarding design, research, and other activities are aligned with this approach.

This Call to Action does not provide specific solutions, nor imply that there are quick fixes to systemic issues. Instead, it signifies Cognia's commitment to keep *all* students at the center of our work, while we embark on a journey to evaluate claims that challenge our current thinking; enact a robust research agenda, and apply our learning to create equitable assessment practices that serve and empower every learner.

A call to action

Cognia's mission is to improve educational outcomes for *all* students. Meaningful assessment has always been a part of that pursuit. We are proud to now apply our efforts and expertise to the work of developing and enhancing equitable assessment. This Call to Action encourages all in the field to thoroughly investigate claims of damage to marginalized students of color caused by current practices—and to ask difficult questions and reflect on our own work and attitudes. Then we must proactively research, develop, and ensure the efficacy of proposed remediations.

We offer these challenges in a spirit of truth-seeking and collaboration, and invite to the table all those who wish to have a voice in how we collectively can—and must—better serve students who have been structurally disadvantaged in our education and educational assessment systems.

We don't claim to have all the answers. Instead, we assert a commitment to identify needs for change and to develop corresponding solutions grounded in research. We offer this Call to Action as an initial step in this important work, and as a conceptual draft. We welcome feedback. We envision that this will become a robust body of work to continue to engage in over time, and we look forward to the contributions of those who wish to help grow this work.

If you, too, hear the call, please join in the effort as we seek truth, challenge the status quo where needed, and develop better methods. If you are undertaking initiatives to dismantle the systemic practices in assessment that perpetuate racism, we want to know about and support your efforts. Please email us at equitable-assessment@cognia.org.

References

- Appiah, K. (2020, June 19). The Case for Capitalizing the 'B' in Black. Retrieved December 07, 2020, from <https://www.theatlantic.com/ideas/archive/2020/06/time-to-capitalize-blackand-white/613159/>
- Arbuthnot, K. (2020). Reimagining Assessments in the Postpandemic Era: Creating a Blueprint for the Future. *Educational Measurement, Issues and Practice*.
- Dixon-Román, E. J., & Gergen, K. J. (2013). Epistemology in measurement: Paradigms and practices. Vol. Princeton NJ: *The Gordon Commission*.
- Goldman, E. (1952). *Rendezvous with destiny*. New York. Random House.
- Herrnstein, R. J., & Murray, C. (1996). *The bell curve: Intelligence and class structure in American life*. Simon and Schuster.
- Hughes, S. (2003). An early gap in Black-White mathematics achievement: Holding school and home accountable in an affluent city school district. *The Urban Review*, 35, 297–322.
- Levy, D. J., Heissel, J. A., Richeson, J. A., & Adam, E. K. (2016). Psychological and biological responses to race-based social stress as pathways to disparities in educational outcomes. *American Psychologist*, 71(6), 455.
- Lubienski, S. T. (2002). A closer look at Black-White mathematics gaps: Intersections of race and SES in NAEP achievement and instructional practices data. *Journal of Negro Education*, 269–287.
- Mehan, H. (2008). A sociological perspective on opportunity to learn and assessment. In P. A. Moss, D. C. Pullin, J. P. Gee, E. H. Haertel, & L. J. Young (Eds.), *Assessment, equity, and opportunity to learn* (pp. 42–75). Cambridge: Cambridge University Press.
- Mendoza-Denton, R. (2014). A social psychological perspective on the achievement gap in standardized test performance between White and minority students: Implications for assessment. *The Journal of Negro Education*, 83(4), 465–484.
- National Academies of Sciences, Engineering, and Medicine. (2018). *How people learn II: Learners, contexts, and cultures*. National Academies Press.
- National Assessment of Education Progress (NAEP). (2009). Washington, DC: U.S. Department of Education.
- Nettles, M. (2000, October). *The status and consequences of admissions test performance for the nation's demographically diverse population of aspiring students*. Paper presented at the Fulbright's Educational Experts Seminar, Ann Arbor, MI.
- Ortiz, C. (2016, November 15). Racism and inequity are products of design. They can be redesigned. Retrieved from <https://medium.com/equity-design/racism-and-inequity-are-products-of-design-they-can-be-redesigned-12188363cc6a>.
- Reynolds, C. R., & Suzuki, L. A. (2012). Bias in psychological assessment: An empirical review and recommendations. *Handbook of Psychology, Second Edition*, 10.
- Silverberg, C. (2008). *IQ testing and tracking: The history of scientific racism in the American public schools: 1890–1924*. University of Nevada, Reno.

MODULE 1

Moving Toward Culturally Sustaining Classroom Assessment

Classroom assessment is the most important place to begin the work of addressing systemic inequities. We argue that culturally sustaining classroom assessment practice and good instruction are mutually supportive in a learning environment.

This module explores how culturally sustaining assessment attends to the social and cultural embeddedness of learning through explicit demonstration of:

- Valuing student cultures and identities
- Supporting student agency by engaging students in the assessment process
- Fostering critical consciousness and social action through problem-based tasks

These three attributes of culturally sustaining classroom assessment support the students' development as cultural, socio-politically situated beings who have agency to confront structural systems of power and oppression and effect change (Behizadeh & Pang, 2015).

Understanding the sociocultural embeddedness of learning

Culturally sustaining classroom assessment is grounded in the sociocultural perspective on how people learn. Humans learn by developing and maintaining mental schema that support sense-making. We integrate new knowledge by searching for meaning and relevance, building on our prior understandings organized in mental structures informed by our lived experiences and social interactions (see National Academies of Sciences, Engineering, and Medicine, 2018). Culture has a deep impact on the way our

mental schema develop through the information “we’ve taken in, interpreted, and categorized based on our cultural norms, beliefs, and ways of being” (Hammond, 2014, p. 23).

To understand cognition and academic achievement, we must also understand the context in which students are constructing knowledge (Basterra, 2011). Any act of cognition can be interpreted as a specific reaction to an individual set of social and cultural experiences (Resnick, 1991). Schools with strict or narrow ways of operationalizing standards for behavior and learning—standards that privilege White cultural norms and ways of knowing—often mistakenly treat cultural differences in the ways students learn as learning deficits (Graham, 2020; Hammond, 2014).

Sociocultural learning perspective also accounts for the role of the student in the learning process, on the basis of an understanding that learning is dependent on student motivation, engagement, and sense of efficacy (National Academies of Sciences, Engineering, and Medicine, 2018). Authentic engagement in the practices of the academic disciplines relies on the emotional, motivational, and relational aspects of the student’s identity, not only on cognitive resources (Holland & Lave, 2009; Shepard, 2019). Structural racism and implicit bias that manifest throughout the schooling and life experiences of students of color undermine the development of academic mindsets for learning (Hammond, 2014).

Culturally sustaining assessment validates the cultural embeddedness of learning and explicitly attends to the sociopolitical reality of students in marginalized populations. It affirms their cultures and identities, creates counter-narratives, and ultimately builds student agency for understanding, critiquing, and confronting systems of social injustice.

Defining the features of culturally sustaining assessment

What are the features that define culturally sustaining assessment practice? We offer an approach with three requirements: students are valued, students are engaged, and students are empowered.

Students are valued

Culturally sustaining classroom assessment occurs within a learning environment where students are safe to be who they are, feel valued, and have a sense of belonging to the learning community (Mendoza-Denton, 2014). Culturally sustaining classroom assessment affirms students of color as descendants of people with rich intellectual histories to be studied and carried forth (Lee, 1998). Prior knowledge, lived experiences, and students' homes and communities are sources of relevant expertise that contribute to meaning making and understanding of the academic content (Moll, Amanti, Neff & Gonzalez, 1992).

Culturally sustaining classroom assessment draws on the cultural wealth that students bring with them into the classroom, for example the skills and assets associated with speaking more than one language or language variant, being connected to a large familial or community network, and the ability to resist negative messages about one's own value or prospects (Yosso, 2005). Culturally sustaining assessments are designed to recognize the brilliance of Black, brown, and indigenous children—interrupting the marginalization and criminalization of children in the classroom (Stuart-Wells, 2019).

For example, as teachers gather evidence of prior knowledge when they begin an instructional unit, the students' own funds of knowledge should be invited into the curriculum (Cowie, Gerzon & Jones, 2020). What experiences or concepts from their histories, homes, and communities can inform their learning? Heritage and Harrison (2020) offer the example of creating a learning environment that encourages a student to connect her observations from gardening with her grandmother to making hypotheses about how plants get food, as part of an introduction to photosynthesis. In this simple example, the teacher's pre-assessment—a discourse about prior knowledge—draws on the diversity of experiences,

skills, and values that students bring with them into the classroom, contributing to their academic learning.

Students are engaged

Culturally sustaining assessments engage students in demonstrating their knowledge in ways that are connected to who they are. Culturally sustaining assessments provide opportunities for meaningful engagement with the academic content and disciplinary practices as well as involving students in the assessment process. Students are engaged as agents of their own learning through embedded opportunities for peer feedback, self-assessment, and even student-led assessment. For example, students can lead conferences with teachers and families in which they discuss their personal and academic goals, the schoolwork that is meaningful to them, and their learning progress. This is a powerful practice for supporting culturally sustaining assessment. Students and experts describe the power of student-led conferences [in this video](#).

Cognia has long emphasized the importance of student engagement. Cognia's Student Engagement Surveys ask students to reflect on and evaluate their experiences and opportunities related to learning, and their relationships with teachers, peers, and the broader school community.

Students are empowered

Culturally sustaining assessments focus on the development of the student as a cultural, situated being who has power to contribute to the confrontation of structural systems of power and oppression (Behizadeh & Pang, 2015; Penuel & Watkins, 2019). Culturally sustaining assessments empower students as agents of change in their lives and communities, to advocate for and advance social justice. In her anti-racist validity framework, Jennifer Randall challenges us to go beyond connecting the assessed content to students' lives by advancing anti-racism as part of the assessment process. For example, to assess knowledge of rates and proportionality, test items could be related to racial disparities in dollars earned per hour, maternal and infant mortality, COVID-19 death, and sentencing for crimes (J. Randall, personal communication, November 30, 2020).

Assessments that invite students to deeply engage with the content while also developing the students' sense of efficacy for creating change are often sustained, project-based or performance-based tasks. Culturally sustaining project-based assessments are opportunities for students

to apply and extend their learning through challenging real-world problem solving, often in collaboration with peers, teachers, or community members. Pullin (2008) suggests that socio-culturally informed learning, and by extension, assessment, fosters “the construction of deep understanding of meaningful knowledge within a learning community and develops transferable and transformative, reflective and critical thinking skills in a truly democratic context” (p. 350). For example, chemistry students at Leaders High School in Brooklyn, New York, worked with their peers and experts in the community to develop a new corrosive inhibitor that would prevent lead leaching into the water supply in Flint, Michigan—a powerful demonstration of the tragic and toxic effects of environmental racism (PBLWorks, 2019). In this powerful example, students are authentically engaged in the disciplinary practices of science—counteracting traditional notions of “who does science,” which Penuel and Watkins (2019) refer to as epistemic justice—all while engineering scientific solutions to advance social justice.

Keesing-Styles (2003) provides an elegant description of how these three features of culturally sustaining assessment (i.e., students are valued, engaged, and empowered) can work together to transform the classroom assessment experience for students and teachers:

To achieve a critical approach to assessment, [the classroom assessment experience] must be centered on dialogic interactions so that the roles of teacher and learner are shared and all voices are validated. It must foster an integrated approach to theory and practice, or what Freire would preferably term as praxis—theory in action. It must value and validate the experience students bring to the classroom and importantly, situate this experience at the centre of the classroom content and process in ways that problematize it and make overt links with oppression and dominant discourses. It must reinterpret the complex ecology of relationships in the classroom to avoid oppressive power relations and create a negotiated curriculum, including assessment, equally owned by teachers and students (para. 42).

Pedagogical reform movements related to culturally sustaining pedagogy, ethnic studies, expeditionary learning, assessment for learning, student-centered learning, and project-based learning each have the potential for transformative impact to improve racial equity within schools, in part through advancing culturally sustaining assessment practice. The challenge in the educational measurement community is to ensure our efforts in large-scale assessment are not undercutting local change efforts but instead provide support for and create coherence with these important instructional and assessment transformations. The next four modules in this Call to Action explore opportunities to adapt large-scale assessment practices to better reflect the sociocultural model of learning and to advance racial equity.

Discussion Questions

1. Describe a time in your educational experience when you felt motivated to learn. What was especially engaging about that learning experience? Why do you think your motivation was high?
 - a. Based on your group's responses to the first question, what do you think are the enabling conditions for a motivating learning experience?
 - b. How might we replicate those conditions in an assessment? What features of an assessment might be important for supporting student motivation and engagement?
2. Some of the earliest publicly funded schools in the United States were Indian Residential Schools where Native American children were separated from their families with the goal of assimilating them into the Euro-American culture. The Indian Residential Schools attempted to eradicate the indigenous culture and removed students' signifiers by cutting their hair, banning their native language, and changing their names.
 - a. In what ways does the legacy of forced assimilation continue in schools today?
 - b. How might culturally sustaining pedagogy and assessment change what we view as important in schools?
3. [Listen to Dr. Chris Emdin](#) talk about his work with seventh- and eighth grade students and [watch examples of their work](#). How do these examples embody the elements of culturally sustaining assessment discussed in this module?
4. Where do you see areas of opportunity in your own work to move towards more culturally sustaining assessments?
 - a. What incremental changes can you start making right away?
 - b. Where are the opportunities for more transformational change? What conditions and resources do you need to support those transformations?

References

- Baker-Bell, A. (2020). Dismantling anti-black linguistic racism in English language arts classrooms: Toward an anti-racist black language pedagogy. *Theory into Practice*, 59(1), 8–21.
- Basterra, M. R. (2011). Cognition, culture, language, and assessment. In M. Basterra, E. Trumbull, & G. Solano-Flores (Eds.), *Cultural validity in assessment: Addressing linguistic and cultural diversity* (pp. 72–95). New York: Routledge.
- Behizadeh, N., & Pang, M. E. (2016). Awaiting a new wave: The status of state writing assessment in the United States. *Assessing Writing*, 29, 25–41.
- Blanc, S., Christman, J. B., Liu, R., Mitchell, C., Travers, E., & Bulkley, K. E. (2010). Learning to learn from data: Benchmarks and instructional communities. *Peabody Journal of Education*, 85(2), 205–225.
- Cowie, B., Gerzon, N., & Jones, B. (2020, December 1). An ecological view of assessment for learning [Webinar]. WestEd. <https://www.wested.org/resources/ecological-view-of-assessment-for-learning/>
- Fine, M., & Ruglis, J. (2009). Circuits and consequences of dispossession: The racialized realignment of the public sphere for US youth. *Transforming anthropology*, 17(1), 20–33.
- González, N. (2005). Beyond culture: The hybridity of funds of knowledge. *Funds of knowledge: Theorizing practices in households, communities, and classrooms*, 29–46.
- Graham, E. J. (2020). “In Real Life, You Have to Speak Up”: Civic Implications of No-Excuses Classroom Management Practices. *American Educational Research Journal*, 57(2), 653–693.
- Hammond, Z. (2014). *Culturally responsive teaching and the brain: Promoting authentic engagement and rigor among culturally and linguistically diverse students*. Corwin Press.
- Heritage, M., & Harrison, C. (2019). *The power of assessment for learning: Twenty years of research and practice in UK and US classrooms*. Corwin.
- Holland, D., & Lave, J. (2009). Social practice theory and the historical production of persons. *Actio: An International Journal of Human Activity Theory*, 2(1), 1–15.
- Keesing-Styles, L. (2003). The relationship between critical pedagogy and assessment in teacher education. *Radical Pedagogy*, 5(1). Available at http://radicalpedagogy.icaap.org/content/issue5_1/03_keesing-styles.html
- Lee, C. D. (1998). Culturally responsive pedagogy and performance-based assessment. *Journal of Negro Education*, 268–279.
- Mendoza-Denton, R. (2014). A social psychological perspective on the achievement gap in standardized test performance between White and minority students: Implications for assessment. *The Journal of Negro Education*, 83(4), 465–484.
- Moll, L. C., Amanti, C., Neff, D., & Gonzalez, N. (1992). Funds of knowledge for teaching: Using a qualitative approach to connect homes and classrooms. *Theory into practice*, 31(2), 132–141.
- National Academies of Sciences, Engineering, and Medicine. (2018). *How people learn II: Learners, contexts, and cultures*. National Academies Press.
- PBLWorks. (2019, June 5). *Water quality project* [Video]. YouTube. https://www.youtube.com/watch?v=OE_GYEaq5Xg&feature=emb_logo

- Penuel, W. R., & Watkins, D. A. (2019). Assessment to promote equity and epistemic justice: A use-case of a research-practice partnership in science education. *The ANNALS of the American Academy of Political and Social Science*, 683(1), 201–216.
- Pullin, D. C. (2008). Assessment, equity, and opportunity to learn. In P. A. Moss, D. C. Pullin, J. P. Gee, E. H. Haertel, & L. J. Young (Eds.), *Assessment, equity, and opportunity to learn* (pp. 76–108). Cambridge: Cambridge University Press.
- Ruffin, T. M. (2020). Remove Systemic Barriers, Engage in Systemic Reform, and Implement Systemic Solutions: Transformative Justice, Good Teachers, and Identity Safe Classrooms. *Moja: An Interdisciplinary Journal of Africana Studies*, 1(1), 16–29.
- Shepard, L. A. (2019). Classroom assessment to support teaching and learning. *The ANNALS of the American Academy of Political and Social Science*, 683(1), 183–200.
- Steele, D. M., & Cohn-Vargas, B. (2013). *Identity Safe Classrooms, Grades K–5: Places to Belong and Learn*. Corwin Press.
- Stuart-Wells, A. (2019, April). *An inconvenient truth about the new Jim Crow of education*. Presidential Address at the Annual Meeting of the American Educational Research Association, Toronto, Canada.
- 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.

MODULE 2

Innovating in Large-Scale Test Design

In our first module we discuss strategies for moving toward culturally sustaining classroom assessment practice that engages students, values their cultural identities, and empowers students as agents of change in the world. While the classroom is arguably the most important place for this work, large-scale assessment has a major influence on classroom practice. How do we leverage the power of standardized assessment to signal the shifts that are important for classrooms and address the race-related inequities in large-scale test design? We offer five opportunities in the sections below:

- Embrace multiculturalism within the assessed constructs
- Recontextualize item writing
- Diversify the workforce at testing organizations
- Refine item and test bias analyses
- Return to matrix sampling

Embracing multiculturalism within the assessed constructs

In her keynote address at the 2019 NCME Special Conference on Classroom Assessment, learning sciences professor Bang called for a new view of disciplinary learning, arguing “we have often constructed disciplinary learning as acultural, or culturally neutral,” and in doing so, “we have often operationalized a false neutrality” (Bang, 2019). She introduced this idea to challenge the notion of objectivity in our current academic values and standards for learning. Claims to objectivity often serve as “a camouflage for the self-interest, power, and privilege of dominant groups in U.S. society” (Yosso, 2005, p. 7). This perspective challenges us to examine the centrality of Whiteness in the types of behaviors and learning we reward in schools. In their critique of the types of knowledge valued in higher

education, scholars Sensoy and DiAngelo (2017) make an argument that can be easily transferred to our current values in K–12 schools:

The modern university—in its knowledge generation, research, and social and material sciences and with its “experts” and its privileging of particular forms of knowledge over others (e.g., written over oral, history over memory, rationalism over wisdom)—has played a key role in the spreading of the colonial empire. In this way, the university has validated and elevated positivistic, White Eurocentric knowledge over non-White, Indigenous, and non-European knowledges (Battiste, Bell, & Findlay, 2002; Carvalho U& Florez-Florez, 2014; Grosoguel, Hernandez, & Velasques, 2016; Mignolo, 2002; p. 561).

Embracing multicultural ways of knowing in our schools would involve confronting the ways in which our current values and learning standards are steeped in White cultural norms. A July 2020 article in the New York Times Magazine offered in-depth insight into this perspective, elevating voices of those who contend that “society’s primary intellectual values are bound up with this marginalization”—of dismissing non-White ways of knowing (Bergner, 2020). As an example, the feature quotes Darnisa Amante-Jackson, founder of the Disruptive Equity Education Project, who states, “Eighty-eight percent of the entire world are people of color... 96 percent of the world’s historical content is white.” She contends that an overemphasis on the written word is a characteristic of Whiteness. Marcus Moore, from *Courageous Conversations*, puts it, “In school and on into the working world... tremendous harm is done by the pervasive rule that Black children and adults must ‘bend to whiteness, in substance, style and format.’” The feature ends with a perspective from Ibrahm X. Kendi, who says that it is incumbent on us to decide if we are a multicultural society that values multiple cultural standards and perspectives, or if we will continue as a unicultural nation where systemic racism is embedded in our most

foundational understandings of what is valued, taught, and rewarded (Bergner, 2020).

These issues persist despite the changing demographics of the U.S. In 2014, children of color comprised the majority of students in public schools in the U.S. for the first time (Avineri et al., 2015). Within the next 30 years, Whites will no longer be the majority population in this country (Johnson, 2014). The success of young people depends on our ability to embrace and adapt to increasing diversity and multiculturalism. “To achieve equity the curriculum needs to include valued knowledge and skills consisting of different kinds of cultural knowledge and experience, reflective of all groups, not privileging one group to the exclusion of others” (Klenowski, 2009, p. 83).

For example, Black students often face anti-black linguistic racism (Baker-Bell, 2020). White Americans have perpetuated linguistic hegemony, insisting that the dominant variety of English is the ticket for social and economic mobility (Avineri et al., 2015). Students speaking Black English arrive at school with five distinct present tenses only to be told that their language—a key part of their identity—is “less than” and wrong, and are forced to speak and write in White Mainstream English to avoid discrimination (Morrison, 2000; Baker-Bell, 2020). We must ask: Why are English language arts exclusively valued in our schools when students could be engaging in a rich curriculum centered around language arts more generally? In a truly multicultural language arts program, the knowledge and skills of students who are multilingual or who speak multiple dialects of English would be valued and developed as assets within our pluralistic society. Paris (2012) urges us to envision pedagogies that support linguistic pluralism by actively sustaining the cultural and linguistic competence of minoritized students while offering access to dominant cultural competence.

If the standards and assessments are designed to privilege the knowledge and culture of those in power, those with different experiences and values who are not members of the dominant group will continue to face barriers and cultural oppression (Klenowski, 2009). Re-examining our content standards through a lens of multiculturalism—valuing ethnic studies, cultural diversity, and multiple ways of being, doing, and knowing—would powerfully address the pervasive White hegemony in our schools and assessments.

Recontextualizing item writing

Developing items that connect to and affirm students’ cultures and identities puts racial justice at the center of large-scale tests. Currently, item writing processes for large-scale assessments include steps to guard against insensitive or culturally specific language that would introduce bias. However, the premise that all cultural context can be removed from an assessment is false. When we believe we can “decontextualize” our assessment items, we are simultaneously adopting outdated assumptions about the nature of competence (Resnick & Resnick, 1992) and developing “White” items (J. Randall, personal communication, November 30, 2020). Gordon, Aber and Berliner (2012) predict:

The exactness and precision gained by decontextualization in the past will be challenged by the situative and existential sensitivities that are necessary when contextualism and perspectivism are required for understanding, as well as knowing. It is worthy of recall that the reason qualitative methods became so prominent at the end of the 20th century was because we finally understood how inadequately we coped with contexts and multicausality.... Comprehensive and valid assessment in education will, in the future, have to be more sensitive to subjective phenomena, i.e., to affect, attribution, existential state, emotion, identity, situation, etc., as will also the teaching and learning transactions in which learners are engaged (p. 5).

What we know about how students learn today is much richer and more complex than what we understood when the practice of decontextualized, standardized assessment item writing first emerged. Sociocultural learning theory tells us that learning is inherently connected to student identity and culture, and it requires active participation of the learner. Students engage in knowledge building by integrating new content into existing schema—complex knowledge structures that students have formed to make meaning out of their lived experiences.

Assessment items that draw on and affirm students’ own experiences, cultural references, and emerging identities are likely to elicit rich information about student understanding and have the potential to disrupt the long-standing racial patterns in performance.

Research shows student performance improves when items are contextually appropriate for students by allowing for connections between the content and their lived experiences—a quality Solano-Flores (2011) refers to as cultural validity. Grounding the content in students' own real-life experiences not only makes it relevant for them, but also easier for them to connect and demonstrate their knowledge (Mislevy & Oliveri, 2019; van de Vijver, 2006).

Rather than trying to strip all cultural context from our assessments, an approach that would more closely align to this science of learning would be to meaningfully connect the items to familiar reference points and important aspects of Black and other minority students' lives. Solano-Flores and Nelson-Barber (2001) suggest we develop a common set of items that can be translated locally to fit the sociocultural context of students' lives. Mislevy and Oliveri (2019) offer the example of assessing algebra in the context of bus schedules and bus stops for students living in Los Angeles but changing the contextualization to assess the same content for Inuit students in the northwest.

Another way to ensure that students can see themselves reflected in the assessments would be to select passages and design items that not only represent diverse authors and perspectives, but attend to the sociopolitical context of the lives of students of color. For example, *What Lane?* by Maldonado features a black male protagonist who is experiencing a shift in how people's perception of him changes as he matures, from being a charming child to being a potential threat.

One of the most challenging aspects of recontextualizing assessment items will be figuring out how to appropriately modify items to reflect the authentic, cultural touchpoints and references within a diverse, multicultural society. Part of the solution is diversifying the workforce at testing organizations.

Diversifying the workforce at testing organizations

Research has documented just how challenging it can be to develop items that are culturally relevant for students. There are design blind spots when all of the item writers are from the same cultural tradition, especially when their cultural and lived experiences differ from the population of examinees. Designers of any kind have to make assumptions, often implicit, and in item design those assumptions may be limiting the relevance of the

assessment for students from non-dominant cultures. An analogy can be drawn to the well-documented limitations of facial recognition technology for female, non-binary, and non-White people (Lohr, 2018). In this case, these limitations are not intentional but are rather a reflection of the unknown blind spots of the mostly White male computer engineers who design the software.

As testing companies seek to diversify their content development teams, they find that people of color are disproportionately underrepresented in the pool of qualified candidates due to the systemic racism in our country that denies people of color the level of access to education and employment that Whites receive. Companies seeking to develop anti-racist hiring policies are examining their qualification criteria and onboarding practices to find opportunities to expand hiring beyond the current pool and to promote retention. For example, when hiring educators for item writing, a testing company may choose to prioritize candidates with direct experience working with youth in marginalized communities, extending beyond those who have formal teaching or assessment industry experience.

Reflecting the racial and cultural makeup of examinees among the item writing teams is only one step toward a more culturally valid assessment. To be effective, the work may require a more collaborative item-development process where item writers are sharing ideas and discussing items as a team at every step. Solano-Flores and Nelson-Barber (2001) even suggest including cultural anthropologists as part of the assessment development team to ensure that a sociocultural perspective is incorporated into decision-making throughout the assessment development process.

Diversifying the workforce will improve the practical abilities of measurement professionals to connect to and reflect the value of the experiences and ways of knowing of students who are culturally and linguistically diverse. Testing companies with diverse workforces will also benefit from the diversity itself through research-documented benefits such as increased creativity and better problem solving (Bowen & Bok, 1998). Additionally, Mendoza-Denton (2014) hypothesizes that there could be a powerful symbolic importance of signaling representation to minority examinees that may result in improved student performance.

Refining item and test bias analyses

Across U.S. society, we are becoming increasingly aware of how intersectionality affects all of our social and educational interactions. For example, Black children all suffer from the damaging effects of systemic racism. Black girls further experience a specific kind of sexism in a patriarchal society that undermines, devalues, and deprioritizes the perspectives, education, and health of black women (see Lorde, 2020). In other words, the intersection of race, gender, and class often lead to compounding (or exponential) forms of marginalization and oppression (Crenshaw, 1989).

Because knowledge acquisition is inextricably dependent on our social experiences in the world, the interpretive frameworks through which we read and respond to assessment items is shaped by our intersectional identities. Class, race, ethnicity, language, and gender diversity are all possible influences on the manner in which knowledge is acquired and demonstrated on an assessment (Gordon, 1995). This means that examining differential item functioning (DIF) separately by gender, socioeconomic status, and race is not only insufficient, but counter-productive in that cross-sectional views of item DIF are washing out the within-group intersectional effects such as low socioeconomic status among Black females (M. Russell, personal communication, October 27, 2020). Looking for differences in item functioning for each intersectionally defined subgroup will be challenging at an item-by-item level due to limitations in statistical power, however the assessment field should be able to quickly move to detecting intersectional effects in estimates of cumulative test bias, or differential test functioning.

For human-scored items, a proactive approach for helping guard against the potential for bias in would be to ensure that the anchor papers selected by rangefinders for each score or score range represent responses that reflect diverse experiences and context. The anchor papers are exemplars that inform how scorers assign a value to all responses. It is unacceptable practice to allow a single subgroup to comprise all or the majority of anchor papers for any score range. Doing so inherently limits the potential for diversity in how scores may be earned. Mandatory reporting of the demographic characteristics of the students whose responses are selected as anchor papers,

along with the criteria upon which these selections are based, must be standard practice.

These practices will provide greater insight into the degree of systematic bias in our reported scores. Detecting intersectional effects would push the field to understand and address the underlying causes of that bias. The difficult task of understanding the underlying causes of group differences in item response patterns is a reminder that equity and fairness in assessment are inherently qualitative concerns for justice; they are sociocultural issues, not only technical ones. (Stobart, 2005).

Returning to matrix sampling

As policymakers work to draft the next authorization of the Elementary and Secondary Education Act, the educational measurement community should more vocally advocate for the option of matrix sampling. The National Center for the Improvement of Educational Assessment and the National Research Council committee responsible for recommendations related to statewide science assessment (Marion, Domaleski & Brandt, 2020; National Research Council, 2014) have already voiced this recommendation. In the 1980s and -90s, many state assessment programs included matrix-sampling approaches to report on achievement at the school level—the unit of interest for our current models of state accountability under the Every Student Succeeds Act (see National Research Council, 2010). The primary benefit of a matrix-sampling approach is that it provides latitude for innovation in statewide standardized assessment content. Due to the very appropriate demand that the statewide assessment consume as little instructional time as necessary, these census tests typically include only the most efficient item types, and the fewest of those that support reliable results. Matrix sampling opens the door for exploring how to incorporate richer, more authentic, culturally sustaining tasks such as those we are advocating for in this Call to Action. Additionally, matrix sampling results in cost savings by reducing testing time and cutting down on the number of items to be scored (Shoemaker, 1975; Popham, 1993), which could free up resources for research, development, and implementation of assessment approaches centered on racial justice.

Those who oppose matrix sampling are adamant that estimates of student-level achievement are essential for ensuring that all students have the opportunity to learn and that the achievement of subgroups of students can be tracked over time. However, this is a false choice as subgroup performance can continue to be monitored using a stratified sampling approach in the matrix design. Beyond this, districts are moving toward more fully realizing balanced systems of assessment, making student scores from statewide summative assessments less necessary (National Research Council, 2001). Cognia collaborates with state and district partners to develop coherent balanced assessment systems that support and monitor learning at the student level. Cognia prioritizes the role of formative assessment within a broader system that also includes interim monitoring assessments. Local classroom assessments are better suited than statewide annual assessments to provide the timely and specific information that informs teaching and learning and provides progress updates to students, teachers, and guardians. The interim assessments function as point-in-time, external monitoring instruments that can provide overall achievement estimates relative to grade-level standards, when needed.

Removing student-level scores reports from statewide accountability systems also drastically reduces the potential for negative consequences associated with students' performance on the statewide assessments—negative consequences that are disproportionately experienced by students of color. The damaging student-level consequences resulting from performance on standardized assessments are far ranging. Examples of the domains that are impacted by individual-level reporting in high-stakes achievement tests include:

- Decreased feelings of self-efficacy, motivation, and engagement (Gergen & Dixon-Román, 2014)
- Limitations on educational attainment associated with test-based tracking, promotion, and graduation (Heubert & Hauser, 1999)
- Disciplinary retaliation by schools for poor performance in the form of suspensions, ultimately bolstering the school-to-prison pipeline for low-scoring students (Stuart-Wells, 2019; Annamma, Morrison & Jackson, 2014)
- Altered relationship dynamics between parents and children (Gergen & Dixon-Román, 2014)

Reporting standardized, criterion-referenced student achievement scores to families has some advantages, but we believe that those benefits do not outweigh the harms. Matrix sampling is a relatively simple technical solution for maintaining the policy priority of school-wide achievement monitoring while creating widespread, systemic transformation in the movement toward a more equitable future.

While we have a moral and perhaps existential imperative to elevate diverse voices, increasing diversity alone will not solve systemic issues of racial injustice in the field of educational measurement. We must all commit to raising the collective critical consciousness of the profession so that we can effectively transform the future of assessment into an actively anti-racist enterprise

Discussion Questions

1. In this module, we state that “Embracing multicultural ways of knowing in our schools would involve confronting the ways in which our current values and learning standards are steeped in White cultural norms.” This sentence is intentionally provocative, urging the educational measurement community to question the neutrality of our current content standards for non-White students.
 - a. Can you think of any examples of values expressed in schools that may be imposing and perpetuating White normative values?
 - b. How might we think about assessment design to broaden what we deem acceptable as evidence of learning?
2. Traditionally, content developers are trained that items should avoid culturally specific references and topics that may be more familiar to some examinees than others. This module challenges that notion, suggesting that the effort to remove cultural references is not possible and serves only to perpetuate the idea that dominant, White norms are “acultural.” Conversely, the module suggests that items should be as culturally relevant as possible for examinees.
 - a. What possibilities do you see with this approach to cultural relevance?
 - b. Can you envision opportunities for how you might recontextualize item writing?
3. This module presents strategies and rationales for diversifying the workforce at testing companies. What policies or programs may be effective for hiring a more diverse workforce?
4. The final two sections of the module offer technical solutions for addressing fairness and equity in educational assessment, which the module refers to as “qualitative concerns for justice.” This means that equity and fairness are not technical characteristics of an assessment; rather, they are the result of human decisions about the degree to which the assessment satisfies qualitative criteria for equity and fairness. What might be some criteria for fairness and equity that we would want to hold ourselves accountable to across our assessment programs?

References

- Annamma, S., Morrison, D., & Jackson, D. (2014). Disproportionality fills in the gaps: Connections between achievement, discipline and special education in the school-to-prison pipeline. *Berkeley Review of Education*, 5(1).
- Avineri, N., Johnson, E., Brice-Heath, S., McCarty, T., Ochs, E., Kremer-Sadlik, T., ... & Paris, D. (2015). Invited forum: Bridging the “language gap”. *Journal of Linguistic Anthropology*, 25(1), 66–86.
- Baker-Bell, A. (2020). Dismantling anti-black linguistic racism in English language arts classrooms: Toward an anti-racist black language pedagogy. *Theory into Practice*, 59(1), 8–21.
- Bang, M. (2019, September). *Making Assessment Responsive to Culturally and Linguistically Diverse Identities*. Keynote presentation at the 2019 NCME Special Conference on Classroom Assessment, Boulder, CO.
- Bergner, D. (2020). White fragility” is everywhere. But does antiracism training work. *The New York Times Magazine*, 15.
- Bowen, D., & Bok, G. (1998). *The shape of the river: Long term consequences of considering race in college and university admissions*, Princeton: Princeton University Press.
- Childs, R. A., & Jaciw, A. P. (2002). Matrix sampling of items in large-scale assessments. *Practical Assessment, Research, and Evaluation*, 8(1), 16.
- Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *u. Chi. Legal f.*, 139.
- Gergen, K. J., & Dixon-Román, E. J. (2014). Social epistemology and the pragmatics of assessment. *Teachers College Record*, 116(11), 1–22.
- Gordon, E. W. (1995). Toward an equitable system of educational assessment. *Journal of Negro Education*, 360–372.
- Heubert, J. P., & Hauser, R. M. (1999). *High stakes: Testing for tracking, promotion, and graduation*. National Academy Press, 2101 Constitution Avenue, NW, Washington, DC 20418.
- Johnson, J. (2014, November 10). Disruptive demographics. Retrieved from <https://www.kenan-flagler.unc.edu/news/disruptive-demographics/>
- Klenowski, V. (2009). Australian Indigenous students: Addressing equity issues in assessment. *Teaching Education*, 20(1), 77–93.
- Levy, D. J., Heissel, J. A., Richeson, J. A., & Adam, E. K. (2016). Psychological and biological responses to race-based social stress as pathways to disparities in educational outcomes. *American Psychologist*, 71(6), 455–484.
- Lohr, S. (2018, February 9). Facial recognition is accurate, if you’re a white guy. *The New York Times*. Retrieved from <https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html>
- Lorde, A. (2020). *The Selected Works of Audre Lorde*. (R. Gay, Ed.). W. W. Norton & Company.
- Marion, S., Domaleski, C., & Brandt, C. (2020). *Assessment and accountability recommendations for the next reauthorization of the elementary and secondary education act*. National Center for the Improvement of Educational Assessment. Retrieved from: https://www.nciea.org/sites/default/files/inline-files/Center%20for%20Assessment_ESEA.ReauthorizationReport_0.pdf
- Mendoza-Denton, R. (2014). A social psychological perspective on the achievement gap in standardized test performance between White and minority students: Implications for assessment. *The Journal of Negro Education*, 83(4), 465–484.

- Mislevy, R. J., & Oliveri, M. E. (2019). Digital Module 09: Sociocognitive Assessment for Diverse Populations <https://ncme.elevate.commpartners.com>. *Educational Measurement: Issues and Practice*, 38(4), 110–111.
- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press.
- National Research Council. (2010). *Best Practices for State Assessment Systems, Part I: Summary of a Workshop*. Washington, DC: The National Academies Press.
- National Research Council. (2014). *Developing assessments for the next generation science standards*. National Academies Press.
- Morrison, T. (2000). As quoted in, Rickford, J. R., & Rickford, R. J. *Spoken soul: The story of Black English*. New York, NY: Wiley.
- Paris, D. (2012). Culturally sustaining pedagogy: A needed change in stance, terminology, and practice. *Educational researcher*, 41(3), 93–97.
- Popham, W. J. (1993). Circumventing the high costs of authentic assessment. *Phi Delta Kappan*, 74, 470–470.
- Sensoy, Ö., & DiAngelo, R. (2017). “We Are All for Diversity, but...”: How Faculty Hiring Committees Reproduce Whiteness and Practical Suggestions for How They Can Change. *Harvard Educational Review*, 87(4), 557–580.
- Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 38(5), 553–573.
- Stobart, G. (2005). Fairness in multicultural assessment systems. *Assessment in Education: Principles, Policy & Practice*, 12(3), 275–287.
- Stuart-Wells, A. (2019, April). *An inconvenient truth about the new Jim Crow of education*. Presidential Address at the Annual Meeting of the American Educational Research Association, Toronto, Canada.
- van de Vijver, F. J. (2006). Cultural differences in abstract thinking. *Encyclopedia of Cognitive Science*.
- 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.

MODULE 3

Reconceptualizing Psychometrics

The first two modules in the Equitable Assessment Call to Action have focused primarily on the potential for test content and format to interrupt patterns of privilege and oppression in educational assessment. We now explore how our psychometric methods and values contribute to perpetuating racial power structures. We question the potential legacies of racism in our most foundational quantification methods and challenge our beliefs about the optimal conditions for achieving score comparability. Specifically, we explore:

- Interrogating our quantification methodologies
- Changing our conceptions of comparability
- Pursuing promising psychometric advances
- Shifting to pragmatic evaluation

Interrogating our quantification methodologies

In his 1997 chapter on postmodern test theory in *Transitions in Work and Learning: Implications for Assessment*, Mislevy compares the post-positivistic advances in our understandings related to human subjectivity and constructivism to the major 20th-century paradigm shifts in physics—moving from Newtonian and Euclidian understandings of the nature of matter to the shattering new ideas of relativity and quantum mechanics. It had been assumed that our understanding of the universe was limited only by the specificity of our instruments for measuring its components.

...improved instrumentation devised to finalize the modern research program revealed that its fundamental models were not in fact the universe's. Mathematical descriptions of observations departed increasingly from such intuitive notions as simultaneity

and definitive locations of persistent entities. Just as ironically, while we obtain better accuracy in modeling phenomena and more power to solve applied problems than the “modern” physicists of the nineteenth century dreamed, we feel farther away from ultimate understanding. The universe is not only stranger than we imagine, mused the mathematician J.B.S. Haldane, it is stranger than we can imagine! (Mislevy, 1997, p. 182).

As our methods in psychological and educational measurement have advanced, so have our understandings about the limitations of our psychometric scales for capturing the complexity of our cognitive resources. The sociocultural perspective complicates the interpretations we can draw from our current educational instruments.

Additionally, we are just beginning to understand how ideas related to racial hierarchies and White supremacy may have impacted the very core of our previously considered objective quantitative methodologies. Dixon-Román's July 2020 article, in *Educational Measurement: Issues and Practice*, calls into question our most foundational psychometric methods and values. Dixon-Román speculates about the connections between the racist ideologies of our psychometric founders and the assumptions about human difference that underlie the construction of widely used psychometric tools such as correlations and the logistic model. He posits that our quantification methods might have been constructed differently if the developers had not viewed Blackness as inferior, if difference had not been “scientifically” pathologized.

Modern psychological measurement assumes that human mental resources can be behaviorally observed and quantitatively ordered to reveal some meaningful estimation of truth—an assumption that can be legitimately questioned. (For a critical analysis of the measurability

hypothesis, see Mitchell, 1999.) Dixon-Román challenges us to question what our methods of quantitative analysis could be if we fully internalized and operationalized the notion that Black lives matter.

In her forthcoming book entitled “Discriminating Data” Professor Chun (2021) argues that the eugenicist history of our psychometric methods matters, “not because it predisposes all uses of correlation towards eugenics, but rather because when correlation works, it does so by making the future coincide with a highly curated past” (in press). Our factor analytic methods—based on correlations among student behavioral responses to a set of contrived stimuli—project those responses onto the most common underlying dimension that cuts across the set of items.

By engaging in a principled design process and gathering validity evidence, we make arguments that assert our degree of confidence that the unidimensional construct underlying item responses can be interpreted as student achievement in the targeted content domain. However, current methods for gathering validity evidence fail to question our reliance on correlation. To increase the reliability of our assessments, we remove or modify items that produce unpredictable responses, or “noise,” due to lack of model fit. We must examine the extent to which this paradigm of model “fit” and the professional goal to remove “construct-irrelevant variance,” is privileging a White, normative definition of academic achievement by excluding expressions of achievement that may reflect different ways of knowing. (See Randall & Huff, 2021.)

Moss, Pullin, Gee & Haertel (2005) remind us, “beliefs and practices informed by psychometrics have become so deeply engrained in the American educational system that it has become difficult to see them as choices arising in particular sociocultural circumstances or to imagine that things could be otherwise” (p. 66). In the next section, we offer the example of comparability as one psychometric construct that could be conceptualized differently.

Changing our conceptions of comparability

In psychometrics, comparability can imply score interchangeability, meaning equivalent scaled scores carry the same interpretations about what students know and can do regardless of the student or the form administered. However, this strict definition of comparability requires

robust evidence that is difficult to obtain in operational practice for statewide assessment programs. Instead of viewing comparability as a dichotomy, we have come to accept that it exists along a continuum with varying degrees to which scores can be meaningfully compared (Winter, 2010). Like validity, score comparability depends on the sufficiency of evidence for supporting the inferences and actions related to student performance based on the test scores (DePascale & Gong, 2020). In practice, this acknowledgement of comparability as a continuum permits practical test design while upholding an ideal, gold standard that would create the exact same conditions for all administrations, in an effort to control for all possible differences in achievement that are construct-unrelated (Geisinger, 2000).

Our notions of comparability in psychometrics derive from a positivistic epistemology that privileges a single, dominant interpretation of the construct and demands a single acceptable form of behavioral evidence to demonstrate achievement on that construct. A particular way of being and knowing and doing is codified in how items are developed to measure assumed-to-be universal constructs. The adequacy of the “one-size-fits-all” presumption of standardized assessment is challenged by the science of cognitive psychology (Moss, 1996); also, this way of instrumentalizing cultural understanding and the values of the dominant White culture is perpetuating White hegemony (Dixon-Román, 2019). Sireci argues that the hyper-focus on standardization in educational measurement leads to exclusion, “and the goal of educational measurement is not to measure the students who are easiest to measure and who conform to the most dominant culture associated with the measurement enterprise, but rather to obtain the best measure of each and every student’s proficiencies” (Sireci, 2020, p. 101).

DePascale & Gong (2020) note that we have already made significant progress in relaxing our notions of standardization in response to federal assessment requirements by providing accommodated and modified forms for students with disabilities. More broadly, the evolution of special education policy and practice in the United States may provide a roadmap for better reconciling the current structures of schooling and assessment with our democratic ideals (Pullin, 2008).

Rather than using standardization as our pathway to score comparability, we may instead choose to imagine a world in which we define comparability in terms of the degree

to which the assessment allowed for each student to meaningfully and authentically engage with the content to demonstrate their knowledge and skills. From a sociocultural perspective, a student is not presumed to have had the same opportunity to demonstrate their knowledge just because they are exposed to the same stimulus. Instead, the assessment environment must afford opportunities that attend to the learners' particular contexts (Gee, 2008). As Herman & Cook (2019) hypothesize in their chapter on fairness in Classroom Assessment and Educational Measurement, "By better responding to student identity, culture, interests, and the interactive processes through which students develop capability, variations in the surface features of an assessment—such as holding students to the same criteria but permitting choice—may yield a better and fairer estimate of student capability" (p. 261).

One example of a large-scale assessment program that challenges traditional conceptualizations of comparability is the AP Art Portfolios, in which students develop portfolios of their work over the course of the year to be evaluated using a common set of rating rubrics. Student writing accompanies the submissions, grounding the art in the local materials, processes, and ideas that informed the work. The meaning of scores is constantly mediated through conversations among the teacher scorers about how to evaluate and rate the immense diversity in submissions.

Mislevy (1997) characterizes the AP Art Portfolios as more of a social phenomenon than an exercise in measurement. If we are to truly examine psychometric assumptions underlying our current educational measurements—construct uniformity, cultural neutrality, decontextualization, unidimensionality, quantitative objectivity, and many others—it may be more useful to recast all of our efforts in educational assessment as social phenomena, rather than exercises in scientific measurement of mental attributes.

Pursuing promising psychometric advances

Advances in machine learning have created promising pathways to better account for the sociocultural ways in which people learn and for the limitations of our current methods for capturing the complexity of human understanding. Parisi & Dixon-Román (2020) envision a future where we abandon our desire to minimize error in

fitting our human-constructed models to the observed data and instead leverage the power of artificial intelligence to gain insight into the richness of human expression in the observed data, a process that would be fundamentally driven by what we now consider "noise." They critique existing computational models that serve only to reproduce human biases embedded within the data matrices the machine is given. Instead, Parisi & Dixon-Román (2020) have hope for computational, self-regenerating algorithms where what was previously understood to be statistical error, is now considered part of the spectrum of variation—where randomness is neither inside nor outside the model.

- Putting forth an approach to measurement modeling more closely within reach, Mislevy (2018) calls for "an argument-structured, socio-cognitively-framed, constructive-realist, subjectivist-Bayesian variant of latent-variable measurement modeling" (p. 415).
Argument-structured: The assessment design is supported by a validity argument that fully investigates the full range of possible explanations for observed item responses other than construct-related variance—such as negative stereotype bias.
- Socio-cognitively-framed: The contextual framing of assessment items reflects the lived experiences of the students, so that the content is connected to real-life situations that are familiar and meaningful to them, and assessment results provide for only partial and time-limited interpretations with situated meanings that depend on each student (Mislevy, Moss, & Gee, 2009).
- Constructive-realist: We understand that the students' real capabilities that they bring to bear when engaging in the assessment are not the same as the variables we have constructed to model and estimate the students' capabilities.
- Subjectivist-Bayesian: We can draw probabilistic approximations from data generated in complex environments in which students have different assessment experiences, producing different types of information.
- Latent-variable modeling: We no longer interpret our assessments as measuring some unseen attribute that exists to some degree in all students, but instead using statistical modeling tools to approximate unique patterns of resources shared by everyone.

Mislevy and Oliveri (2019) argue that this type of model contextualizing is especially important when constructs are most susceptible to differences in cross-cultural

interpretation, such as assessments designed to measure skills like collaboration. However, given what we know about the situative nature of all knowledge acquisition and sense making, it can be easily argued that we need to take this kind of deliberate approach to developing and interpreting our psychometric models for any assessments delivered in a multicultural society such as ours—particularly those used to make high-stakes decisions about students, schools, and systems.

Cognia adheres to a principled-design approach, namely the Principled Approach to Design, Development, and Implementation, or “PADDI” (Ferrara, Lai, Reilly, and Nichols, 2016). At a high level, the process follows these steps:

1. Define assessment targets, and intended score interpretations and uses (SIUs)
2. Develop assessment activities and testing procedures to provide information on intended SIUs
3. Select, develop, and implement aligned measurement models
4. Create validity argument-based technical documentation

When we understand the purposes and SIUs of our assessments, these purposes guide all processes and also logically and comfortably yield relevant validity arguments. When diversity, equity, accessibility, and inclusion are all targets, they will be reflected in all aspects of an assessment approach. As illustrated by the example below, if it is important that students can see themselves at the center of their assessments, there will be evidence of this throughout assessment design. Because these conditions can and should result in documentation that forms strong validity arguments, we hold that equity in assessment is a precursor to all aspects of a principled design approach, from the identification of assessment targets through the documentation of the validity of the assessment.

For example, if the target of a reading comprehension assessment is that students are able to cite textual evidence in support of inferences that one can make from a text, and if equity is a desired key component of the score interpretations and uses of the assessment, then acknowledgment of equity will be present throughout the approach. Selection of assessment texts will ensure diversity, accessibility, and inclusion, meaning that texts will be culturally relevant and responsive to the tested population and that content development and advisement will involve professionals with direct understanding the population’s culture. Further, all assessment items will be developed to align with the assessed standard as well as with the equity goals. Beyond considering how questions are asked, the development process will take into account how student responses will be scored. Diligently following these approaches will thus result in the ability to create validity argument documentation that emphasizes the overarching importance of equity.

No matter how complex the model and how nuanced our interpretations, however, there is and always will be “a fundamental disconnect between the personalized, content-based, actional information that stakeholders seek from assessment and the psychometric procedures designed to model group performance” (DePascale, 2020, para. 28). DePascale (2020) further warns that we must be humble in the claims that we make about individuals, understanding that within-group variation will always be greater than differences between groups, even as we design, report, and use assessments in more culturally responsive ways.

Shifting to pragmatic evaluation

While we work to advance our psychometric techniques to better reflect our latest understandings of the sociocultural-embedded ways people learn, and to fully account for the views of White supremacy held by our methodological founders, we must shift our view of current assessment practice from a public perception of scientific truth (e.g., “valid and reliable”) to something more pragmatic. We should encourage the perception of assessments as tools for indicating information that may be useful as part of an evaluation. They should be viewed as providing information that must be interpreted alongside qualitative approaches to understanding what students know and can do (Dixon-Román & Gergen, 2013).

Understandings of students' learning and programs' effects are enriched by multiple perspectives and diverse sources of evidence, some new or previously neglected but others with familiar (albeit reconceived) forms.... And as long as we in education purport to help other people's children learn, at other people's expense, we bear the duty of gaining and using as broad an understanding as we can to guide our actions and of conveying our reasoning and results as clearly as we can to those to whom we are responsible (Mislevy, 1997, p. 197).

If we understand that assessments do not reveal some unseen attributes inside the heads of students, but are instead instruments that can be practically useful, we then all assume the extra burden to fully understand the value of the reported scores and the consequences on individuals and society of their use (Gergen & Dixon-Román, 2014). Modules 4 and 5 explore issues of racial equity in score reporting and test use, respectively.

Discussion Questions

1. Despite advances in measurement, psychological tests of intelligence continue to produce scores that show racial group mean differences.
 - a. How do you make sense of this?
 - b. What conclusions can you draw about the limitations of our psychological instruments for revealing an approximation of the “truth” about the complex structures of human capacity?
2. This module proposes that we change our conception of comparability to be predicated on the degree to which the assessment allows for each student to meaningfully engage in demonstrating what they know and can do. This understanding of comparability would have implications for how we think about validity evidence.
 - a. What kinds of validity evidence would be useful for evaluating the degree to which an assessment experience allowed students to demonstrate their learning?
 - b. What inspiration can be drawn from the AP Studio Art assessment example for innovating in statewide standardized assessment?
3. This module offers a couple of examples of promising new advances in psychometrics.
 - a. What are you most optimistic about for the future of psychometrics?
 - b. What emerging technologies in educational assessment will help move us into a more equitable future?
4. The last section of this module encourages us to have more humility in reporting results by acknowledging the limitations of our instrumentation and encouraging the use of additional, more qualitative measures when making decisions about students and school programs.
 - a. How might these recommendations impact your work?
 - b. What implications do they have on the work of your organization?

References

Chun, W. (2021). *Discriminating Data*. The MIT Press.

DePascale, C. (2020, August 31). *I can see the writing on the wall* [Blog Post]. Retrieved from <https://charliedepascale.blog/2020/08/31/i-can-see-the-writing-on-the-wall/>

DePascale, C., & Gong, B. (2020). Comparability of individual students' scores on the "same test." In A. I. Berman, E. H. Haertel, & J. W. Pellegrino (Eds.), *Comparability of large-scale educational assessments: Issues and recommendations* (pp. 25–48). Washington, DC: National Academy of Education.

Dixon-Román, E. (2020). A Haunting Logic of Psychometrics: Toward the Speculative and Indeterminacy of Blackness in Measurement. *Educational Measurement: Issues and Practice*, 39(3), 94–96.

Dixon-Román, E. J. (2019). Validation as Hegemony: A Response to Camara et al. (2019). *Educational Measurement: Issues and Practice*, 38(4), 31–32.

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

Gee, J. P. (2008). A sociocultural perspective on opportunity to learn. In P. A. Moss, D. C. Pullin, J. P. Gee, E. H. Haertel, & L. J. Young (Eds.), *Assessment, equity, and opportunity to learn* (pp. 76–108). Cambridge: Cambridge University Press.

Geisinger, K. F. (2000). Psychological testing at the end of the millennium: A brief historical review. *Professional Psychology: Research and Practice*, 31, 117–118.

Gergen, K. J., & Dixon-Román, E. J. (2014). Social epistemology and the pragmatics of assessment. *Teachers College Record*, 116(11), 1–22.

Gordon, E. W., Aber, L., & Berliner, D. (2012). *Changing Paradigms for Education: From Filling Buckets to Lighting Fires to Cultivation of Intellectual Competence*. A Gordon Commission Report. Retrived from: https://www.ets.org/Media/Research/pdf/gordon_gordon_berliner_aber_changing_paradigms_education.pdf.

Herman, J., & Cook, L. (2019). Fairness in classroom assessment. In S. M. Brookhart & J. H. McMillan (Eds.), *Classroom assessment and educational measurement* (pp. 243–264). Routledge.

Mislevy, R. J. (1997). Postmodern test theory. *Transitions in work and learning: Implications for assessment*, 180–199.

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

Mislevy, R. J., & Oliveri, M. E. (2019). Digital Module 09: Sociocognitive Assessment for Diverse Populations <https://ncme.elevate.commpartners.com>. *Educational Measurement: Issues and Practice*, 38(4), 110–111.

Michell, J. (1999). *Measurement in Psychology: A Critical History of a Methodological Concept*. New York NY: Cambridge University Press.

Moss, P. A. (1996). Enlarging the dialogue in educational measurement: Voices from interpretive research traditions. *Educational researcher*, 25(1), 20–29.

Moss, P. A., Pullin, D., Gee, J. P., & Haertel, E. H. (2005). The idea of testing: Psychometric and sociocultural perspectives. *Measurement: Interdisciplinary research and perspectives*, 3(2), 63–83.

Parisi, L., & Dixon-Román, E. (2020). Data capitalism, sociogenic prediction and recursive indeterminacies. In P. Mörténböck & H. Mooshammer (Eds.), *Data publics: Public plurality in an era of data determinacy* (pp. 48–62). New York: Routledge.

- Pullin, D. C. (2008). Individualizing assessment and opportunity to learn. In P. A. Moss, D. C. Pullin, J. P. Gee, E. H. Haertel, & L. J. Young (Eds.), *Assessment, equity, and opportunity to learn* (pp. 76–108). Cambridge: Cambridge University Press.
- Randall, J., Huff, K. (2021). *“Color-Neutral” is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens*. Center for Educational Assessment Research Report No. 988.
- Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In *Changing assessments* (pp. 37–75). Springer, Dordrecht.
- Sireci, S. G. (2020). Standardization and UNDERSTANDARDization in Educational Assessment. *Educational Measurement: Issues and Practice*, 39(3), 100–105.
- Solano-Flores, G. (2011). Assessing the cultural validity of assessment practices: An introduction. In M. Basterra, E. Trumbull, & G. Solano-Flores (Eds.), *Cultural validity in assessment: Addressing linguistic and cultural diversity* (pp. 3–21). New York: Routledge.

MODULE 4

Reframing Reporting

In this module we challenge the language we use to describe individual student achievement and group-level performance. We show how language choices can undermine racial equity by perpetuating notions of fixed intelligence and reinforcing racial stereotypes. By changing its language, the education community can powerfully affect perceptions and realities related to persistent racial group disparities in test scores and the influence of stereotype threat on marginalized students of color. What role does the educational measurement community play in communicating to students about their capacity to learn? We argue that student-facing reports can be redesigned to implicitly and explicitly communicate these messages to students.

Social psychologists have found that non-White students' academic achievement is depressed, in part, by the cumulative effect of racist interactions that students of color experience throughout their schooling and lives. These interactions communicate negative messages about their intelligence and capabilities. The effect, known as "stereotype threat," results in an underestimation of performance for minoritized students, especially under high-stakes conditions (Aronson, 2002; Walton & Spencer, 2009).

Steele and Aronson (1995) show that performance for Black students worsens under high-pressure testing conditions—while dominant group performance holds steady, an effect attributed to the sensitivity of minoritized students to the psychological threat associated with confirming one's stereotype. A 2008 meta-analysis found that race-based stereotype threat has a dramatic effect on performance with a standardized effect size of .43 (Nguyen & Ryan, 2008). This large effect size provides a clear threat to comparability and calls into question the distribution of energy spent on addressing threats to comparability. Other, comparatively minor threats, such as item rendering differences across digital devices receive considerable attention and effort (see Dadey, Lyons & DePascale, 2018)

while little has been done to address the real and large impacts of stereotype threat in high-stakes testing.

Mendoza-Denton (2014) suggests that one of the most powerful ways we can counteract the effect of stereotype threat is by actively promoting the perspective that intelligence is malleable. Mendoza-Denton, Kahn, & Chan (2008) find evidence that assumptions of fixed intelligence contribute to widening the academic achievement gap by hampering performance in the context of a negative stereotype, while bolstering performance where there may be a positive stereotype (e.g., Asians are good at math). In this way, persistent cultural beliefs that intelligence is an unchangeable, inherited trait are likely contributing to score discrepancies by race, an effect that may be mediated by promoting the view that our cognitive resources can develop through the processes of learning.

Research has confirmed that the belief that one's intellectual abilities can be developed, referred to as a "growth mindset" (Dweck, 2008), is related to improved performance (Dweck, 2000; Claro, Paunesku & Dweck, 2016; Yeager et al., 2019). Growth mindset is associated with feelings of self-efficacy and student motivation—important factors in student learning (National Academies of Sciences, Engineering, and Medicine, 2018). Learning about the malleability of intelligence has been shown to be particularly powerful in improving outcomes for racial and ethnic minority students (Aronson, Fried, & Good, 2002; Blackwell, Trzesniewski & Dweck, 2007; Broda et al., 2018).

Promoting a Growth Mindset

Providing more useful feedback to students would support growth mindsets and student motivation. In large-scale summative assessment, one way we could improve the feedback we provide is by aligning the evidence of student learning to well-articulated learning progressions. If we begin the test design process by specifying models of

competence development within the domains, we could use those underlying theories of cognition to show students where they are currently in their learning and where they are going—describing how their expertise can develop. Learning progressions shift the paradigm from reporting a static measure of achievement to a more coherent framework for conceptualizing criterion-referenced growth (Briggs & Peck, 2015). Reporting student learning relative to learning progressions has the potential to blur the lines between summative and formative assessment, resulting in score reports that are useful for both providing information about achievement and supporting student learning.

Assessments based on learning progressions have exciting potential to support students as agents of their own learning, by clarifying learning goals and enabling self-assessment and self-regulated learning toward those goals as discussed in Module 1 (Goral & Bailey, 2019). These changes would benefit all students; however, they are likely to be particularly beneficial for students who continue to suffer from outdated and racist notions of inherited racial hierarchies in their capacity to learn.

Another, more immediately accessible opportunity for promoting growth mindset is in examining the language we use to communicate to students about their abilities in our achievement levels. O'Donnell (2020) looks at the achievement level labels in statewide assessments, labels that are ascribed to students an average of seventeen times between grades 3 and 12. While some of these labels may be favorable for promoting a growth mindset (e.g., approaching expectations, developing learner), many can be viewed as undermining a growth mindset (e.g., inadequate, unsatisfactory, and substantially below proficient). Rewording achievement-level labels to reflect learning as an ongoing process is a simple way to signal to students that their learning is not fixed nor finite.

These strategies offer opportunities for advancing racial equity through the language of assessment reporting at the individual level, but what about when we report and discuss test performance by groups? The next section of this module explores the power of our language when reporting and discussing disparities in test scores by race.

Changing the discourse on group test score differences

Equity-minded scholars have been urging us to change the way we label and discuss the persistent racial disparities in student performance, calling for us to shift to “systems centered language” by reorienting term achievement gap as education debt or opportunity gap (Flores, 2018; Ladson-Billings, 2006; O'Reilly, 2020). Public discourse related to score discrepancies has explicitly and implicitly attributed racial group differences in achievement scores to students' cultures, communities, or individual shortcomings, fueling racist ideas related to fixed hierarchies in intelligence by race (Kendi, 2016; Suzuki & Aronson, 2005; Valenica, 1997). Attributing score differences to deficits associated with individuals or groups ignores the potential race-related limitations in the measures themselves and, more broadly, overlooks the systemic racial oppression in our society that underlies persistent differences in assessment scores by race.

While changing the way we discuss group score differences may seem semantic and inadequate for addressing the underlying causes, research suggests that framing score discrepancies as the “achievement gap” contributes to racist stereotypes and demotivates people to make the systemic changes needed to close the gaps. Quinn (2020) found that participants who viewed media coverage discussing the Black–White achievement gap subsequently viewed Black students as less competent than their White peers. Shifting the language from the “achievement gap” to “racial inequality in educational outcomes,” however, increased the prioritization that educators place on addressing those inequalities (Quinn, Desruisseaux & Nkansah-Amankra, 2019). This implies that how we report racial disparities in achievement scores matters. It matters for how people view students and for garnering support to address the persistent differences in student outcomes. “It's an argument that we need to think carefully about the way we talk about [inequalities], because the way in which we talk about them affects the way that people understand the issues” (Mehnken, 2020).

The educational measurement community can be part of the solution by offering language that clearly describes differences in scores while avoiding language that contributes to perpetuating negative stereotypes. We can help lead the way in changing our discourse by offering anti-racist language in our score reporting, technical documentation, and research related to racial group disparities in achievement scores.

Discussion Questions

1. What is your understanding of how stereotype threat perpetuates and exacerbates the achievement gap?
 - a. Does it make sense for the assessment industry to identify and address this as a threat to score comparability and validity?
 - b. Why or why not?
2. Mendoza-Denton (2014) argues that promoting a growth mindset is an effective way to combat the negative effects of stereotype threat.
 - a. What is a growth mindset?
 - b. How might current practice in educational assessment undermine messaging related to growth mindset?
 - c. What opportunities exist for educational assessments to reinforce messaging that promotes growth mindset?
3. This module suggests that the educational measurement community has a role to play in offering anti-racist language to describe mean score differences by race—often referred to as the achievement gap.
 - a. How might using “achievement gap” language perpetuate inequities?
 - b. What language might we use instead when discussing racial group score differences?
4. Score reporting will continue to evolve with advancing technologies and with commitment to the actions described here.
 - a. What are you most optimistic about for the future of score reporting?
 - b. What opportunities do you see for score reporting to promote equity in assessment?

References

- Annamma, S., Morrison, D., & Jackson, D. (2014). Disproportionality fills in the gaps: Connections between achievement, discipline and special education in the school-to-prison pipeline. *Berkeley Review of Education*, 5(1).
- Aronson, J. M. (Ed.). (2002). *Improving academic achievement: Impact of psychological factors in education*. San Diego, CA: Academic Press.
- Aronson, J., Fried, C. B., & Good, C. (2002). Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence. *Journal of Experimental Social Psychology*, 38(2), 113–125. doi:10.1006/jesp.2001.1491
- Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development*, 78(1), 246–263. doi:10.1111/j.1467-8624.2007.00995.x
- Briggs, D. C., & Peck, F. A. (2015). Using learning progressions to design vertical scales that support coherent inferences about student growth. *Measurement: Interdisciplinary Research and Perspectives*, 13(2), 75–99.
- Broda, M., Yun, J., Schneider, B., Yeager, D. S., Walton, G. M., & Diemer, M. (2018). Reducing inequality in academic success for incoming college students: A randomized trial of growth mindset and belonging interventions. *Journal of Research on Educational Effectiveness*, 11(3), 317–338.
- Claro, S., Paunesku, D., & Dweck, C. S. (2016). Growth mindset tempers the effects of poverty on academic achievement. *Proceedings of the National Academy of Sciences*, 113(31), 8664–8668.
- Dadey, N., Lyons, S., & DePascale, C. (2018). The comparability of scores from different digital devices: A literature review and synthesis with recommendations for practice. *Applied Measurement in Education*, 31(1), 30–50.
- Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality, and development*. Psychology press.
- Dweck, C. S. (2008). *Mindset: The new psychology of success*. Random House Digital, Inc.
- Flores, O. J. (2018). (Re) constructing the language of the achievement gap to an opportunity gap: The counternarratives of three African American women school leaders. *Journal of School Leadership*, 28(3), 344–373.
- Goral, D. P., & Bailey, A. L. (2019). Student self-assessment of oral explanations: Use of language learning progressions. *Language Testing*, 36(3), 391–417.
- Kendi, I. (2016, October 20). Why the Academic Achievement Gap is a Racist Idea. Retrieved from <https://www.aaihs.org/why-the-academic-achievement-gap-is-a-racist-idea/>
- Ladson-Billings, G. (2006). From the achievement gap to the education debt: Understanding achievement in U.S. schools. *Educational Researcher*, 35(7), 3–12. doi:10.3102/0013189X035007003
- Mahnken, K. (2020, August 11). *The achievement gap has driven education reform for decades. Now some are calling it a racist idea*. The 74. Retrieved from: <https://www.the74million.org/article/the-achievement-gap-has-driven-education-reform-for-decades-now-some-are-calling-it-a-racist-idea/>.
- Mendoza-Denton, R. (2014). A social psychological perspective on the achievement gap in standardized test performance between White and minority students: Implications for assessment. *The Journal of Negro Education*, 83(4), 465–484.
- Mendoza-Denton, R., Kahn, K., & Chan, W. Y. (2008). Can fixed views of ability boost performance in the context of favorable stereotypes? *Journal of Experimental Social Psychology*, 44, 1187–1193.

- National Academies of Sciences, Engineering, and Medicine. (2018). *How people learn II: Learners, contexts, and cultures*. National Academies Press.
- O'Donnell, F. (2020). What's in a Label? Unpacking the Meaning of Achievement Labels from Tests. *Doctoral Dissertation*. https://scholarworks.umass.edu/dissertations_2/1856
- O'Reilly, M. (2020, June 5). Systems Centered Language: Speaking truth to power during COVID-19 while confronting racism. *Medium*. <https://medium.com/@meagoreillyphd/systems-centered-language-a3dc7951570e>
- Quinn, D. M. (2020). Experimental effects of “Achievement Gap” news reporting on viewers’ racial stereotypes, inequality Explanations, and inequality prioritization. *Educational Researcher*, 49(7), pp. 482–492.
- Quinn, D. M., Desruisseaux, T. M., & Nkansah-Amankra, A. (2019). “Achievement Gap” Language Affects Teachers’ Issue Prioritization. *Educational Researcher*, 48(7), 484–487.
- Shepard, L. A., Diaz-Bilello, E. K., Penuel, W. R., & Marion, S. F. (2020). *Classroom assessment principles to support teaching and learning*. Boulder, CO: Center for Assessment, Design, Research and Evaluation, University of Colorado Boulder.
- Sireci, S. G. (2020, September). *Psychometricians in the hands of an angry mob*. Presidential Address for the National Council on Measurement in Education.
- Steele, C.M. & Aronson, J. (1995). Stereotype threat and the intellectual performance of African Americans. *Journal of Personality and Social Psychology*, 69, 797–811.
- Suzuki, L. A., & Aronson, J. (2005). The cultural malleability of intelligence and its impact on the racial/ethnic hierarchy. *Psychology, Public Policy, and Law*, 11(2), 320–327.
- Valencia, R. R. (1997). Conceptualizing the notion of deficit thinking. In R. Valencia (Ed.), *The evolution of deficit thinking: Educational thought and practice* (pp. 113–131). London, England: Routledge Falmer.
- Walton, G. M., & Spencer, S. J. (2009). Latent ability: Grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science*, 20, 1132–1139.
- Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., ... & Paunesku, D. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, 573(7774), 364–369.

MODULE 5

Addressing Inequities in Test Use

In our final module of this Equitable Assessment Call to Action, we address how structures of privilege and oppression in our society are replicated through test use and misuse. Connell (1992) calls our approach to assessment hegemonic, meaning “both that it is culturally dominant, connected with the society’s central structure of power; and that it functions to maintain the social power and prestige of dominant groups” (p. 75). Since 1992, tests have become more prominent fixtures in schools and as the subject of public policy. We suggest that the measurement community:

- Take a vocal stance against test uses that perpetuate inequity
- Broaden and routinize our collection of evidence related to the individual and societal consequences of test use
- Advocate for redesigning our school accountability systems to expressly empower communities that have been systemically oppressed

Taking a stance against test uses that perpetuate inequity

The validity of an assessment depends on the quality of theory and evidence supporting the intended interpretations and uses of tests scores (Messick, 1989). This indicates that validity is not a feature of an assessment, but a judgment about the appropriateness of a particular interpretation or use of an assessment score. For example, there may be evidence that a sixth-grade math assessment produces scores that support valid claims about student proficiency. Different or additional evidence might be needed, however, to support the use of the scores for placing students into advanced coursework in mathematics. This use would require substantial

theoretical and empirical evidence due to the high-stakes nature of such placement decisions.

Misuses of educational assessment scores are far too common, and it has been shown that many of these assessment malpractices perpetuate racial inequities in our schools. This is in part due to the failure of assessment vendors to clearly articulate the range of uses an assessment is designed to support and to actively warn against potential misuse. The Standards for Educational and Psychological Testing (the Standards) clearly make it incumbent on the test developer to caution against potential uses that are not supported by validity evidence, which should be especially true for those uses with potential negative consequences for students (AERA, APA, & NCME, 2014).

One example of misuse is the pervasive practice of sorting and tracking students in schools. Extensive research indicates that tracking is not necessary for advancing learning and has devastating long-term consequences for racial equity in academic attainment, particularly in mathematics (Braddock, 1990; Burris, Wiley, Wilner & Murphy, 2008; Darling-Hammond, 1994; Glasner, 2018; Oakes, 1985; Smith-Maddox, 1998).

The continued inappropriate use of educational assessment to support and justify tracking is perpetuating racial inequities in schools. Research has shown that tracking disproportionately limits access to rigorous content and high-quality teachers for students of color (Heubert & Hauser, 1999). The use of tests in this case is particularly problematic in that it lends a perceived scientific legitimacy for creating a within-school stratification system—a system that is largely maintained due to the desires of middle- and high-income parents to gain advantage for their own children (Lucas, 1999). As Caroline Gipps (1999) puts it “examinations have a legitimating role in that they

allow the ruling classes to legitimate the power and prestige they already have” (p. 361).

The National Council of Teachers of Mathematics (2020) has taken a strong position against tracking in their latest book series, *Catalyzing Change*. Similarly, the American Educational Research Association has addressed value-added modeling and other controversial issues. The field and industry of educational assessment should similarly adopt a vocal stance against tracking and against the inappropriate use of psychological and educational assessments to support this practice. Statements disavowing specific test uses that perpetuate structures of racial inequity would send powerful messages to schools and educators who continue to use tests for this purpose.

Collecting evidence related to the consequences of test use for racial equity

There is a longstanding debate in the measurement community about whether the consequences of test use can impact the validity of the assessment. Despite the lack of consensus in this area, most experts agree with the notion that the consequences of test use are important and worthy of study. Even Cizek (2020), a vocal contrarian on the subject, emphasizes that “[c]onsequences must be incorporated in a comprehensive framework for defensible testing” (p. 88). Given this unanimity, assessment vendors should begin collecting evidence of the consequences of test use as part of routine technical maintenance of an assessment program. Evidence related to test consequences must prioritize the impact of test use on minoritized students.

Studying the impact of test use as part of principled approaches to assessment design, development, and implementation (Ferrara et al., 2016) would rapidly expand our understanding of the role of educational measurement in perpetuating or interrupting racial inequity. A Principled Design for Efficacy approach (PDE; Nichols, Ferrara, & Lai, 2016) that details a chain of reasoning to describe the intended outcomes, or consequences, of the assessment can be used to articulate of a theory of action for the testing program. As an example, the Nebraska Department of Education leads a multi-state collaborative in proposing a Stackable, Instructionally embedded, Portable Science (SIPS) assessment system in their 2020 application for a federal state assessment grant. The Nebraska

Department of Education (2020) uses a theory of action to hypothesize that the proposed assessment system will lead to “improved student achievement,” “enhanced student engagement in science,” and “enriched capacity for life-long science learning” (p. e26). As part of the evaluation effort for this assessment system, evidence related to these outcomes will be collected, and should be closely examined for differences in program impact across racial groups.

The Nebraska example highlights how principled assessment design approaches can articulate and evaluate the direct consequences of assessment use in creating equitable outcomes for students related to the programmatic goals (e.g., improved learning, improved engagement). Collecting and reviewing evidence in light of racial equity is a step in the right direction. But what about evaluating the impact of the assessment system on areas of influence that are outside the direct, intended effects of the assessment?

We argue that these types of indirect consequences—especially those with disproportionate negative impact on students of color—are also the responsibility of the measurement community. Some of these consequences include:

- Using tests to define expectations and determine access to quality courses (e.g., low-track placement, basic skills remediation, ineffective “test prep” pedagogy, scripted curricula), and and to assign rich curriculum (e.g., social studies, art, music, and physical education)
- Using tests to support models of schooling that perpetuate existing racial power structures (for example, see Graham, 2020)
- Using high-stakes tests to justify placement tracking and school segregation.

Gergen and Dixon-Román (2014) suggest we go even further in our evaluation of consequences and look at how our testing programs propagate cultural ideologies and societal structures that perpetuate inequality. For example, in allowing the dominant culture to define what is to be learned and how mastery is demonstrated, are we undermining cultural pluralism in our society? And what are the potentially damaging, long-term impacts on children when they internalize the evaluative hierarchy created by test scores?

A deliberate and consistent approach to gathering evidence of the direct and indirect outcomes of test use would greatly enhance the conversations about racial inequity at state departments of education, assessment companies, technical advisory committee meetings, and professional conferences. Among the measurement community and in broad public discourse, these evidence-based understandings and conversations are necessary to design improved solutions for a more equitable future.

Centering racial justice in accountability system redesign

Many education reformers argue that equity in educational opportunities is a driving factor in the design of federally mandated test-based accountability systems that have dominated the education landscape for the past two decades. Although test-based accountability has been effective at shining a spotlight on the glaring racial disparities in test scores present throughout our educational system, wholesale adoption of this reform movement has not been successful in inspiring the meaningful and lasting changes that would effectively address those racial disparities. Federally mandated accountability and test-based reform have not worked as intended, and in some cases, the disparities have even widened (Jennings & Sohn, 2014).

Additionally, a growing body of literature documents the harmful effects of top-down accountability that dictates strict rewards and consequences for schools. Not only are these effects well-documented and wide ranging, but they have disproportionate negative impacts on minoritized students and communities of color. Examples of these effects include:

- Overemphasis on low-level, remedial instruction in predominantly Black and Hispanic schools (Davis & Martin, 2008)
- Retaliation for low test scores using behavioral punishments disproportionately doled out to students of color (Stuart-Wells, 2019)
- Rise in models of schooling that shame and pressure students regarding test scores, and apply strict “White” behavior codes for students of color (Graham, 2020)

- Increased racial segregation in the composition of schools (Knoester & Au, 2014)
- School closures that disproportionately disrupt Black students’ schooling and their community engagement (Ewing, 2018)

Accountability is aimed at ensuring that federal Title 1 dollars are being spent wisely on effective programs that result in improved student test scores (DePascale, 2015). But the current systems are premised on a false assumption that school accountability for student achievement will inspire program effectiveness.

We are faced with the reality that accountability alone will not improve educational programs for students, no matter how severe the consequences. Raising the stakes on statewide assessment has led to misdirected and distorted efforts to try and improve performance on a narrow and incomplete set of indicators—resulting in the disproportionate negative impacts described above. For example, Pedulla et al. (2004) found that more than 60% of teachers reported teaching in ways that contradicted their own ideas of sound educational practice as a result of accountability testing, with percentages rising to 76% in schools that faced the harshest consequences. This may be partly explained by the widespread adoption of pedagogically constraining scripted curriculum products, which has been shown to be particularly oppressive in urban districts (Kavanagh & Fisher-Ari, 2020).

Centering our accountability systems on racial justice requires a reconceptualization of the mechanisms by which we expect to make meaningful progress. We need a different framework for approaching Title 1 program evaluation and for addressing the systemic inequities in educational opportunity and achievement. Gergen and Dixon-Román (2014) propose an empowerment evaluation approach. Fetterman (2001) provides the original definition of empowerment evaluation, which is “the use of evaluation concepts, techniques, and findings to foster improvement and self-determination” (p. 3).

The concept of empowerment evaluation was further clarified in Wandersman et al. (2005) by providing ten principles intended to guide the conceptualization and implementation of empowerment evaluations (Wandersman et al., 2005 as cited in Fetterman & Wandersman, 2007):

- Improvement
- Community ownership
- Inclusion
- Democratic participation
- Social justice
- Community knowledge
- Evidence-based strategies
- Capacity building
- Organizational learning
- Accountability.

Empowerment evaluation involves providing the program stakeholders, in our case communities and schools, with the tools to support ongoing self-evaluation and improvement (Wandersman et al., 2005). School accountability reimaged as an effort in empowerment evaluation would involve engaging with communities in understanding their goals, priorities, and values for schooling; partnering to provide resources and tools to formatively evaluate progress toward those goals; and ultimately benefitting all students through locally driven, sustainable school program improvements. Assessments of student learning would certainly continue to play an important role in empowerment evaluation efforts, but the nature of the evidence and the ways it is used to inform improvements may change dramatically.

Most state departments of education likely do not have the resources to engage with the communities they serve in the intensive manner that empowerment evaluation requires. However, state departments of education may be able to partner with technical providers to provide tools and supports for this work by facilitating “networked improvement communities” in which school districts collaborate as resources for one another within an empowerment evaluation framework (Bryk, Gomez & Grunow, 2011; C. Brandt, personal communication, January 11, 2021).

Our current model² of accountability was adopted with the passage of the No Child Left Behind Act in 2001. Conceptual and methodological advances achieved since then in the fields of assessment and evaluation can be leveraged to inform more equitable, enabling, and empowering policies than it was possible to envision 20 years ago. For the next reauthorization of the Elementary and Secondary Education Act, the educational measurement community should strongly advocate for federal policies that advance racial equity through new models of program evaluation. Assessment programs can and must redistribute power and resources to those who have been systemically oppressed and marginalized.

2 While features of accountability policy changed under the 2015 Every Student Succeeds Act, the same underlying model was maintained.

Discussion Questions

1. Educational assessment providers have a significant influence across many aspects of schooling.
2. What responsibility do educational assessment providers hold to ensure their products are not misused in ways that perpetuate racial inequity.
 - a. How might these organizations more effectively communicate with users about supported uses and warn against harmful misuses?
3. This module contends that collection of data related to the direct and indirect consequences of test use should be a routine part of the technical maintenance of any assessment program.
 - a. Do you agree?
 - b. Reflecting on your own work, what data related to the consequences of test use might be relevant to collect?
4. Test-based accountability is viewed by its proponents as a civil rights issue, ensuring that all students have an opportunity to learn rigorous academic content standards.
 - a. How do we reconcile the intentions of this policy with the real-world unintended consequences felt most acutely by communities of color?
 - b. How might we continue to collect data for tracking student achievement without the harmful effects of test-based accountability that we have seen over the last two decades?
 - c. What are the potential benefits and limitations of matrix sampling discussed in Module 2?
5. Just as Module 1 urges us to leverage students' power as agents of their own learning, this module argues that communities are the most important agents of improvement in an empowerment evaluation framework.
 - a. What kinds of supports, tools, and partnerships would encourage this kind of locally driven, continuous improvement?
 - b. What role might educational assessment play in an empowerment evaluation model?

References

- AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- American Educational Research Association. (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educational Researcher*, 44(8), 448–452.
- Boykin, A. W. (2014). Human diversity, assessment in education and the achievement of excellence and equity. *The Journal of Negro Education*, 83(4), 499–521.
- Braddock, J. H., II. (1990, February). *Tracking: Implications for student race-ethnic subgroups* (Report No. 1). Baltimore, MD: The John Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students.
- Bryk, A. S., Gomez, L. M., & Grunow, A. (2011). Getting ideas into action: Building networked improvement communities in education. In *Frontiers in sociology of education* (pp. 127–162). Springer, Dordrecht.
- Burris, C. C., Wiley, E. D., Welner, K., & Murphy, J. (2008). Accountability, rigor, and detracking: Achievement effects of embracing a challenging curriculum as a universal good for all students. *Teachers College Record*, 110(3), 571–607.
- Cizek, G. J. (2020). *Validity: An Integrated Approach to Test Score Meaning and Use*. Routledge.
- Connell, R. W. (1992). Social justice in education. In *Schools and Social Justice* (pp. 11–19). Toronto: Our Schools/Our Selves Education Foundation.
- Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review*, 64, 5–30.
- Davis, J., & Martin, D. B. (2008). Racism, assessment, and instructional practices: Implications for mathematics teachers of African American students. *Journal of Urban Mathematics Education*, 1(1), 10–34.
- Democrats for Education Reform. (2020, July 20). DFER Urges DNC Platform Committee on Revisions. Retrieved December 10, 2020, from <https://dfer.org/press/dfer-urges-dnc-platform-committee-on-revisions/>
- DePascale, C. (2015, December 8). ESEA – It’s so much more than a test. Retrieved December 12, 2020, from <https://charliedepascale.blog/2015/12/08/esea-its-so-much-more-than-a-test/>
- Ewing, E. L. (2018). *Ghosts in the schoolyard: Racism and school closings on Chicago’s South Side*. University of Chicago Press.
- Ferrara, S., Lai, E., Reilly, A., & Nichols, P. D. (2016). Principled approaches to assessment design, development, and implementation. *The Handbook of Cognition and Assessment, Frameworks, Methodologies and Applications*, 41–74.
- Fetterman, D. M. (2001). *Foundations of empowerment evaluation*. Thousand Oaks, CA: Sage.
- Fetterman, D., & Wandersman, A. (2007). Empowerment evaluation: Yesterday, today, and tomorrow. *American Journal of Evaluation*, 28(2), 179–198.
- Gergen, K. J., & Dixon-Román, E. J. (2014). Social epistemology and the pragmatics of assessment. *Teachers College Record*, 116(11), 1–22.
- Gipps, C. (1999). Chapter 10: Socio-cultural aspects of assessment. *Review of research in education*, 24(1), 355–392.
- Glasner, D. P. (2018). *The Impact of Tracking Students in Mathematics on Middle School Student Achievement Outcomes* (Doctoral dissertation, Cleveland State University).

- Graham, E. J. (2020). "In Real Life, You Have to Speak Up": Civic Implications of No-Excuses Classroom Management Practices. *American Educational Research Journal*, 57(2), 653–693.
- Heubert, J. P., & Hauser, R. M. (1999). *High stakes: Testing for tracking, promotion, and graduation*. National Academy Press, 2101 Constitution Avenue, NW, Washington, DC 20418.
- Jennings, J., & Sohn, H. (2014). Measure for measure: How proficiency-based accountability systems affect inequality in academic achievement. *Sociology of Education*, 87(2), 125–141.
- Kavanagh, K. M., & Fisher-Ari, T. R. (2020). Curricular and pedagogical oppression: Contradictions within the juggernaut accountability trap. *Educational Policy*, 34(2), 283–311.
- Knoester, M., & Au, W. (2017). Standardized testing and school segregation: like tinder for fire? *Race Ethnicity and Education*, 20(1), 1–14.
- Lee, C. D. (1998). Culturally responsive pedagogy and performance-based assessment. *Journal of Negro Education*, 268–279.
- Lucas, S. R. (1999). *Tracking Inequality: Stratification and Mobility in American High Schools*. *Sociology of Education Series*. Teachers College Press, 1234 Amsterdam Avenue, New York, NY 10027.
- Mendoza-Denton, R. (2014). A social psychological perspective on the achievement gap in standardized test performance between White and minority students: Implications for assessment. *The Journal of Negro Education*, 83(4), 465–484.
- Mendoza-Denton, R., Kahn, K., & Chan, W. Y. (2008). Can fixed views of ability boost performance in the context of favorable stereotypes? *Journal of Experimental Social Psychology*, 44, 1187–1193.
- Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher*, 18(2), 5–11.
- National Council of Teachers of Mathematics. (2020, May 26). *NCTM releases new reports that aim to help students with mathematics beginning as early as preschool* [Press release]. Retrieved from <https://blog.apastyle.org/apastyle/2010/09/how-to-cite-a-press-release-in-apa-style.html>
- Nebraska Department of Education. (2020). *Stackable, Instructionally-embedded, Portable Science (SIPS) assessments: A proposal submitted in response to the Request for Proposals under the Competitive Grants for State Assessments Program, CFDA 84.368A*. Retrieved from: https://oese.ed.gov/files/2020/10/Nebraska-Department-of-Education_Redacted.pdf
- Nichols, P. D., Ferrara, S., & Lai, E. (2016). Principled design for efficacy: Design and development for the next generation of assessments. In R. Lissitz & H. Jiao (Eds.), *The next generation of testing: Common core standards, smarter balanced, PARCC, and the nationwide testing movement* (pp. 49–81). Baltimore: Information Age Publishing.
- Oakes, J. (1985). *Keeping track: How schools structure inequality*. New Haven, CT: Yale University Press.
- Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers. Chestnut Hill: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.
- Smith-Maddox, R. (1998). Defining culture as a dimension of academic achievement: Implications for culturally responsive curriculum, instruction, and assessment. *Journal of Negro Education*, 302–317
- Wandersman, A., Snell-Johns, J., Lentz, B., Fetterman, D., Keener, D. C., Livet, M., et al. (2005). The principles of empowerment evaluation. In D. M. Fetterman & A. Wandersman (Eds.), *Empowerment evaluation principles in practice* (pp. 27–41). New York: Guilford.

