Check for updates

Research Report

Culturally Responsive Assessment: Provisional Principles

ETS RR–23-11

Michael E. Walker
Margarita Olivera-Aguilar
Blair Lehman
Cara Laitusis
Danielle Guzman-Orth
Melissa Gholson

December 2023

ETS Research Report Series

EIGNOR EXECUTIVE EDITOR

Daniel F. McCaffrey
Lord Chair in Measurement and Statistics

ASSOCIATE EDITORS

Usama Ali
Senior Measurement Scientist

Beata Beigman Klebanov
Principal Research Scientist

Heather Buzick
Senior Research Scientist

Tim Davey
Director Research

Larry Davis
Director Research

Jamie Mikeska
Senior Research Scientist

Jonathan Schmidgall
Senior Research Scientist

Jesse Sparks
Senior Research Scientist

Michael Walker
Distinguished Presidential Appointee

Klaus Zechner
Senior Research Scientist

PRODUCTION EDITORS

Kim Fryer
Manager, Editing Services

Ayleen Gontz
Senior Editor

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

23308516, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12374, Wiley Online Library on [29/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ETS Research Report Series ISSN 2330-8516

23308516, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12374, Wiley Online Library on [29/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

RESEARCH REPORT

Culturally Responsive Assessment: Provisional Principles

Michael E. Walker, Margarita Olivera-Aguilar, Blair Lehman, Cara Laitusis, Danielle Guzman-Orth, & Melissa Gholson

ETS, Princeton, NJ

Recent criticisms of large-scale summative assessments have claimed that the assessments are biased against historically excluded groups because of the assessments' lack of cultural representation. Accompanying these criticisms is a call for more culturally responsive assessments—assessments that take into account the background characteristics of the students; their beliefs, values, and ethics; their lived experiences; and everything that affects how they learn and behave and communicate. In this paper, we present provisional principles, based on a review of research, that we deem necessary for fostering cultural responsiveness in assessment. We believe the application of these principles can address the criticisms of current assessments.

Keywords culturally responsive assessment; large-scale assessment; summative assessment; test bias; test fairness

doi:10.1002/ets2.12374

This paper concerns itself primarily with large-scale summative assessments in the United States, including accountability exams, such as the Smarter Balanced summative assessment (Smarter Balanced Assessment Consortium [Smarter Balanced], 2022); college entrance exams, such as SAT® (College Board, n.d.) and ACT (n.d.); and licensure exams, such as Praxis® (ETS, n.d.). The principles discussed in this paper could easily be applied to interim assessments, such as MAP (NWEA, n.d.) or the ETS Testlets (Edulastic, n.d.), as well as to large-scale group score assessments, such as NAEP (National Center for Educational Statistics, n.d.). We choose to focus on individual-level summative assessments for two reasons. First, these assessments are often associated with high-stakes decisions such as college admissions or hiring. Second, because of their high-stakes nature, these assessments may be difficult to adapt in a way that will be mutually acceptable to the majority of stakeholders. We believe summative assessments must change to remedy the criticisms of high-stakes standardized tests causing negative consequences for historically excluded students.[1] These criticisms focus primarily on three aspects of the tests: (a) their lack of cultural representation, (b) their role in perpetuating racial stereotypes, and (c) their use as gatekeepers limiting educational advancement for historically excluded students.

Each of these perceived shortcomings may result from a good faith effort to produce fair assessments. Consider the first criticism, lack of cultural representation. To facilitate comparison of students of different backgrounds, the gold standard in measurement has been to "keep everything the same for everyone" (Sireci, 2020). Cultural references are removed with the goal of minimizing construct-irrelevant variance. Randall (2021) made the case that this attempt to eliminate cultural references removes the context that is relevant to non-White students. She argued that no context can truly be neutral; thus, what remains in the assessments is a White-centered context that benefits primarily White students.[2] Several papers criticizing large-scale assessments have claimed that traditional assessment practices center White culture (e.g., Lyons et al., 2021; Sireci, 2020). Some researchers argue that White-centeredness seems almost inevitable given that assessment designers are disproportionally White (see Packman et al., 2010; Randall, Rios, & Jung, 2021).

A second criticism of assessments is that the reporting of achievement gaps perpetuates negative stereotypes about the academic abilities of Black, Hispanic, and Native American students. This issue has been widely acknowledged and discussed (Cross, 2007; Harper, 2015; Ladson-Billings, 2006; Montenegro & Jankowski, 2017), and this claim is supported by a recent small-scale experimental study (Quinn, 2020). By focusing on achievement gaps (e.g., by race, gender, disability) and employing these categories as main explanatory factors, deficit-based narratives are perpetuated that do not acknowledge the role of structural inequities (Delpit, 2012; Noguera, 2003; Quinn, 2020).

Corresponding author: Michael Walker, E-mail: mwalker@ets.org

23308516, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12374, Wiley Online Library on [29/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The messaging behind achievement gaps may not only affect others' stereotypes of Black, Hispanic, and Native American students, but it may also negatively contribute to students' perceptions of themselves (Delpit, 2012), as they learn that their language and identities are not valued and that they need to conform to the predominant cultural way of talking and thinking (Randall, Poe, & Slomp, 2021). Labels applied to students based on academic achievement reflect a struggling educational system (Baldridge, 2014; Gorski, 2011; Milner, 2011). These labels also help sustain deficit narratives that may already be applied to certain groups. Subsequently, many incoming and veteran teachers enter classrooms with negative biases and racist ideologies based on these same narratives that they are unconsciously reinforcing through their interactions with students (Simson, 2014). Note that we do not argue against reporting disparities that could inform policy and teaching practice. Rather, we argue that these differences are better framed as "inequality in educational outcomes" rather than "achievement gaps." Quinn and Desruisseaux (2022) found that choice of these seemingly equivalent phrases did not affect teachers' explanations for why inequalities exist. However, use of the phrase "achievement gaps" primed deficit mindsets in teachers, leading them to deprioritize resolving the issue. By contrast, focusing on inequality tends to reduce preconceptions that feed biased views and treatment.

A third criticism is that standardized tests have unintended negative consequences that adversely affect students who have been historically excluded from higher education, including students of color, students living in economically disadvantaged communities, English language learners, and students with disabilities. In high school, low test scores may lead to exclusion from high-quality education such as Advanced Placement® courses.[3] Many critics claim that, because of their use as college admission criteria, standardized tests create barriers to academic advancement for large segments of the population (e.g., Couch II et al., 2021).

Negative effects are also seen in accountability exams. Test-based accountability was intended to close racial and socioeconomic achievement gaps by focusing on groups rather than individual students. Policymakers believed that requiring standards-based state testing for basic skills of reading, mathematics, and science would lead to educational improvement (No Child Left Behind [NCLB], 2002). NCLB did yield some benefits, most notably ensuring that some groups of lower performing students were not systematically excluded from accountability systems, increasing the accessibility of state assessments for students with disabilities and English learners, and directing resources to schools the most in need. Still, critics of educational accountability policies have cited a variety of unintended negative consequences within the educational system that exacerbate inequities in education (Emler et al., 2019; Lane et al., 1998). Accountability assessments have led to a narrowing of the curriculum with an emphasis on skills (Jennings & Bearak, 2014; Lane, 2020), increased pressure on educators (Ro, 2019), a lack of educator autonomy (Kavanagh & Fisher-Ari, 2020), an overemphasis on test preparation (Sonnert et al., 2019), the inappropriate use of test scores (Tavassolie & Winsler, 2019), questionable practices in reassignment of teachers or principals (Lane, 2020; Martin, 2012), tracking of students (Giersch, 2018; Lane & Stone, 2002), penalizing teachers and schools (Arcia et al., 2011; Nichols & Harris, 2016), and encouraging cheating (Aronson et al., 2016). These results potentially have differential impact on historically excluded students.

Randall (2021) pointed to the intentional neutrality of test items as a contributor toward some of the more recent assessment challenges. Students are not inherently neutral; they represent multiple intersecting identities and preferences. The measurement field's traditional approach of creating a neutral test that does not mirror the diversity of the target students fails to incorporate the diversity of who the learners are, what they have learned, and how they have learned. To combat this barrier to authentic, representative assessment content, Randall (2021) proposed an enhanced heuristic consisting of collective steps toward a justice-oriented, antiracist[4] approach to construct definition and representation, consisting of markers of organization and individual positionality, identification of who is being assessed and their identities, power dynamics (e.g., who holds the power, or who is being privileged or harmed by nature of the selected construct definition), processes for evaluating the construct-definition procedures, and careful evaluation of consequences emerging from the construct definition (refer to Table 1 in Randall, 2021). The specific combination of enhanced heuristics, other approaches, or combinations of approaches needed to remove the barriers raised in assessment remains to be determined, but what is clear is that there is a critically overdue need to revise traditional practices that have promoted narrowly defined constructs that function exclusively as windows to allow students to see the world from the perspective of mainstream America, when constructs should also serve as mirrors for diverse learners to see themselves represented in the purpose and context of the assessment (Style, 1988/1996). Learning about their own as well as others' cultures helps students to engage more meaningfully and effectively with people of diverse backgrounds.

23308516, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12374, Wiley Online Library on [29/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 1 Principled Questions to Ask When Designing and Using Assessments

Stage of testing	Principle	Questions to ask
Population	Shared power; high expectations	Who is eligible to take this test? Who is excluded?
Purpose	Shared power; flexibility; high expectations	Why is this test needed? What knowledge will this test describe?
	Shared power; high expectations	Who is included in the early conversation about the need for a test?
	High expectations; asset-based	How will the test provide opportunity? Who may be harmed?
	Engagement	How desirable is this test to all involved parties?
Design	Shared power	What structures allow for co-design by all involved parties?
	Shared power	How are group differences in the definition of knowledge incorporated?
	Engagement; flexibility	How will the test incorporate test taker background?
	Engagement; flexibility; asset-based	How will the test work to test taker strengths?
	Shared power	How to ensure that the test allows everyone to show what they know?
	Shared power	What mechanisms allow for feedback from all parties at every stage of design?
Administration	Engagement; high expectations	How do administration conditions help every test taker feel safe and welcome?
	Shared power; flexibility; asset-based	How are individual differences among test takers accommodated?
Scoring	Flexibility	What feedback is needed from this test?
	Flexibility; asset-based	How do scoring models account for differences in mode of expression?
	Shared power	How are views of involved parties incorporated into scoring models?
Use	Shared power	What input to test takers have in how their test information is used?
	Shared power	How can test users give input on consequences of information use?
	Shared power	How are test taker rights protected?
	High expectations; asset-based	How are possible adverse consequences of score use avoided?
	High expectations; asset-based	How does score use promote opportunity for all test takers?

What Is Cultural Responsiveness?

In response to the issues listed above, we suggest several principles that may facilitate the development of more culturally responsive assessments. We want to clarify from the outset that we view culture as more than identification with race or gender. Instead, as Montenegro and Jankowski (2017) indicated, culture involves

> (1) the explicit elements that make people identifiable to a specific group(s) including behaviors, practices, customs, roles, attitudes, appearance, expressions of identity, language, housing region, heritage, race/ethnicity, rituals, religion; (2) the implicit elements that combine a group of people which include their beliefs, values, ethics, gender identity, sexual orientation, common experiences (e.g. military veterans and foster children), social identity; and (3) cognitive elements or the ways that the lived experiences of a group of people affect their acquisition of knowledge, behavior, cognition, communication, expression of knowledge, perceptions of self and others, work ethic, collaboration, and so on. (pp. 8–9)

The idea of culturally responsive education and assessment is not new. It has appeared in many arenas under many names over the years: “Culturally relevant pedagogy” (Ladson-Billings, 1995), “culturally responsive teaching” (Gay, 2000), “culturally responsive assessment” (Hood, 1998; Landl, 2021; Slee, 2010), and “culturally sustaining pedagogy (Baker-Bell et al., 2017). Some researchers have used the term “socioculturally responsive” (e.g., Lee & Quijada Cerecer, 2010) as a reminder that learning cannot be separated from the social context in which it occurs (Gutiérrez, 2012). Researchers of second language acquisition often have used the term “socioculturally responsive” in the tradition of Vygotsky’s (1978) sociocultural theory of mind (Yuksel & Inan, 2013); but not always (e.g., Karem & Washington, 2021, speak of a “culturally responsive” approach to testing dual language learners). Other emerging terms include “socially just and culturally responsive” (Peters et al., 2020) and “culturally and socially responsible” (Taylor & Nolen, 2022). We prefer to use the term “culturally responsive,” which has a long history and appears to be most prevalent in the mainstream educational space as of this writing. We emphasize, as do other researchers using the term, that the issues surrounding differential impact of assessments are broader than culture: They include issues of equity, access, and social justice.

Landl (2021) defined a culturally responsive assessment as one that evaluates students' knowledge, skills, and understandings in a way that takes into account their unique cultural identities. A culturally responsive assessment incorporates flexibility and choice so that students can leverage their own cultural perspectives to demonstrate their mastery of a given subject area. Similarly, Randall, Poe, and Slomp (2021) stated:

> Culturally sustaining approaches to assessment (a) draw on BIPOC [Black, Indigenous, and People of Color] students' funds of knowledge (Vélez-Ibáñez & Greenberg, 1992), (b) are connected to the lives of these students, (c) allow them to demonstrate their competence in a variety of ways through community cultural wealth (Yosso, 2005), and (d) are embedded within a culturally sustaining curriculum. (p. 596)

Why Does It Matter?

Our ultimate goal is to create culturally responsive assessments that emphasize new or enhanced processes or procedures and to incorporate these enhanced processes in principled assessment design approaches (e.g., evidence-centered design; Mislevy et al., 2003) across all stages of test design and development. In this paper, we propose five design principles of culturally responsive assessment, which we believe can address the criticisms of current assessments:

1. Culturally responsive assessments require a process that shares power across all concerned parties at all stages of the assessment process. Giving a voice to all stakeholders will help ensure that the assessments are culturally inclusive and will help prevent test misuses.
2. Assessments should be designed to foster academic engagement and belonging in academic environments. This principle reflects a desire to design assessments that reflect the test taker's identity and lived experience.
3. Culturally responsive assessments should reflect the expectation that all students have the potential to perform at high levels. Such high expectations for all students will help to negate the influence of negative biases that often hinder student performance.
4. Assessments should be designed to maximize flexibility to account for individual differences in culture, interests, and identities of all learners. By accommodating the test takers' diverse backgrounds, culturally responsive assessments will allow test takers to use their particular talents to maximum advantage.
5. Culturally responsive assessments are designed to reflect asset-based perspectives that measure what students know and can do and disrupt traditional deficit-narratives.

We see these five principles as foundational to support a culturally responsive assessment triangle model. This assessment triangle has been used as both a framework for designing assessments and as a tool for establishing validity (Marion & Pellegrino, 2006; National Research Council, 2001). The five principles represent pillars of the assessment design structure, supporting the assessment triangle (refer to Figure 1). This triangle consists of the three vertices of cognition, observation, and interpretation. Cognition refers to models of how students learn and represent knowledge. Observation refers to the kinds of tasks that will elicit student behaviors that demonstrate key knowledge. Interpretation refers to how we make sense of the information collected from observations. The five design principles of culturally responsive assessment provide a framework upon which to rest cognition, observation, and interpretation. The principles support transparent processes to ensure that every test taker is maximally represented in each phase of design and development. The principles are dynamic, and the interactions have been called out in more detail in the sections that follow.

Design Principles of Culturally Responsive Assessment

Shared Power Principle

> *Culturally responsive assessments require a process that shares power across all concerned parties, including them in all stages of the assessment process. This shared power includes monitoring the consequences of assessments and their uses.*

Human development is a cultural phenomenon. What, in what order, and by what means a child learns, as well as how that learning manifests itself, very much depend upon the culture in which the child develops. These cultural differences can be quite large. Rogoff (2003) noted, for example, that in many U.S. families, children are not considered capable

23308516, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12374, Wiley Online Library on [29/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

23308516, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12374, Wiley Online Library on [29/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

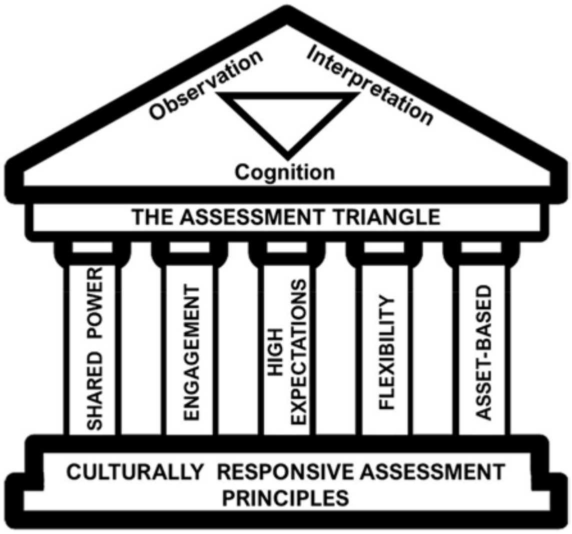


Figure 1 Five principles of cultural responsiveness and their relationship to assessment design.

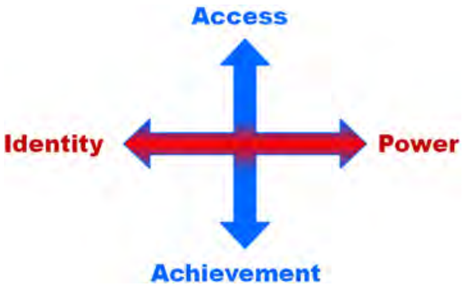


Figure 2 Four dimensions of equity.

of caring for themselves or others until about age 10; whereas in places in Oceania, 3-year-old children are gardening, tending to household duties, and taking care of younger siblings. It stands to reason, then, that children viewed through a cultural lens other than their own could appear deficient by the standards of the observer (Randall, Poe, & Slomp, 2021). In a similar way, standardized tests that tend to focus on one culture to the exclusion of others potentially put students outside that target culture at a disadvantage (Sireci, 2020). The disconnect between the mainstream way of teaching and testing and other cultures' ways of knowing can lead students from these other cultures to approach tests with fear and anger rather than with creativity (Trumbull & Nelson-Barber, 2019). The result is an inequitable situation in which those students are left vulnerable and powerless (e.g., see Hartocollis, 2019). Such uses of tests as instruments of power and control violate the fundamental democratic values upon which our society is based (Shohamy, 2001).

A desired outcome of culturally responsive assessments is that they will create greater equity among students from different backgrounds. According to Gutiérrez (2012), equity has four dimensions: access, achievement, identity, and power. Access refers to resources that provide the student with the opportunity to learn, including quality teachers, facilities, curriculum, and supports. Achievement includes measurable outcomes such as test scores. Identity includes the students' opportunity to see themselves reflected in the curriculum and assessments. Power concerns who has input on the curriculum and who defines what knowledge is. Gutiérrez saw access and achievement as composing the dominant axis, which maintains the status quo. Identity and power define the critical axis, acknowledging the legitimacy of the students' social and cultural perspectives (see Figure 2). Greater control lies along this critical axis and ultimately requires empowering the student with major decisions about multiple aspects of the test and its uses. In other words, power is the key to equity for historically excluded individuals. This shared power involves acknowledging the legitimacy of the student's viewpoint.

In cultural research, it is important to understand other people's (and even one's own) ways of doing things before placing value judgments on them. Even if value judgments become necessary at some point, those judgments will be

23308516, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12374, Wiley Online Library on [29/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

better informed if behavioral patterns are first observed without prejudice (Rogoff, 2003). This same attitude of separating understanding from judgment can help testing professionals in the development of culturally responsive assessments. Understanding results from keen observation and sincere dialog without criticism. Such dialog builds trust, which is essential to forming alliances among various stakeholders who feel that they are equals in the assessment development process. Assessment practices based on such democratic principles delineate a partnership between the tester and the tested. The design of tests, and even the definition of knowledge, become shared responsibilities among multiple stakeholders, including not only students and testing professionals, but also parents, teachers, and community members (Shohamy, 2001).

Such partnerships are the key to achieving cultural validity in assessments. First, stakeholders jointly attempt to understand students' social, cultural, and educational backgrounds and ways of communicating (Solano-Flores, 2011). Then the stakeholders can identify and eliminate any aspects of an assessment that may prevent students from a given cultural group from demonstrating what they know (Solano-Flores & Nelson-Barber, 2001). Cultural validity becomes the foundation for the validity argument and is integral to the development of the assessment and the interpretation of results (Kūkea Shultz & Englert, 2021).

When designing a science assessment for their Hawaiian Language Immersion Program, the Hawaii State Department of Education (DOE) employed such practices. The Hawaii DOE assembled a team of parents, teachers, community members, administrators, and testing professionals to design an assessment centered on Hawaiian language and culture (Kūkea Shultz et al., 2019). The group began by jointly formulating a set of science standards. They then built the assessment based on those standards. The result was an assessment that the group believed better reflected Hawaiian immersion students' knowledge, skills, and understanding than did the previous assessment based solely on mainstream culture.

A key component to ensuring that the science assessment allowed students to appropriately demonstrate what they knew was the cognitive interview. Through structured interviews, test developers learned how students interpreted test items as they worked through them. These interviews helped in understanding student interpretations of the material as well as how well the items corresponded to the intended construct (Kūkea Shultz & Englert, 2021). The success of the exercise did not lie so much in the cognitive interview, which is standard test development practice. Rather, the most important step lay in choosing representative student samples from the target population.

In the context of human-computer interaction, McCarthy and Wright (2015) have advocated for what they call experience-centered design. In this process, users as well as researchers and designers explore something together with the goal of changing it for the better. Especially when participants enter the process from differential positions of power, it is important to build mutual trust to facilitate equal sharing of ideas. According to the authors, "Participation is not a one-off event; experience-centered participatory design has to become embedded in organizational values and practices, and specific participatory projects always have to be seen with respect to the particular context in which they operate" (pp. 80–81).

Araneda and Sireci (2021) extended this idea to test construction. They advocated creating a test in the same way that game designers create new games—by focusing on the user experience. The authors spoke of test users[5] rather than test takers. In the user experience framework, participants explore all aspects of the test experience, including cognitive, emotional, and physical. They do this through free-associating thoughts and behaviors[6] during, immediately after, and upon reflecting on the test-taking experience. They also do this by engaging in dialog with the test designer. As mentioned, these events cover multiple modes of experience: not just thoughts, but also feelings as well as physical reactions (e.g., feeling hungry or cold during the test). The authors believed that an experiential process could benefit not only test development but also test validation. As with cognitive interviews, the experience-centered design approach intimately involves concerned parties in the development process, so that the resulting assessment can best reflect their culture.

Choice (e.g., of what item to answer, on the mode of response) on an assessment might also be seen as empowering. Insofar as students view choice (see the Flexibility Principle section in this report) as a means to increase their performance, then allowing choice provides the perception of personal control over outcomes. This perceived control can lead to greater personal engagement (see the Engagement Principle section in this report), which in turn would positively affect performance (Skinner et al., 1990; Wise & Smith, 2016). Choice by itself, however, is insufficient. Whether or not choice is empowering depends upon who defines the choice. If the choice is defined unilaterally by some entity outside of the culture (e.g., an outside test developer), then the choice still represents external control. Likewise, some collectivist cultures may prefer that choices are made at the group rather than individual level (Ryan & Deci, 2000). By contrast, sharing

23308516, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12374, Wiley Online Library on [29/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

power requires that stakeholders be free to determine for themselves the proper course of action. Thus, even the kinds of choices that appear on an assessment should be made in partnership with stakeholders.

Engagement Principle

> *Culturally responsive assessments are designed to foster active participation, use of productive cognitive strategies (e.g., self-regulation), and feelings of belonging in the academic environment when completing the assessment*.

The engagement principle is relevant to all assessment types, as promoting engagement facilitates students to perform to the best of their abilities (Wise & Smith, 2016). Designing to maximize engagement (active participation) may be viewed as a best practice more generally for assessments, but this practice is not always implemented fully. For example, the practice of creating context-free assessment content will likely lessen engagement for all students because the content does not align with students' beliefs, interests, or values (Ryan & Deci, 2000). The demands-capacity model of test-taking effort (Wise & Smith, 2011) posits that engagement during tests represents an interaction between item characteristics (demands) and student characteristics (capacity). Students' effort represents their willingness or ability to provide an engaged response to an item. Thus, a general best practice should be to design assessment content to increase students' willingness to meaningfully engage with the content.

Designing engaging assessments should benefit everyone generally, but it may have a particularly positive impact on historically excluded students. Critics believe that so-called context-free assessments are not devoid of context; instead, they default to the dominant White culture in the United States and only removing context and references that are specific to identities or cultures that do not align with the dominant culture (Montenegro & Jankowski, 2017; Randall, 2021). It is then possible that the removal of context from assessment content is particularly disadvantaging historically excluded students by limiting their ability to connect with assessment content and meaningfully engage with it. Context-free assessments may also lead to the choice of context-free learning materials as a consequence of teaching to the test, which would only further disadvantage historically excluded students (Thompson & Allen, 2012). Developing context-rich assessment content will require the inclusion of historically excluded students and community members to ensure that their experiences are centered (Araneda & Sireci, 2021) and avoid the use of stereotypes or other inappropriate contexts (see the Shared Power Principle section in this report). The engagement principle would then promote the use of context-rich assessment content that represents diverse perspectives and lived experiences.

An example of context-rich assessment content can be seen in the recently developed culturally relevant ETS Testlets (Edulastic, n.d.), which are online-administered scenario-based assessments that are intended to support classroom instruction. Scenario-based assessments are inherently context-rich as they are made up of items and tasks situated within specific contexts, situations, topics, and interaction styles (e.g., dialogs with virtual agents). The design of scenario-based assessments purposefully aims to promote maximal engagement by situating items as necessary to achieving some larger, meaningful goal and feelings of belonging through simulated social contexts and interactions (O'Reilly & Sabatini, 2013; Sabatini et al., 2018, 2019). However, the presence of context-rich assessment content does not guarantee that the assessment will be culturally relevant. To increase their cultural relevance, the development process of the Testlets elicited scenarios, contexts, and topics from historically excluded students and community members (e.g., ETS business resource groups) through interviews and focus groups. These efforts produced a wide range of scenarios, contexts, and topics that are potentially relevant and engaging to students from a variety of backgrounds, with several topics, contexts, and scenarios reflected in later Testlets (e.g., voting rights, urban community gardens, labor movements). These efforts also identified the design of avatars representing nonbinary/gender nonconforming individuals and students with disabilities as another important design consideration to help promote engagement (O'Dwyer et al., 2023).

Self-determination theory is a general theory of motivation (Ryan & Deci, 2000) that can provide guidance for how to include context-rich assessment content but also guide the implementation of other design decisions to promote engagement for historically excluded students. Tasks that elicit intrinsic motivation have often been promoted as the ideal, but research on different types of extrinsic motivation has revealed that extrinsically motivating tasks can result in positive effects similar to intrinsic motivation when the task promotes autonomy for students and aligns with their identities, beliefs, and values. Aligning with students' identity, beliefs, and values goes beyond the assessment content to all aspects of the assessment. For example, inclusion of historically excluded students and community members in the assessment

23308516, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12374, Wiley Online Library on [29/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

development process (see the Shared Power Principle section in this report) and alternate response format options and opportunities for student agency (see the Flexibility Principle section in this report) can allow for assessments to be perceived as aligning with students' identities and for students to view the assessment as valuing their identities. Alternate response options that allow students to respond in the format in which they are most comfortable and confident can improve measurement of their knowledge and skills and promote the value of their culture (Baker-Bell, 2020; Randall, Poe, & Slomp, 2021). The engagement principle would then promote the use of alternate response option formats in assessments and more generally evaluating the degree to which assessment content promotes autonomy for students.

The engagement principle can also inform the way in which feedback is provided to students. Feedback should be designed to motivate all students to continue their learning journeys. However, feedback on some standardized assessments utilize deficit-based language (e.g., "unsatisfactory," "far below proficient"; O'Donnell & Sireci, 2021) that could cause students to feel less motivated due to lowered expectancies for success in the future and diminished feelings of autonomy over their learning (Ryan & Deci, 2019; Wigfield & Eccles, 2000). The engagement principle would then promote the use of asset-based language (see the Asset-Based Principle section in this report) and the inclusion of suggested next steps to guide addressing areas for improvement. These design considerations are likely to benefit all students but may particularly benefit historically excluded students. Systemic inequalities, negative stereotypes, and negative prior experiences can all have a negative impact on feelings of autonomy and competence (Lewis & Hunt, 2019; Rojas & Liou, 2017) and these are disproportionately experienced by historically excluded students. Thus, more positively worded feedback that promotes feelings of autonomy and competence (e.g., "novice," "beginning learner"; O'Donnell & Sireci, 2021) will facilitate historically excluded students to feel motivated to continue their learning journeys.

Theories of engagement and motivation noted in this section can be used to guide the design of culturally relevant and responsive assessments; however, it is important to note that the majority of these theories primarily focus on learning contexts and do not specifically address the needs of historically excluded students. Thus, there is a need for empirical research to better understand how these theories can be applied to support historically excluded students while completing assessments. It is also the case that engagement should be viewed as a key outcome when evaluating culturally relevant and responsive assessments (in conjunction with more typical psychometric evaluations). Engagement can be used as a metric in small-scale cognitive laboratory research (e.g., think-alouds, cognitive interviews) and large-scale field testing to better understand the effectiveness of designs aimed at cultural relevance and responsiveness. For example, engagement may serve as a moderator for context-rich assessment content impacting assessment performance, which could facilitate understanding how to leverage flexibility to give each student their best opportunity to show what they know and can do.

High-Expectations Principle

> *Culturally responsive assessments should reflect the expectation that all students have the potential to perform at high levels.*

Negative stereotypes of historically excluded students' academic abilities may result in low expectations for what they can accomplish, which may negatively affect students' self-perceptions and lower their academic achievement (Lewis & Hunt, 2019; Rojas & Liou, 2017). One way to combat negative perceptions about historically excluded students' academic ability is through maintaining high expectations of students' performance in a learning environment that acknowledges and values their cultural experiences (Delpit, 2012; see the Asset-Based Principle section in this report).

Although traditional reporting of standardized test results may have contributed to negative stereotypes of historically excluded students' academic abilities, assessments can signal the expectation that all students have the potential to perform at high levels if we commit to designing standardized assessments with cultural responsiveness in mind. We propose that culturally responsive assessments, with high expectations of all students' performance, provide opportunities for students to (a) *engage in higher order thinking* and (b) *reflect on and revise their thinking*. It should be noted that while there is vast research on the effect of high expectations on students' self-perceptions and achievement (Babad, 1993; Brophy, 1982; Lundberg et al., 2018; Rojas & Liou, 2017; Rubie-Davies et al., 2006; Vega et al., 2012), less attention has been given to how to signal high expectations in assessments. Hence, most of the literature reviewed in formulating this principle is not specific to assessments, but it is in the teaching and learning research literature (e.g., culturally responsive education). Next, we review the research supporting the two key aspects that we propose as part of the high-expectations principle.

23308516, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12374, Wiley Online Library on [29/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Higher Order Thinking: Interpretation and Transformation of Information

In culturally responsive education, high expectations about students' learning and performance are reflected through rigorous environments where students are expected to learn at high levels, receive the appropriate support to achieve high levels of learning, and demonstrate those high levels (Blackburn, 2017). Rigor is critically important here, as it cannot be disassociated from engagement—a core theme of culturally responsive education (and a principle of culturally responsive assessments; see the Engagement Principle section in this report). In a rigorous environment, students become engaged with the content being learned and take ownership of their own learning process. Conversely, student engagement is low without a rigorous learning environment (Washor & Mojkowski, 2007). For example, Delpit (2012) pointed out that African American students are often exposed to less rigorous content and are expected to complete simple tasks that do not facilitate engagement and learning.

Rigor tends to be confused with difficulty, but it is much more than that. Activities in a rigorous academic environment involve the interpretation and transformation of the content being taught, such that students are given opportunities to build their knowledge by making meaningful connections not only with the knowledge they already possess but also with their identity and cultural background (Stembridge, 2020). The goal of academic rigor is to prepare students to be lifelong learners who own their learning process (see the Shared Power Principle section in this report) and engage in authentic and complex academic tasks both in professional and real-life events (Matusevich et al., 2009). As such, rigor is frequently operationalized as the higher order thinking needed to apply knowledge and skills in novel contexts and applications (Edmunds et al., 2017; Matsumura et al., 2008; Mitchell et al., 2005).

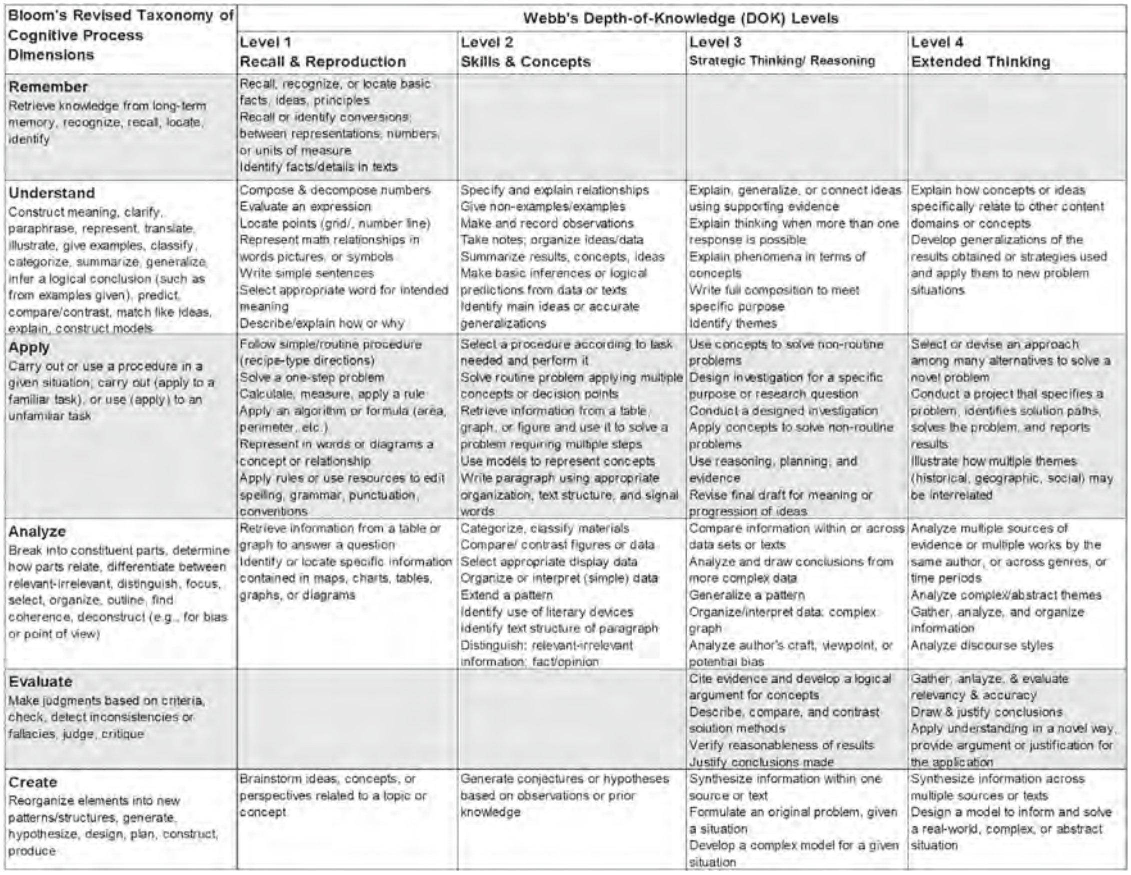
In the context of assessments, Evans (2021) pointed out that rigor is achieved by encouraging students to employ higher order thinking skills that lead students to construct meaning and interpretations from content when completing the assessment. Webb's depth of knowledge (DOK; Webb, 1997, 1999) is a classification of thinking complexity that can be used to identify the activities that encourage the transformation of skills and concepts emphasized by culturally responsive education. The four levels of DOK are

1. Recall
2. Skills and concepts
3. Strategic thinking
4. Extended thinking

Levels 3 and 4 lend themselves to culturally responsive assessments that target the higher order thinking needed to relate the content to students' backgrounds. Importantly, these higher levels of thinking are associated with higher engagement in the classroom (Paige et al., 2013; see the Engagement Principle section in this report). Despite the potential that items at Levels 3 and 4 offer in measuring the full extent of students' knowledge and abilities, their development and use pose several challenges, resulting in their underutilization (McClellan et al., 2016). While items at Levels 1 and 2 may have unambiguous correct answers, there could be multiple valid interpretations at Levels 3 and 4, given that they require students to make connections between multiple concepts, analyze information, and apply critical thinking skills. As a result, developing effective items at Levels 3 and 4 requires significant resources (i.e., time and effort). Further, the ambiguity in the correct responses to these items may also complicate their scoring and may potentially increase the opportunities to develop biased items which favor specific ways of thinking and interpretations.

Bloom's taxonomy of educational objectives stratifies cognitive skills by difficulty (remember, understand, apply, analyze, evaluate, create). Important work has been conducted to extend Bloom's taxonomy to emphasize the cultural relevance of the items and tasks developed, via the Bloom-Bank matrix, which aims to reflect the goals, objects, and perspectives of multicultural education. This matrix ranges from low multicultural levels in which the content rarely provides students with multicultural growth and substance, to content that encourages students to analyze content from different cultural perspectives, critique social and cultural issues, and propose a plan of action seeking social change (Ford & Scott, 2021; Scott, 2014).

This taxonomy, in combination with DOK, can be used as a guide to build rigorous assessments emphasizing activities that lead to meaningful transformation of the content. Hess et al. (2009) proposed a cognitive rigor matrix that maps the DOK needed to perform tasks requiring different cognitive skills (see Figure 3). Because Bloom's dimensions are ordered in terms of difficulty, the cognitive rigor matrix illustrates the relative independence of higher order thinking and task difficulty.



Bloom's Revised Taxonomy of Cognitive Process Dimensions	Webb's Depth-of-Knowledge (DOK) Levels			
	Level 1 **Recall & Reproduction**	**Level 2** **Skills & Concepts**	**Level 3** **Strategic Thinking/ Reasoning**	**Level 4** **Extended Thinking**
Remember Retrieve knowledge from long-term memory, recognize, recall, locate, identify	Recall, recognize, or locate basic facts, ideas, principles Recall or identify conversions: between representations, numbers, or units of measure Identify facts/details in texts			
Understand Construct meaning, clarify, paraphrase, represent, translate, illustrate, give examples, classify, categorize, summarize, generalize, infer a logical conclusion (such as from examples given), predict, compare/contrast, match like ideas, explain, construct models	Compose & decompose numbers Evaluate an expression Locate points (grid/, number line) Represent math relationships in words pictures, or symbols Write simple sentences Select appropriate word for intended meaning Describe/explain how or why	Specify and explain relationships Give non-examples/examples Make and record observations Take notes; organize ideas/data Summarize results, concepts, ideas Make basic inferences or logical predictions from data or texts Identify main ideas or accurate generalizations	Explain, generalize, or connect ideas using supporting evidence Explain thinking when more than one response is possible Explain phenomena in terms of concepts Write full composition to meet specific purpose Identify themes	Explain how concepts or ideas specifically relate to other content domains or concepts Develop generalizations of the results obtained or strategies used and apply them to new problem situations
Apply Carry out or use a procedure in a given situation; carry out (apply to a familiar task), or use (apply) to an unfamiliar task	Follow simple/routine procedure (recipe-type directions) Solve a one-step problem Calculate, measure, apply a rule Apply an algorithm or formula (area, perimeter, etc.) Represent in words or diagrams a concept or relationship Apply rules or use resources to edit spelling, grammar, punctuation, conventions	Select a procedure according to task needed and perform it Solve routine problem applying multiple concepts or decision points Retrieve information from a table, graph, or figure and use it to solve a problem requiring multiple steps Use models to represent concepts Write paragraph using appropriate organization, text structure, and signal words	Use concepts to solve non-routine problems Design investigation for a specific purpose or research question Conduct a designed investigation Apply concepts to solve non-routine problems Use reasoning, planning, and evidence Revise final draft for meaning or progression of ideas	Select or devise an approach among many alternatives to solve a novel problem Conduct a project that specifies a problem, identifies solution paths, solves the problem, and reports results Illustrate how multiple themes (historical, geographic, social) may be interrelated
Analyze Break into constituent parts, determine how parts relate, differentiate between relevant-irrelevant, distinguish, focus, select, organize, outline, find coherence, deconstruct (e.g., for bias or point of view)	Retrieve information from a table or graph to answer a question Identify or locate specific information contained in maps, charts, tables, graphs, or diagrams	Categorize, classify materials Compare/ contrast figures or data Select appropriate display data Organize or interpret (simple) data Extend a pattern Identify use of literary devices Identify text structure of paragraph Distinguish: relevant-irrelevant information; fact/opinion	Compare information within or across data sets or texts Analyze and draw conclusions from more complex data Generalize a pattern Organize/interpret data: complex graph Analyze author's craft, viewpoint, or potential bias	Analyze multiple sources of evidence or multiple works by the same author, or across genres, or time periods Analyze complex/abstract themes Gather, analyze, and organize information Analyze discourse styles
Evaluate Make judgments based on criteria, check, detect inconsistencies or fallacies, judge, critique			Cite evidence and develop a logical argument for concepts Describe, compare, and contrast solution methods Verify reasonableness of results Justify conclusions made	Gather, anlayze, & evaluate relevancy & accuracy Draw & justify conclusions Apply understanding in a novel way, provide argument or justification for the application
Create Reorganize elements into new patterns/structures, generate, hypothesize, design, plan, construct, produce	Brainstorm ideas, concepts, or perspectives related to a topic or concept	Generate conjectures or hypotheses based on observations or prior knowledge	Synthesize information within one source or text Formulate an original problem, given a situation Develop a complex model for a given situation	Synthesize information across multiple sources or texts Design a model to inform and solve a real-world, complex, or abstract situation

Figure 3 Hess' cognitive rigor matrix with curricular examples: applying Webb's depth-of-knowledge levels to Bloom's cognitive process dimensions. From Hess et al. (2009). Copyright © Karin K. Hess: Hess' Cognitive Rigor Matrix.

Another implication of the cognitive rigor matrix by Hess et al. (2009) is that, though challenging, higher order thinking (Levels 3 and 4) could be assessed with multiple-choice (MC) items, which tend to map into the lower dimensions of Bloom's cognitive processes. This contrasts with the perspective that MC items are usually not associated with higher order thinking (Matusevich et al., 2009; Scully, 2017) and that open-ended questions are more appropriate to elicit the transformation and transfer of learning (Baker et al., 1993; Edmunds et al., 2017; Matsumura et al., 2008). Despite the criticisms of MC items, a body of research has found consistent student performance on MC and constructed-response (CR) items (Hickson et al., 2012; Thissen et al., 1994; Walstad & Becker Jr., 1994), suggesting that in some situations a set of MC items may cover the same processes tapped by open-ended questions (Bennett et al., 1991; Tractenberg et al., 2013). It is likely that observed performance differences between MC and CR items may be associated with the overrepresentation of MC items tapping recall or recognition due to the difficulty of building practical MC items assessing higher order thinking (Scully, 2017).

We should clarify that we do not intend to advocate for the sole use of MC items, as there are other reasons to prefer varied item types and assessment formats For example, not only do scenario-based assessments and other performance tasks lend themselves to the measurement of higher order thinking skills, but they also facilitate students' engagement with the assessment by using content with real-world relevance, require students to apply their knowledge and skills in context, and provide an opportunity to build on students' strengths (Barlowe & Cook, 2016; Linn & Burton, 1994; Tung, 2017), hence increasing the external validity or authenticity of the assessments (see the Flexibility Principle section in this report). Similarly, CR items may be preferable for measuring higher order thinking and for engaging students in creating their own content based on their funds of knowledge and cultural identity (see the Asset Principle section in this report) and they could be built in a way that mimics the advantages of MC items (e.g., increased reliability, ease of scoring) by creating a series of short CR items each covering a small component scored using a more analytic rubric (Walker, 2007). One of the most significant disadvantages of CR items, that is, scoring, has been overcome by technological advances that have

23308516, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12374, Wiley Online Library on [29/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

made automated scoring possible. However, automated scoring has challenges that may result in inequitable results for specific student subpopulations. Bridgeman et al. (2012), for example, have shown that automated scoring may result in favoring some student subgroups over others, compared to human scoring. Research on artificial intelligence has flagged the underrepresentation of certain subgroups in the data used to train the models, resulting in algorithms that reproduce existing biases (Bender et al., 2021; Nordling, 2019). Similarly, in the context of standardized assessments, the algorithms for scoring CRs may only be as good as the data used to train the models. Training data that fails to incorporate not only the diversity in background characteristics of the students, but also in the diversity in their language and ways of demonstrating knowledge, may reproduce and exacerbate existing biases. Culturally responsive assessments address this by overrepresenting historically excluded groups in all phases of test design to increase the likelihood that differences in culture and background are built into task designs, scoring rubrics, and scoring algorithms.

Reflect on and Revise Thinking

Culturally responsive assessments should value practice and growth (Delpit, 2012; Ladson-Billings, 2021), not intelligence as a fixed trait (Dweck & Yeager, 2019). As such, they should provide opportunities for students to reflect on and revise their thinking (Boston & Wolf, 2006). The conceptualization of rigor as opportunities to show growth is demonstrated in the rigor rubric developed by the North Carolina State Board of Education (Matusevich et al., 2009). Regarding classroom assessments, higher levels of rigor imply an increasing number of opportunities for students to grow by receiving feedback in multiple assessment opportunities. There may be value in exploring these ideas in summative assessments as well.

Opportunities for students to show growth in their learning not only require multiple assessment occasions, but also effective feedback that they can act upon. Despite its potential advantages, several considerations are granted to increase its positive impact on learning and achievement (Hattie & Timperley, 2007). A vast body of research on feedback indicates that, in general, immediate feedback is more effective than delayed feedback (Anderson et al., 2001; Buzhardt & Semb, 2002; Corbett & Anderson, 2001; Mason & Bruning, 2001) and that elaboration feedback (i.e., explaining why an answer was wrong and providing cues about the correct response) is more beneficial to students' learning than verification feedback (i.e., only indicating whether an answer was correct or incorrect; Bangert-Drowns et al., 1991; Mason & Bruning, 2001; Shute, 2008). Providing immediate elaboration feedback may be particularly beneficial for students who approach a task with incorrect or incomplete information. Without adequate feedback, erroneous thoughts can be reinforced if they are not corrected in time. Timely feedback is also relevant to correct students' overconfidence misconceptions (e.g., overconfidence in understanding course material or their performance on an upcoming test), resulting in less-than-optimal outcomes (Finn & Tauber, 2015). Because a critical principle of culturally responsive assessments involves empowering students to make choices about their assessment experiences (see the Shared Power Principle section in this report), it is essential to provide feedback that corrects overconfidence misconceptions and helps students make better decisions.

Culturally responsive assessments could also provide opportunities for self-reflection by creating situations leading to productive struggle, that is, with challenging tasks where answers are not readily apparent (Basileo & Lyons, 2019; Edwards & Beattie, 2016; Hiebert & Grouws, 2007) and that are in students' zone of proximal development (Vygotsky, 1978). While the usefulness of creating productive struggle in assessments depends on the test's purpose, computer adaptive assessments and assessments involving student choice may be well suited to introducing such challenge as long as detailed feedback is provided that students can act upon. In a computer adaptive format, productive struggle may be elicited by presenting items at or slightly above students' ability levels. Similarly, giving students the choice to select a harder or easier item than the one previously answered can create productive struggle by encouraging them to try more complex items associated with stated standards of performance. These choices could foster self-regulation, ownership of the learning and assessment process, and self-efficacy. Through the repeated administration of assessments that incorporate opportunities of productive struggles, it could be possible to track students' changes in responses and growth in their learning.

Flexibility Principle

> *Culturally responsive assessments are designed to allow for flexibility to account for individual differences in culture, interests, and identities for all learners. Flexibility should consider the interests, preferences, and needs of all*

23308516, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12374, Wiley Online Library on [29/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

members of society while prioritizing those who have been historically excluded from assessments, education, and opportunity.

Standardized tests are frequently criticized as being inflexible to the interests, preferences, and needs of Black, Hispanic, and Native American students; multilingual learners; and students with disabilities. In the interest of standardization, traditional large-scale summative assessments have used the same form for each student, with narrow exceptions in the form of computer adaptive tests, accommodations for test takers with disabilities and in some cases multilingual learners. For example, computerized adaptive tests only alter the difficulty of the presented items to match the student's performance on previously presented items; such tests do not generally vary the content of the items or the response format to match the characteristics of the students (see Meijer & Nering, 1999, for an overview of computerized adaptive tests). Accommodations for students with disabilities may involve extended time, magnification, read-aloud protocols, or other format changes to a test whose content has not changed (Cormier et al., 2010). Multilingual learners benefit when English dictionaries and glossaries accompany the otherwise unaltered test (Abedi et al., 2001; Francis et al., 2006). In these examples, when flexibility has been provided on tests, alternatives have been primarily used to retrofit existing assessments with the intent of reducing construct-irrelevant variance.[7]

Assessments that maximize flexibility (in both context and content) from the outset are more likely than traditional standardized tests to meet students' basic psychological needs and to foster academic engagement and a sense of belonging (see the Engagement Principle section in this report) while also allowing students' cultural differences to be viewed as assets rather than deficits (see the Asset-Based Principle section in this report). Self-determination theory provides the empirical basis for how social-contextual factors can increase or decrease engagement through the satisfaction of basic psychological needs of autonomy, competence, and belonging (Ryan & Deci, 2017). Culturally responsive assessments can be designed to meet (or partially meet) each of these basic psychological needs through the development of student models that are grounded in cultural interests, identities, and needs of all learners and that provide greater flexibility through choices (autonomy). When flexibility is focused on a student's own preferences, it has the potential to increase autonomy, competence, and belonging and lead to greater engagement in learning. A recent instantiation of this principle may be found in the model of caring assessments (Lehman et al., 2018; Zapata-Rivera et al., 2020), which proposes a system that adapts to student characteristics and behaviors before (format and design features), during (task selection and supports), and after (actionable feedback) the assessment.

The most recent version of the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA] et al., 2014), or *Standards*, suggests that changes to an assessment can vary across a continuum from minor to substantial. Minor changes include "changes to the presentation and/or format of the test, test administration, or response procedures that maintain the original construct and result in scores comparable to those on the original test" (pp. 58–59). Substantial changes include a variety of adaptations that change the construct being measured (e.g., an alternate assessment based on alternate academic achievement standards for students with significant cognitive disabilities). These more substantial changes reflect the goal of obtaining a "reasonable measure of a somewhat different but appropriate construct" (p. 59). This type of flexibility, designed to reduce construct-irrelevant variance, has two negative impacts on students. One involves the requirement that the student self-identify as having a need that may be perceived as different from the student's peers. Another negative impact is the practice of isolating and segregating students who require accommodations into separate testing rooms or locations (Wentz et al., 2011). These unintended outcomes are at odds with culturally responsive teaching and learning, which aim to foster a community of learners (Stembridge, 2020).

Instead, we propose a type of flexibility that expands on the principles of universal design for learning (UDL; CAST, 2018) but prioritizes the interests, needs, and preferences of all stakeholders. This will require the development of student models that will in turn inform the design of task models that provide choices. This design should be done in collaboration with members of cultural groups that have been historically excluded from research, assessment, education, and the opportunity to have a seat at the table (see the Shared Power Principle section in this report). Flexibility for students from diverse cultural groups has rarely been considered from the start of assessment design. Flexibility is a key element of the UDL guidelines that includes three overarching principles to allow for multiple means of presenting content (audio, video, text, images), multiple ways for students to respond (writing a report or doing an oral presentation), and multiple means of engagement. Flexibility in accessible presentation of test content has been widely implemented in K–12 assessments where students are offered an array of "universal tools" that allow for test content to be provided

23308516, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12374, Wiley Online Library on [29/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

23308516, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12374, Wiley Online Library on [29/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

in audio, tactile, and visual formats (e.g., Smarter Balanced, 2022). Likewise, there are several models of flexibility of response wherein students can respond via speech to text, typing, or even in an alternate language (e.g., American Sign Language or Spanish). These approaches are an inclusive message to students, teachers, parents, and policy makers that each way of knowing is equally valued.

Within the assessment community, flexibility has been viewed by some as impractical and potentially detrimental to the purpose of many standardized assessments. One challenge of fully implementing the UDL guidelines in assessment is a concern that allowing too much individual flexibility will prevent establishing equivalence of scores across people. Another concern has been the effect of choice on student performance. While several studies in the late 1990s explored choice on large scale assessments, the findings were mixed when looking at changes in test-taker performance (Bridgeman et al., 1997; Campbell & Donahue, 1997). However, these same studies found some evidence that test-taking experiences are enhanced by choice. For example, in the 1997 NAEP study by Campbell and Donahue, students who chose their reading passages perceived the assessment as easier than students in the nonchoice group. This perception may increase student's sense of competence, which is one of the important elements in self-determination theory related to autonomy. In fact, the research on choices outside of large-scale assessment contexts has demonstrated that choice in task leads to greater intrinsic motivation for most cultures around the world, although more research is needed on the impact of personal choice versus collective choice in collectivist cultures (Ryan & Deci, 2017). One possible study could explore the impact on performance, engagement, and sense of belonging when the choice in passage topics is made by the individual, classroom cohort, or test developer.

Despite the lack of strong evidence that choice improves test scores, the use of choice in testing has been implemented in UDL, where flexibility in response, presentation, and engagement have been core principles for instructional and assessment practices. Recently, several psychometricians and assessment experts have also voiced support for assessments that provide this level of flexibility. Sireci (2020) has called for psychometricians "stuck in the 20th century" to embrace more flexible practices to provide more valid assessments. Likewise, Bennett (2021) included greater flexibility in his framework for socioculturally responsive assessments. Bennett pointed to the work of Mislevy (2018), who coined the phrase "conditional sense of fairness" to refer to greater flexibility in assessment procedures in return for more comparable evidence across individuals. One flexible design approach highlighted by Mislevy is UDL. He provided guidance on how to implement UDL to achieve comparability. He included three conditions for comparability: (a) it must be possible for targeting capabilities to be measured through alternate work products; (b) students must understand how to demonstrate them in the form they are using; and (c) the evaluation procedures must be applied in equivalent, suitably adapted, ways to all forms. We describe several examples of how this shift in thinking should be considered to allow for greater flexibility in the administration (presentation and response) and content (stimuli, questions, and responses) of assessments that acknowledge diversity in cultural and individual differences in test takers.

The universal design framework, commonly used in test development, emphasizes maximizing accessibility for the widest range of individuals. Universal design began as a set of architectural guidelines for developing physical spaces that are maximally accessible for all users of the space (Story et al., 1998). Since then, universal design has been extended to both learning and assessment spaces. The fairness chapter of the *Standards* makes several references to the architectural concept of universal design as well as to the principles for universal design of assessment (UDA), developed by the National Center on Educational Outcomes (n.d.). While the UDL guidelines are included in laws and policies that impact curriculum and teacher preparation, they are not included in the *Standards*. Yet Mislevy (2018) pointed to the UDL guidelines as a framework for providing flexibility, not the architectural or UDA guidelines referenced in the *Standards*. As we consider the next version of the *Standards*, it will be important to define how the concept of universal design can be implemented within an assessment context and which guidelines are best situated to provide flexibility for individual differences that extend beyond accessibility to include differences in culture, ethnicity, language, and lived experiences.

Asset-Based Principle

Culturally responsive assessments are designed to measure what students know and can do by using asset-based perspectives that disrupt traditional deficit-narratives.

In culturally responsive assessments, the phrase "asset-based"[8] can be defined in a variety of ways; but valuing students—the whole student—and their linguistic, cultural, and social characteristics and seeing their diversity as strengths in the classroom environment rather than challenges to overcome are at the heart of asset-based practices (e.g., Ramasubramanian et al., 2021). Further, an asset-based approach recognizes and appreciates the diversity in the "learning and lived experiences of all students" (p. 23). Assets include personal characteristics, interests, and everyday experiences to support learning and demonstrate knowledge.

Asset-based thinking can also be extended to culture. Culturally relevant assets are bound within the test taker and are associated within their "culture, community and family" (Roe, 2019, p. 5). We consider culture to be a complex set of time-specific and contextualized systems of beliefs, attitudes, and rituals (Gutiérrez & Rogoff, 2003; Nasir & Hand, 2006). Additionally, religion, socioeconomic status, and location (i.e., regions within a country) contribute to one's culture (Cohen, 2009). Other forms of cultural perspectives include individualist versus collectivist tendencies (Burn & Thongprasert, 2005). Acculturation—the extent to which students of a nondominant culture choose to maintain their culture over time by integrating or assimilating into the dominant culture—is another facet of culture (Van de Vijver & Phalet, 2004). Additional complex feelings of marginalization or separation may also emerge and can influence how students view and identify with their own culture.

Students have a range of learning and lived experiences (Ramasubramanian et al., 2021); thus, they may identify at the intersection of multiple characteristics (e.g., gender, race, ethnicity, language, disability, religion; see Bešić, 2020; Collins, 2015; Crenshaw, 1989, 1991, 2017). Using an asset-based lens, we can view learning as a mechanism to help students discover their self-identity across both formal and informal learning environments (Esteban-Guitart & Moll, 2014; Hogg & Volman, 2020). Learners' self-identity in this context is not restricted to labels but is a set of resources or tools. "These tools have been historically accumulated and culturally developed; they are socially distributed and transmitted; and they are essential for constructing one's identity and for defining and presenting oneself" (Esteban-Guitart, 2012, p. 177). Students' identity and culture may have additional impact on how tests are designed and developed. These impacts may be seen through various means, such as the representation of different cultural identities (e.g., race, language, ethnicity, religions) or culturally situated behaviors (e.g., individualist, collectivist) on background questionnaires or the diversity of character or setting reflected in reading passages, graphics, task scenarios, or activities designed to promote culturally and justice-oriented assessments (Randall, 2021). Additionally, students' assets could potentially be incorporated into scoring models, where multiple means of evidence of what students know and can do are given value. Ultimately, these multifaceted, integrated concepts of culture and identity are foundational to valuing students' diversity as assets, and there are multiple ways to apply these considerations to make assessments asset-based and more culturally responsive.

Why We Need Asset-Based Approaches for Culturally Responsive Assessments

Using student assets in education is not a new concept, and it is an approach that may be especially suited to address the traditional criticisms around testing perpetuating deficit narratives about diverse learners. Examples such as culturally relevant pedagogy (Ladson-Billings, 1995) and culturally responsive frameworks (Gay, 2000) incorporate student assets and use asset-based framing when referring to who the students are and what they know and can do. In assessment, Mislevy (2018) highlighted similar perspectives, drawing specific attention to the unique linguistic, cultural, and substantive patterns that all people possess and can use to define how and why people do or expect certain things (e.g., broadly encompassing the learners' and test developers' extra and intrapersonal knowledge and activities around a particular domain, such as language type, which can be further distilled down to various subdomains and skills, cultural and regional variations, dialects, and registers).

The *Standards* (AERA et al., 2014) have identified several areas related to student characteristics, such as language and cultural diversity, that have potential to introduce fairness issues in assessments. In this view, the heterogeneity of learners' characteristics, or *assets*, may be influenced by the learners' personal characteristics. While the *Standards* seek to promote fairness for all learners across assessment contexts, they do not prescribe how to address the range of diversity and heterogeneity of the test takers. To reconsider how students' linguistic, cultural, and substantive characteristics (and other characteristics that students bring into the classroom) can be beneficial for equitable assessments, we reframe our test-taker variables as strengths. As a measurement community, we must rethink and redesign traditional assessment development processes to intentionally incorporate diverse student characteristics as assets—and reconsider how these asset-based perspectives can further influence how learners feel empowered and supported in both their personal identity and their social belonging within their community (Volman & Gilde, 2021; see also the Shared Power Principle and Engagement Principle sections in this report).

23308516, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12374, Wiley Online Library on [29/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The design and development of a culturally responsive assessment should consider how an asset-based principle operates within a larger theory of action making explicit how asset-based principles support fairness and equity. These assets should be incorporated through all phases of the design, development, scoring, and test use to measure what students know and are able to do. Adopting this asset-based principle toward test design actively pushes back against the traditional deficit narratives that disproportionately impact historically excluded groups (e.g., people of color, economically disadvantaged people, multilingual and multicultural people, people with disabilities). These deficit narratives typically focus on perceived weaknesses without recognizing the systemic inequities that create barriers and a "hostile obstacle course" that members of historically excluded groups must navigate to achieve commensurate goals compared to their peers (Berhe et al., 2021).

For example, recognizing learner variability and identity as part of understanding who are the target learners helps to inform what types of questions should be asked. Questions should also be asset-based to ensure that students have every opportunity to show what they know and can do. The antideficit framework (Harper, 2010) and asking different questions project (Vora et al., 2022) are examples that seek out and validate diverse perspectives; the former through framing questions around what students can do; the latter through recognizing different perspectives (e.g., Roy, 2008). These perspectives are illustrative of recognizing the strength and value in the diversity of students' perspectives, characteristics of the asset-based principle for culturally relevant assessments.

Language learning and assessment is one such example of student diversity that can be impacted by a deficit view of language and language practices. An asset-based view recognizes and values diversity in language practices. Unintended consequences of these deficit-oriented practices are reflected in the singular prioritization of maintaining linguistic ideologies (e.g., language type, register, dialect; Blum, 2017; Rosa & Flores, 2017). Instead, more contemporary, asset-based approaches redefine linguistic ideologies by representing language systems as integrated and fluid (e.g., Grosjean, 1989), normalizing the flexibility and "appropriateness" of language for all learners (Flores & Rosa, 2015; Rosa, 2016; Rosa & Flores, 2017), and using scoring procedures and interpretations that represent authentic practices, like strategically using the entire linguistic repertoire to communicate (i.e., like translanguaging using multiple languages or gestures; Bedore et al., 2005; García, 2009).

Additional empirical research has been conducted using asset-based approaches and seeks to amplify the voices, perspectives, and lived experiences for diverse groups including Black (Coleman & Davis, 2020; Gravel et al., 2021; Harper, 2010), Hispanic (López, 2017), Native American, and Alaskan Native (Wu et al., 2020). The study by Wu et al. (2020) represents an asset-based research perspective to investigate the strengths of students who are American Indian (AI) and Alaskan Native (AN) on NAEP mathematics in Grades 4 and 8. Rather than focus on the differences between groups, the researchers used an excellence lens to distinguish students who identified as AI/AN and who did better than non-AI/AN peers on 14 questions as measured by the percentage of students who answered a question correctly. A quantitative analysis was done to find items in the content areas of measurement, geometry, and number properties and operations. A qualitative item analysis uncovered important themes for items favoring AI/ANs, including context found in their everyday lives (via context or a visual representation) or relation to a cultural experience (situation or context). Items did not necessarily have the same connotation for non-AI/AN students. This represents important investigations in exploring how tests may favor specific populations and a method to investigate why such differences exist and provides future methods to investigate further why differences occur and the strengths of the learners. Other asset-based research has focused on the use of counter stories to support historically excluded populations (Battey & Franke, 2015; Golombek et al., 2022; Rodela & Rodriguez-Mojica, 2020). Counter stories are "stories of people whose experiences are not often told" (Solórzano & Yosso, 2002, p. 26). Counter stories provide a space for students to articulate their lived-in experiences (Morrison, 2017). The use of counter stories attempts to interrupt deficit ideologies and support indigenous knowledge, as well as value assets of diversity inclusive of racial, cultural context, lived experiences, and cultural wealth (Battey & Franke, 2015; Golombek et al., 2022; Rodela & Rodriguez-Mojica, 2020). Most importantly, counter stories provide students with meaning and legitimacy without detracting from their lived experiences. Counter stories provide a method to support the goal of an asset-based lens to refute external narratives and provide space for students to articulate their lived-in experiences (Morrison, 2017).

Asset-based narratives focus on the relative strengths and ideologies shared within the community group (García & Öztürk, 2017; Howard, 2013; Ramasubramanian et al., 2021). By no means do, we aim to account for all examples or recommendations in this paper, but instead we aim to highlight those that may be most relevant to our goal of developing

23308516, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12374, Wiley Online Library on [29/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

23308516, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12374, Wiley Online Library on [29/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

culturally relevant assessments that build in initial awareness of and value for student diversity as strengths rather than challenges, such as the Hawaiian DOE assessment development effort to value Hawaiian students' language and culture (Kūkea Shultz et al., 2019).

Summary and Considerations for Assessment Development

In this paper, we discussed five design principles of culturally responsive assessments. Research suggests that following these principles will result in a more equitable, inclusive, and informative assessment experience. Key to this process is *shared power*, including all stakeholders in every stage of the assessment process, from conceptualization to use of the resulting information. An inclusive process allows people from different groups to shape who and what should be assessed and how, so that they can be more assured that the results adequately reflect what each person knows and can do. Different groups of stakeholders need to agree on how assessment information will be processed and used, as those uses directly affect lives and livelihoods.

Context-rich assessments will foster *engagement*, which in turn will facilitate people's best performance. Developing such assessments requires participation of people from all segments of the target population in the assessment development process. This includes the way feedback from the assessment is shared with participants. The information should be presented in a way that encourages the individual to continue learning.

In a related way, culturally responsive assessments should reflect *high expectations* for participants. Assessments requiring higher order thinking—the interpretation and transformation of information—will help signal these expectations to participants. So will assessments that give people the opportunity to reflect on and revise their thinking through timely and targeted feedback. The experience of tackling and mastering rigorous material will help forge the mentality of a lifelong learner.

Because of the diversity of backgrounds of learners, assessments should allow for *flexibility* in the way material is presented, as well as in acceptable response formats. This flexibility may better foster academic engagement and sense of belonging. Flexibility will allow people better opportunities to show what they know and can do. Effective design of such assessments necessitates participation from multiple segments of the population.

Finally, assessments grounded in an *asset-based* perspective allow people to utilize their backgrounds and characteristics to demonstrate their capabilities. Such a perspective must allow for the multiple ways in which a person incorporates various cultural background and numerous characteristics to form a unique identity. Acknowledgment of the diversity of learner characteristics during the assessment development process will ideally lead to an assessment that allows an individual to leverage those unique qualities in completing the tasks. Similarly, any use of assessment results will take those individual characteristics into consideration.

We believe that taking these five principles into consideration during assessment development and use will result in more equitable assessment experiences. Table 1 lists some questions that assessment developers and users may want to consider as they engage at various points in the assessment life cycle. These questions are not meant to be exhaustive; rather they are meant to stimulate involved parties to be more inclusive of interested parties and more expansive in the issues considered when designing and using assessments.

The changing demographic, social, and political landscape has highlighted the need for shifting priorities in educational measurement. Instead of continuing traditional approaches to standardization and fairness, concentrated efforts are focused on *UNDERSTANDardization* (Sireci, 2020), conditional fairness (Mislevy, 2018), and cultural relevance and representation as means to drive transformative change toward culturally responsive assessment practices. The five principles highlighted in this framework are an initial step toward designing and developing culturally responsive assessments. Future efforts in this line of research include conceptual, theoretical, and empirical studies to further investigate and evaluate the impact these principles may have on assessment design and development procedures, as well as the ultimate goal of better meeting diverse learners' needs to better show what they know and can do.

Acknowledgments

The authors are listed in reverse alphabetical order. This is a collaborative effort in all respects; all authors contributed equally to this paper.

23308516, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12374, Wiley Online Library on [29/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Notes

[1] We use the phrase “historically excluded” to refer to learners of many groups (characterized by race, ethnicity, gender, culture, language, disability, economic status, or other identities) who have been actively barred from full participation in the educational system. We prefer this term to ones like “underrepresented minorities” or “traditionally underserved,” which may at once mask the differences among individuals under consideration and hide their active exclusion by others.

[2] The editor pointed out that producing a test with a White-centered context makes sense if the goal is to predict performance in White-centered institutions We agree. This is a failure of the entire educational system, a topic beyond the scope of this paper. While modifying the test alone will not completely rectify the situation, we believe the modifications will result in improvements.

[3] As an example, the PSAT/NMSQT® score report includes a section on “Course Recommendations,” which indicates whether a student’s “scores show you are ready to succeed in AP® courses” (College Board, 2022, p. 9).

[4] According to Randall (2021), an antiracist approach to assessment involves a conscious, political decision to confront the historic roots of the inequality that manifests in assessment. In practice, this approach would yield an assessment that is culturally responsive insofar as it focuses on issues and situations relevant to historically excluded groups. The terms “antiracist” and “culturally responsive,” while not synonymous, overlap to a great extent.

[5] In this context, when the authors talk about “test users,” they mean primarily the people who take the test. However, the authors’ argument and approach may apply as well to other stakeholders, who use the test information (e.g., scores) for decision-making or other purposes.

[6] Methodologies used to collect the data may include think-alouds, direct observation, process data, cognitive interviews, surveys, focus groups, and ethnographic studies.

[7] The National Center on Educational Outcomes (n.d.) offers seven elements of universal design of assessments, designed to improve accessibility of tests, although these elements do not address cultural responsiveness.

[8] Other terms, such as “assets-oriented” (e.g., Sharpe et al., 2000), “strengths perspective” (Wieck et al., 1989), or “strength-based” (e.g., Jimerson et al., 2004) can all be used to describe intentional framing of students’ assets to articulate strengths.

References

Abedi, J., Lord, C., Boscardin, C. K., & Miyoshi, J. (2001). *The effects of accommodations on the assessment of limited English proficient (LEP) students in the National Assessment of Educational Progress (NAEP)* (CSE Technical Report No. 537). National Center for Research on Evaluation, Standards, and Student Testing. https://cresst.org/publications/cresst-publication-2909/

ACT. (n.d.). *ACT scores for higher education professionals*. https://www.act.org/content/act/en/products-and-services/the-act-postsecondary-professionals/scores.html

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

Anderson, D. I., Magill, R. A., & Sekiya, H. (2001). Motor learning as a function of KR schedule and characteristics of task-intrinsic feedback. *Journal of Motor Behavior*, *33*(1), 59–66. https://doi.org/10.1080/00222890109601903

Araneda, S., & Sireci, S. (2021, June 9–11). *An experiential approach to test design and validation* [Paper presentation]. Annual meeting of the National Council on Measurement in Education, online.

Arcia, G., Macdonald, K., Patrinos, H. A., & Porta, E. (2011). *School autonomy and accountability*. https://openknowledge.worldbank.org/bitstream/handle/10986/21546/944500WP00PUBL0ing0Background0Paper.pdf;sequence=1

Aronson, B., Murphy, K. M., & Saultz, A. (2016). Under pressure in Atlanta: School accountability and special education practices during the cheating scandal. *Teachers College Record*, *118*(14), 1–26. https://doi.org/10.1177/0161468116118014

Babad, E. (1993). Teachers’ differential behavior. *Educational Psychology Review*, *5*, 347–376. https://doi.org/10.1007/BF01320223

Baker, E. L., O’Neil, H. F., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist*, *48*(12), 1210–1218. https://doi.org/10.1037/0003-066X.48.12.1210

Baker-Bell, A. (2020). We been knowin: Toward an antiracist language & literacy education. *Journal of Language and Literacy Education*, *16*(1), 1–12. https://files.eric.ed.gov/fulltext/EJ1253929.pdf

Baker-Bell, A., Paris, D., & Jackson, D. (2017). Learning Black language matters: Humanizing research as culturally sustaining pedagogy. *International Review of Qualitative Research*, *10*(4), 360–377. https://doi.org/10.1525/irqr.2017.10.4.360

Baldridge, B. J. (2014). Relocating the deficit: Reimagining Black youth in neoliberal times. *American Educational Research Journal*, *51*(3), 440–472. https://doi.org/10.3102/0002831214532514

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*(2), 213–238. https://doi.org/10.3102/00346543061002213

Barlowe, A., & Cook, A. (2016). Putting the focus on student engagement: The benefits of performance-based assessment. *American Educator*, *40*(1), 4–11. https://www.aft.org/ae/spring2016/barlowe-and-cook

Basileo, L. D., & Lyons, M. E. (2019). *The research base supporting the Rigor Diagnostic observation instrument*. LSI Applied Research Center. https://www.learningsciences.com/wpcontent/uploads/2021/08/LSI02-07-Rigor-Diagnostic-Observation-12-18-2019-Digital.pdf

Battey, D., & Franke, M. (2015). Integrating professional development on mathematics and equity: Countering deficit views of students of color. *Education and Urban Society*, *47*(4), 433–462. https://doi.org/10.1177/0013124513497788

Bedore, L. M., Peña, E. D., García, M., & Cortez, C. (2005). Conceptual versus monolingual scoring. *Language, Speech, and Hearing Services in Schools*, *36*(3), 188–200. https://doi.org/10.1044/0161-1461(2005/020)

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? In *FAccT'21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM. https://doi.org/10.1145/3442188.3445922

Bennett, R. E. (2021, April 8–12). *Equity and assessment in the post-Covid-19 era* [Paper presentation]. Annual meeting of the American Educational Research Association, online.

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement*, *28*(1), 77–92. https://doi.org/10.1111/j.1745-3984.1991.tb00345.x

Berhe, A. A., Barnes, R. T., Hastings, M. G., Mattheis, A., Schneider, B., Williams, B. M., & Marín-Spiotta, E. (2021). Scientists from historically excluded groups face a hostile obstacle course. *Nature Geoscience*, *15*, 2–4. https://doi.org/10.1038/s41561-021-00868-0

Bešić, E. (2020). Intersectionality: A pathway towards inclusive education? *Prospects*, *49*, 111–122. https://doi.org/10.1007/s11125-020-09461-6

Blackburn, B. R. (2017). Rigor and assessment in the classroom. *Instructional Leader*, *30*(4), 1–5. https://cdn.ymaws.com/tepsa.site-ym.com/resource/collection/5B4325B4-14A6-4B13-8CE9-5E8F6827E164/july2017.pdf

Blum, S. D. (2017). Unseen WEIRD assumptions: The so-called language gap discourse and ideologies of language, childhood, and learning. *International Multilingual Research Journal*, *11*(1), 23–38. https://doi.org/10.1080/19313152.2016.1258187

Boston, M., & Wolf, M. K. (2006). *Assessing academic rigor in mathematics instruction: The development of the instructional quality assessment toolkit* (CSE Technical Report No. 672). University of California, Los Angeles, National Center for Research on Evaluation, Standards and Student Testing. https://cresst.org/wp-content/uploads/R672.pdf

Bridgeman, B., Morgan, R., & Wang, M. (1997). Choice among essay topics: Impact on performance and validity. *Journal of Educational Measurement*, *34*(3), 273–286. https://doi.org/10.1111/j.1745-3984.1997.tb00519.x

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, *25*(1), 27–40. https://doi.org/10.1080/08957347.2012.635502

Brophy, J. E. (1982). How teachers influence what is taught and learned in classrooms. *The Elementary School Journal*, *83*(1), 1–13. https://doi.org/10.1086/461287

Burn, J. M., & Thongprasert, N. (2005). A culture-based model for strategic implementation of virtual education delivery. *International Journal of Education and Development Using ICT*, *1*(1), 32–52. http://ijedict.dec.uwi.edu/viewarticle.php?id=17

Buzhardt, J., & Semb, G. B. (2002). Item-by-item versus end-of-test feedback in a computer-based PSI course. *Journal of Behavioral Education*, *11*(2), 89–104. https://doi.org/10.1023/A:1015479225777

Campbell, J. R., & Donahue, P. L. (1997). *Students selecting stories: The effects of choice in reading assessment. Results from the NAEP reader special study of the 1994 National Assessment of Educational Progress* (NCES 97-491). U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics. http://nces.ed.gov/naep/pdf/main1994/97491.pdf

CAST. (2018). *Universal Design for Learning Guidelines version 2.2*. http://udlguidelines.cast.org

Cohen, A. B. (2009). Many forms of culture. *American Psychologist*, *64*(3), 194–204. https://doi.org/10.1037/a0015308

Coleman, S. T., & Davis, J. (2020). Using asset-based pedagogy to facilitate STEM learning, engagement, and motivation for Black middle school boys. *Journal of African American Males in Education*, *11*(2), 76–94. https://jaamejournal.scholasticahq.com/article/18095-using-asset-based-pedagogy-to-facilitate-stem-learning-engagement-and-motivation-for-black-middle-school-boys

College Board. (2022). *PSAT/NMSQT, Preliminary SAT/National Merit Scholarship Qualifying Test: Understanding scores 2022–23*. https://satsuite.collegeboard.org/media/pdf/psat-nmsqt-understanding-scores.pdf

College Board. (n.d.). *SAT suite of assessments*. https://satsuite.collegeboard.org/sat

Collins, P. H. (2015). Intersectionality's definitional dilemmas. *Annual Review of Sociology*, *41*(1), 1–20. https://doi.org/10.1146/annurev-soc-073014-112142

Corbett, A. T., & Anderson, J. R. (2001). Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. In J. Jacko & A. Sears (Eds.), *CHI'O1: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 245–252). ACM Press. https://doi.org/10.1145/365024.365111

Cormier, D. C., Altman, J, Shyyan, V., & Thurlow, M. L. (2010). *A summary of the research on the effects of test accommodations: 2007–2008* (Technical Report No. 56). University of Minnesota, National Center on Educational Outcomes. https://nceo.info/Resources/publications/OnlinePubs/Tech56/default.htm

23308516, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12374, Wiley Online Library on [29/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Couch, M., II, Frost, M., Sangtiago, J., & Hilton, A. (2021). Rethinking standardized testing from an access, equity and achievement perspective: Has anything changed for African American students? *Journal of Research Initiatives*, *5*(3), 6. https://digitalcommons.uncfsu.edu/jri/vol5/iss3/6/

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, *1989*(1), 139–167. http://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8

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

Crenshaw, K. W. (2017). *On intersectionality: Essential writings*. The New Press.

Cross, B. E. (2007). Urban school achievement gap as a metaphor to conceal U.S. apartheid education. *Theory Into Practice*, *46*(3), 247–255. https://doi.org/10.1080/00405840701402299

Delpit, L. (2012). *"Multiplication is for White people": Raising expectations for other people's children*. The New Press.

Dweck, C. S., & Yeager, D. S. (2019). Mindsets: A view from two eras. *Perspectives on Psychological Science*, *14*(3), 481–496. https://doi.org/10.1177/1745691618804166

Edmunds, J., Arshavsky, N., Glennie, E., Charles, K., & Rice, O. (2017). The relationship between project-based learning and rigor in STEM-focused high schools. *Interdisciplinary Journal of Problem-Based Learning*, *11*(1), 3. https://doi.org/10.7771/1541-5015.1618

Edulastic. (n.d.). *Introducing testlets: Engaging, scenario-based tasks for assessment or self-directed distance learning*. https://edulastic.com/ets-testlets/

Edwards, A. R., & Beattie, R. L. (2016). Promoting student learning and productive persistence in developmental mathematics: Research frameworks informing the Carnegie pathways. *NADE Digest*, *9*(1), 30–39. https://thenoss.org/resources/Pictures/Digest/NADEDigest_Winter_2016_2.pdf

Emler, T. E., Zhao, Y., Deng, J., Yin, D., & Wang, Y. (2019). Side effects of large-scale assessments in education. *ECNU Review of Education*, *2*(3), 279–296. https://doi.org/10.1177/2096531119878964

Esteban-Guitart, M. (2012). Towards a multimethodological approach to identification of funds of identity, small stories and master narratives. *Narrative Inquiry*, *22*(1), 173–180. https://doi.org/10.1075/ni.22.1.12est

Esteban-Guitart, M., & Moll, L. C. (2014). Funds of identity: A new concept based on the funds of knowledge approach. *Culture & Psychology*, *20*(1), 31–48. https://doi.org/10.1177/1354067X13515934

ETS. (n.d.). *The Praxis tests: Supporting aspiring teachers on their journey to the classroom*. https://www.ets.org/praxis/site/epp/about/core.html

Evans, C. (2021, September) *A culturally responsive classroom assessment framework. The intersections of equity, pedagogy, and sociocultural assessment*. https://www.nciea.org/blog/classroom-assessment/culturally-responsive-classroom-assessment-framework

Finn, B., & Tauber, S. K. (2015). When confidence is not a signal of knowing: How students' experiences and beliefs about processing fluency can lead to miscalibrated confidence. *Educational Psychology Review*, *27*(4), 567–586. https://doi.org/10.1007/s10648-015-9313-7

Flores, N., & Rosa, J. (2015). Undoing appropriateness: Raciolinguistic ideologies and language diversity in education. *Harvard Educational Review*, *85*(2), 149–171. https://doi.org/10.17763/0017-8055.85.2.149

Ford, D. Y., & Scott, M. F. T. (2021). Culturally responsive and relevant curriculum: The revised Bloom-Banks matrix. In K. R. Stephens & F. A. Karnes (Eds.), *Introduction to curriculum design in gifted education* (pp. 331–349). Routledge. https://doi.org/10.4324/9781003235842-20

Francis, D., Rivera, M., Lesaux, N., Kieffer, M., & Rivera, H. (2006). *Practical guidelines for the education of English language learners: Research-based recommendations for the use of accommodations in large-scale assessments*. RMC Research Corporation, Center on Instruction. https://www.centeroninstruction.org/practical-guidelines-for-the-education-of-english-language-learners-research-based-recommendations-for-the-use-of-accommodations-in-large-scale-assessments

García, E. E., & Öztürk, M. 2017). *An asset-based approach to Latino education in the United States: Understanding gaps and advances*. Routledge.

García, O. (2009). Education, multilingualism and translanguaging in the 21st century. In T. Skutnabb-Kangas, R. Phillipson, A. K. Mohanty, & M. Panda (Eds.), *Social justice through multilingual education* (pp. 140–158). Multilingual Matters. https://doi.org/10.21832/9781847691910-011

Gay, G. (2000). *Culturally responsive teaching: Theory, research, and practice*. Teachers College Press.

Giersch, J. (2018). Academic tracking, high-stakes tests, and preparing students for college: How inequality persists within schools. *Educational Policy*, *32*(7), 907–935. https://doi.org/10.1177/0895904816681526

Golombek, P., Olszewska, A. I., & Coady, M. (2022). Humanizing power of counter-stories: Teachers' understandings of emergent bilinguals in rural settings. *Teaching and Teacher Education*, *113*, 103655. https://doi.org/10.1016/j.tate.2022.103655

Gorski, P. C. (2011). Unlearning deficit ideology and the scornful gaze: Thoughts on authenticating the class discourse in education. *Counterpoints*, *402*, 152–173. https://www.jstor.org/stable/42981081

23308516, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12374, Wiley Online Library on [29/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Gravel, B. E., Tucker-Raymond, E., Wagh, A., Klimczak, S., & Wilson, N. (2021). More than mechanisms: Shifting ideologies for asset-based learning in engineering education. *Journal of Pre-College Engineering Education Research (J-PEER)*, *11*(1), 15. https://doi.org/10.7771/2157-9288.1286

Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language*, *36*(1), 3–15. https://doi.org/10.1016/0093-934X(89)90048-5

Gutiérrez, K. D., & Rogoff, B. (2003). Cultural ways of learning: Individual traits or repertoires of practice. *Educational Researcher*, *32*(5), 19–25. https://doi.org/10.3102/0013189X032005019

Gutiérrez, R. (2012). Context matters: How should we conceptualize equity in mathematics education? In B. Herbel-Eisenmann, J. Choppin, D. Wagner, & D. Pimm (Eds.), *Equity in discourse for mathematics education: Theories, practices, and policies* (pp. 17–33). Springer. https://doi.org/10.1007/978-94-007-2813-4_2

Harper, S. R. (2010). An anti-deficit achievement framework for research on students of color in STEM. *New Directions for Institutional Research*, *2010* (148), 63–74. https://doi.org/10.1002/ir.362

Harper, S. R. (2015). Success in these schools? Visual counternarratives of young men of color and urban high schools they attend. *Urban Education*, *50*(2), 139–169. https://doi.org/10.1177/0042085915569738

Hartocollis, A. (2019, November 5). Harvard won a key affirmative action battle. But the war's not over. *The New York Times*. https://www.nytimes.com/2019/10/02/us/harvard-admissions-lawsuit.html

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hess, K. K., Carlock, D., Jones, B., & Walkup, J. R. (2009, June). *What exactly do "fewer, clearer, and higher standards" really look like in the classroom? Using a cognitive rigor matrix to analyze curriculum, plan lessons, and implement assessments* [Paper presentation]. Council of Chief State School Officers meeting, Detroit, MI, USA.

Hickson, S., Reed, W. R., & Sander, N. (2012). Estimating the effect on grades of using multiple-choice versus constructive-response questions: Data from the classroom. *Educational Assessment*, *17*(4), 200–213. https://doi.org/10.1080/10627197.2012.735915

Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 371–404). Information Age.

Hogg, L., & Volman, M. (2020). A synthesis of funds of identity research: Purposes, tools, pedagogical approaches, and outcomes. *Review of Educational Research*, *90*(6), 862–895. https://doi.org/10.3102/0034654320964205

Hood, S. (1998). Culturally responsive performance-based assessment: Conceptual and psychometric considerations. *The Journal of Negro Education*, *67*(3), 187–196. https://doi.org/10.2307/2668188

Howard, T. C. (2013). How does it feel to be a problem? Black male students, schools, and learning in enhancing the knowledge base to disrupt deficit frameworks. *Review of Research in Education*, *37*(1), 54–86. https://doi.org/10.3102/0091732X12462985

Jennings, J. L., & Bearak, J. M. (2014). "Teaching to the test" in the NCLB era: How test predictability affects our understanding of performance. *Educational Researcher*, *43*(8), 381–389. https://doi.org/10.3102/0013189X14554449

Jimerson, S. R., Sharkey, J. D., Nyborg, V. M., & Furlong, M. J. (2004). Strength-based assessment and school psychology: A summary and synthesis. *The California School Psychologist*, *9*(1), 9–19. https://doi.org/10.1007/BF03340903

Karem, R. W., & Washington, K. N. (2021). The cultural and diagnostic appropriateness of standardized assessments for dual language learners: A focus on Jamaican preschoolers. *Language, Speech, and Hearing Services in Schools*, *52*(3), 807–826. https://doi.org/10.1044/2021_LSHSS-20-00106

Kavanagh, K. M., & Fisher-Ari, T. R. (2020). Curricular and pedagogical oppression: Contradictions within the juggernaut accountability trap. *Educational Policy*, *34*(2), 283–311. https://doi.org/10.1177/0895904818755

Kūkea Shultz, P., & Englert, K. (2021). Cultural validity as foundational to assessment development: An indigenous example. *Frontiers in Education*, *6*, 701973. https://doi.org/10.3389/feduc.2021.701973

Kūkea Shultz, P., Englert, K., Krug, K., Ruth, K., Ching, L., & Franco, L. (2019). *Context matters: The promise of cultural and community validity in assessment* [Conference session]. Third Annual NCME Special Conference on Classroom Assessment, Boulder CO, United States. https://drive.google.com/file/d/1Q0JXb7cYwnCuItoY9961yp-F3jFylwWT/view

Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal*, *32*(3), 465–491. https://doi.org/10.3102/00028312032003465

Ladson-Billings, G. (2006). From the achievement gap to the education debt: Understanding achievement in U.S. schools. *Educational Researcher*, *35*(7), 3–12. https://doi.org/10.3102/0013189X035007003

Ladson-Billings, G. (2021). I'm here for the hard re-set: Post pandemic pedagogy to preserve our culture. *Equity & Excellence in Education*, *54*(1), 68–78. https://doi.org/10.1080/10665684.2020.1863883

Landl, E. (2021, October). *Culturally responsive assessment and balanced assessment systems* [Webinar]. REL Pacific Webinar. https://www.youtube.com/watch?v=rT3tHRyCMhA

Lane, S. (2020). *Test-based accountability systems: The importance of paying attention to consequences* (Research Report No. RR-20-02). ETS. https://doi.org/10.1002/ets2.12283

23308516, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12374, Wiley Online Library on [29/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Lane, S., Parke, C. S., & Stone, C. A. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement: Issues and Practice*, *17*(2), 24–28. https://doi.org/10.1111/j.1745-3992.1998.tb00830.x

Lane, S., & Stone, C. A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice*, *21*(1), 23–30. https://doi.org/10.1111/j.1745-3992.2002.tb00082.x

Lee, T. S., & Quijada Cerecer, P. D. (2010). (Re) claiming native youth knowledge: Engaging in socio-culturally responsive teaching and relationships. *Multicultural Perspectives*, *12*(4), 199–205. https://doi.org/10.1080/15210960.2010.527586

Lehman, B., Sparks, J. R., & Zapata-Rivera, D. (2018). When should an adaptive assessment care? In N. Guin & A. Kumar (Eds.), *Proceedings of ITS 2018: Intelligent Tutoring Systems 14th International Conference, Workshop on Exploring Opportunities for Caring Assessments* (pp. 87–94). ITS. https://ceur-ws.org/Vol-2354/w3paper1.pdf

Lewis, E. L., & Hunt, B. (2019). High expectations: Increasing outcomes for Black students in urban schools. *Urban Education Research & Policy Annuals*, *6*(2), 78–89. https://journals.charlotte.edu/urbaned/article/view/893/808

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice*, *13*(1), 5–8, 15. https://doi.org/10.1111/j.1745-3992.1994.tb00778.x

López, F. A. (2017). Altering the trajectory of the self-fulfilling prophecy: Asset-based pedagogy and classroom dynamics. *Journal of Teacher Education*, *68*(2), 193–212. https://doi.org/10.1177/0022487116685751

Lundberg, C. A., Kim, Y. K., Andrade, L. M., & Bahner, D. T. (2018). High expectations, strong support: Faculty behaviors predicting Latina/o community college student learning. *Journal of College Student Development*, *59*(1), 55–70. https://doi.org/10.1353/csd.2018.0004

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

Marion, S. F., & Pellegrino, J. W. (2006). A validity framework for evaluating the technical quality of alternate assessments. *Educational Measurement: Issues and Practice*, *25*(4), 47–57. https://doi.org/10.1111/j.1745-3992.2006.00078.x

Martin, P. C. (2012). Misuse of high-stakes test scores for evaluative purposes: Neglecting the reality of schools and students. *Current Issues in Education*, *15*(3). https://cie.asu.edu/ojs/index.php/cieatasu/article/view/1061/391

Mason, B. J., & Bruning, R. (2001). *Providing feedback in computer-based instruction: What the research tells us*. University of Nebraska-Lincoln, Center for Instructional Innovation.

Matsumura, L. C., Slater, S. C., & Crosson, A. (2008). Classroom climate, rigorous instruction and curriculum, and students' interactions in urban middle schools. *The Elementary School Journal*, *108*(4), 293–312. https://doi.org/10.1086/528973

Matusevich, M. N., O'Connor, K. A., & Hargett, M. V. P. (2009). The nonnegotiables of academic rigor. *Gifted Child Today*, *32*(4), 44–52. https://doi.org/10.1177/107621750903200412

McCarthy, J., & Wright, P. (2015). *Taking [A]part: The politics and aesthetics of participation in experience-centered design*. MIT Press.

McClellan, C., Joe, J., & Bassett, K. (2016). *Still on the right trajectory: State teachers of the year compare former and new state assessments*. National Network of State Teachers of the Year. https://files.eric.ed.gov/fulltext/ED581173.pdf

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement*, *23*(3), 187–194. https://doi.org/10.1177/01466219922031310

Milner, R. H. (2011). Culturally relevant pedagogy in a diverse urban classroom. *Urban Review: Issues and Ideas in Public Education*, *43*(1), 66–89. https://doi.org/10.1007/s11256-009-0143-0

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge. https://doi.org/10.4324/9781315871691

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (Research Report No. RR-03-16). ETS. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Mitchell, K., Shkolnik, J., Song, M., Uekawa, K., Murphy, R., & Garet, M., & Means, B. (2005). *Rigor, relevance, and results: The quality of teacher assignments and student work in new and conventional high schools*. American Institutes of Research and SRI International. https://www.sri.com/publication/education-learning-pubs/teaching-quality-pubs/rigor-relevance-and-results-the-quality-of-teacher-assignments-and-student-work-in-new-and-conventional-high-schools/

Montenegro, E., & Jankowski, N. A. (2017). *Equity and assessment: Moving towards culturally responsive assessment* (Occasional Paper No. 29). National Institute for Learning Outcomes Assessment. https://www.learningoutcomesassessment.org/wp-content/uploads/2019/02/OccasionalPaper29.pdf

Morrison, K. L. (2017). Informed asset-based pedagogy: Coming correct, counter-stories from an information literacy classroom. *Library Trends 66*(2), 176–218. https://doi.org/10.1353/lib.2017.0034

Nasir, N. I. S., & Hand, V. M. (2006). Exploring sociocultural perspectives on race, culture, and learning. *Review of Educational Research*, *76*(4), 449–475. https://doi.org/10.3102/00346543076004449

National Center for Educational Statistics. (n.d.). *NAEP*. https://nces.ed.gov/nationsreportcard/

National Center on Educational Outcomes. (n.d.). *Universal design of assessments*. https://nceo.info/Assessments/universal_design

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. The National Academies Press. https://doi.org/10.17226/10019.

23308516, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12374, Wiley Online Library on [29/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Nichols, S. L., & Harris, L. R. (2016). Accountability assessment's effects on teachers and schools. In G. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 40–56). Routledge.

No Child Left Behind Act of 2001, 20 U.S.C. § 6319 et seq. (2002).

Noguera, P. A. (2003). The trouble with Black boys: The role and influence of environmental and cultural factors on the academic performance of African American males. *Urban Education*, *38*(4), 431–459. https://doi.org/10.1177/0042085903038004005

Nordling, L. (2019). A fairer way forward for AI in health care. *Nature*, *573*, S103–S105. https://doi.org/10.1038/d41586-019-02872-2

NWEA. (n.d.). *MAP growth*. https://www.nwea.org/map-growth/

O'Donnell, F., & Sireci, S. G. (2021). Language matters: Teacher and parent perceptions of achievement labels from educational tests. *Educational Assessment*, *27*(1), 1–26. https://doi.org/10.1080/10627197.2021.2016388

O'Dwyer, E. P., Sparks, J. R., & Nabors Oláh, L. (2023). Enacting a process for developing culturally relevant classroom assessments. *Applied Measurement in Education*, *36*(3), 286–303. https://doi.org/10.1080/08957347.2023.2214652

O'Reilly, T., & Sabatini, J. (2013). *Reading for understanding: How performance moderators and scenarios impact assessment design* (Research Report No. RR-13-31). https://doi.org/10.1002/j.2333-8504.2013.tb02338.x

Packman, S., Camara, W. J., & Huff, K. (2010). A snapshot of industry and academic professional activities, compensation, and engagement in educational measurement. *Educational Measurement: Issues and Practice*, *29*(3), 15–24. https://doi.org/10.1111/j.1745-3992.2010.00180.x

Paige, D. D., Sizemore, J. M., & Neace, W. P. (2013). Working inside the box: Exploring the relationship between student engagement and cognitive rigor. *NASSP Bulletin*, *97*(2), 105–123. https://doi.org/10.1177/0192636512473505

Peters, H. C., Luke, M., Bernard, J., & Trepal, H. (2020). Socially just and culturally responsive leadership within counseling and counseling psychology: A grounded theory investigation. *The Counseling Psychologist*, *48*(7), 953–985. https://doi.org/10.1177/0011000020937431

Quinn, D. M. (2020). Experimental effects of "achievement gap" news reporting on viewers' racial stereotypes, inequality explanations, and inequality prioritization. *Educational Researcher*, *49*(7), 482–492. https://doi.org/10.3102/0013189X20932469

Quinn, D. M., & Desruisseaux, T.-M. (2022). *Replicating and extending effects of "achievement gap" discourse* (ED Working Paper No. 22-628). Annenberg Institute at Brown University. https://doi.org/10.26300/2pky-ch12

Ramasubramanian, S., Riewestahl, E., & Landmark, S. (2021). The trauma-informed equity-minded asset-based model (TEAM): The six R's for social justice-oriented educators. *Journal of Media Literacy Education*, *13*(2), 29–42. https://doi.org/10.23860/JMLE-2021-13-2-3

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

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

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

Ro, J. (2019). Learning to teach in the era of test-based accountability: A review of research. *Professional Development in Education*, *45*(1), 87–101. https://doi.org/10.1080/19415257.2018.1514525

Rodela, K. C., & Rodriguez-Mojica, C. (2020). Equity leadership informed by community cultural wealth: Counterstories of Latinx school administrators. *Educational Administration Quarterly*, *56*(2), 289–320. https://doi.org/10.1177/0013161X19847513

Roe, K. (2019). Supporting student assets and demonstrating respect for funds of knowledge. *Journal of Invitational Theory and Practice*, *25*, 5–13. https://eric.ed.gov/?id=EJ1251817

Rogoff, B. (2003). *The cultural nature of human development*. Oxford University Press.

Rojas, L., & Liou, D. D. (2017). Social justice teaching through the sympathetic touch of caring and high expectations for students of color. *Journal of Teacher Education*, *68*(1), 28–40. https://doi.org/10.1177/0022487116676314

Rosa, J., & Flores, N. (2017). Unsettling race and language: Toward a raciolinguistic perspective. *Language in Society*, *46*(5), 621–647. https://doi.org/10.1017/S0047404517000562

Rosa, J. D. (2016). Standardization, racialization, languagelessness: Raciolinguistic ideologies across communicative contexts. *Linguistic Anthropology*, *26*(2), 162–183. https://doi.org/10.1111/jola.12116

Roy, D. (2008). Asking different questions: Feminist practices for the natural sciences. *Hypatia*, *23*(4), 134–157. https://doi.org/10.1111/j.1527-2001.2008.tb01437.x

Rubie-Davies, C., Hattie, J., & Hamilton, R. (2006). Expecting the best for students: Teacher expectations and academic outcomes. *British Journal of Educational Psychology*, *76*(3), 429–444. https://doi.org/10.1348/000709905X53589

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivation: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54–67. https://doi.org/10.1006/ceps.1999.1020

Ryan, R. M., & Deci, E. L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. The Guilford Press. https://doi.org/10.1521/978.14625/28806

23308516, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12374, Wiley Online Library on [29/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Ryan, R. M., & Deci, E. L. (2019). Brick by brick: The origins, development, and future of self-determination theory. *Advances in Motivation Science*, *6*, 111–156. https://doi.org/10.1016/bs.adms.2019.01.001

Sabatini, J., O'Reilly, T., & Doorey, N. A. (2018). *Retooling literacy education for the 21st century: Key findings of the Reading for Understanding initiative and their implications* [Unpublished manuscript]. ETS.

Sabatini, J., O'Reilly, T., Weeks, J., & Wang, Z. (2019). Engineering a 21st century reading comprehension assessment system utilizing scenario-based assessment techniques. *International Journal of Testing*, *20*(1), 1–23. https://doi.org/10.1080/15305058.2018.1551224

Scott, M. T. (2014). Using the Blooms–Banks matrix to develop multicultural differentiated lessons for gifted students. *Gifted Child Today*, *37*(3), 163–168. https://doi.org/10.1177/1076217514532275

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research, and Evaluation*, *22*, 4. https://doi.org/10.7275/swgt-rj52

Sharpe, P. A., Greaney, M. L., Lee, P. R., & Royce, S. W. (2000). Assets-oriented community assessment. *Public Health Reports*, *115*(2–3), 205–211. https://doi.org/10.1093/phr/115.2.205

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing*, *18*(4), 373–391. https://doi.org/10.1177/026553220101800404

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Simson, D. (2014). Exclusion, punishment, racism and our schools: A critical race theory perspective on school discipline. *UCLA Law Review*, *61*, 506–563.

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

Skinner, E. A., Wellborn, J. G., & Connell, J. P. (1990). What it takes to do well in school and whether I've got it: A process model of perceived control and children's engagement and achievement in school. *Journal of Educational Psychology*, *82*(1), 22–32. https://doi.org/10.1037/0022-0663.82.1.22

Slee, J. (2010). A systemic approach to culturally responsive assessment practices and evaluation. *Higher Education Quarterly*, *64*(3), 246–260. https://doi.org/10.1111/j.1468-2273.2010.00464.x

Smarter Balanced Assessment Consortium. (2022, October 14). *2020–21 summative technical report*. https://technicalreports.smarterbalanced.org/2020-21_summative-report/_book/

Solano-Flores, G. (2011). Assessing the cultural validity of assessment practices: An introduction. In M. del Rosario Basterra, E. Trumbull, & G. Solano-Flores (Eds.), *Cultural validity in assessment: Addressing linguistic and cultural diversity* (pp. 3–21). Routledge.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching*, *38*(5), 553–573. https://doi.org/10.1002/tea.1018

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

Sonnert, G., Barnett, M. D., & Sadler, P. M. (2019). Short-term and long-term consequences of a focus on standardized testing in AP calculus classes. *The High School Journal*, *103*(1), 1–17. https://doi.org/10.1353/hsj.2020.0000

Stembridge, A. (2020). *Culturally responsive education in the classroom: An equity framework for pedagogy*. Routledge. https://doi.org/10.4324/9780429441080

Story, M. F., Mueller, J. L., & Mace, R. L. (1998). *The universal design file: Designing for people of all ages and abilities*. North Carolina State University, Center for Universal Design.

Style, E. (1996). Curriculum as window and mirror. *Social Science Record*, *33*(2), 35–45. (Original work published 1988).

Tavassolie, T., & Winsler, A. (2019). Predictors of mandatory 3rd grade retention from high-stakes test performance for low-income, ethnically diverse children. *Early Childhood Research Quarterly*, *48*, 62–74. https://doi.org/10.1016/j.ecresq.2019.02.002

Taylor, C. S., & Nolen, S. B. (2022). *Culturally and socially responsible assessment: Theory, research, and practice*. Teachers College Press.

Thissen, D., Wainer, H., & Wang, X.-B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement*, *31*(2), 113–123. https://doi.org/10.1111/j.1745-3984.1994.tb00437.x

Thompson, G. L., & Allen, T. G. (2012). Four effects of the high-stakes testing movement on African American K-12 students. *The Journal of Negro Education*, *81*(3), 218–227. https://doi.org/10.7709/jnegroeducation.81.3.0218

Tractenberg, R. E., Gushta, M. M., Mulroney, S. E., & Weissinger, P. A. (2013). Multiple choice questions can be designed or revised to challenge learners' critical thinking. *Advances in Health Sciences Education*, *18*(5), 945–961. https://doi.org/10.1007/s10459-012-9434-4

Trumbull, E., & Nelson-Barber, S. (2019). The ongoing quest for culturally-responsive assessment for indigenous students in the U.S. *Frontiers in Education*, *4*, 40, 1–11. https://doi.org/10.3389/feduc.2019.00040

Tung, R. (2017). Performance-based assessment: Meeting the needs of diverse learners. *Voices in Urban Education*, *46*, 3–5.

23308516, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12374, Wiley Online Library on [29/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Van De Vijver, F. J. R., & Phalet, K. (2004). Assessment in multicultural groups: The role of acculturation. *Applied Psychology*, *53*(2), 215 – 236. https://doi.org/10.1111/j.1464-0597.2004.00169.x

Vega, D., Moore, J. L., III, Baker, C. A., Bowen, N. V., Hines, E. M., & O'Neal, B. (2012). Salient factors affecting urban African American students' achievement: Recommendations for teachers, school counselors, and school psychologists. In J. L. Moore III & C. W. Lewis (Eds.), *African American students in urban schools: Critical issues and solutions for achievement* (pp. 113 – 139). Peter Lang Publishers.

Volman, M., & Gilde, J. (2021). The effects of using students' funds of knowledge on educational outcomes in the social and personal domain. *Learning, Culture and Social Interaction*, *28*, 100472. https://doi.org/10.1016/j.lcsi.2020.100472

Vora, K., McCullough, S., & Giordano, S. (2022). Asking different questions in STEM research: Feminist STS approaches to STEM pedagogy. *ADVANCE Journal*, *3*(1), 1 – 38. https://doi.org/10.5399/osu/ADVJRNL.3.1.10

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Walker, M. E. (2007, April 9 – 13). *Criteria to consider when reporting constructed response scores* [Paper presentation]. Annual meeting of the National Council on Measurement in Education, Chicago, IL, United States.

Walstad, W. B., & Becker, W. E., Jr. (1994). Achievement differences on multiple-choice and essay tests in economics. *American Economic Review*, *84*(2), 193 – 196.

Washor, E., & Mojkowski, C. (2007). What do you mean by rigor? *Educational Leadership*, *64*(4), 84 – 87.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments on mathematics and science education* (Research Monograph No. 6). Council of Chief State School Officers.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states* (Research Monograph No. 18). Council of Chief State School Officers.

Wentz, B., Jaeger, P. T., & Lazar, J. (2011). Retrofitting accessibility: The legal inequality of after-the-fact online access for persons with disabilities in the United States. *First Monday*, *16*(11). https://firstmonday.org/ojs/index.php/fm/article/download/3666/3077

Wieck, A., Rapp, C., Sullivan, W. P., & Kisthardt, S. (1989). A strengths perspective for social work practice. *Social Work*, *34*(4), 350 – 354. https://doi.org/10.1093/sw/34.4.350

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68 – 81. https://doi.org/10.1006/ceps.1999.1015

Wise, S. L., & Smith, L. F. (2011). A model of examinee test-taking effort. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High-stakes testing in education: Science and practice in K-12 settings* (pp. 139 – 153). American Psychological Association. https://doi.org/10.1037/12330-009

Wise, S. L., & Smith, L. F. (2016). The validity of assessment when students don't give good effort. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 204 – 220). Routledge.

Wu, K., Chaphalkar, R., Hecker, M., & Lask, E. (2020). Hidden strengths of American Indian and Alaska Native students in mathematics as measured by the National Assessment of Educational Progress. *Journal of American Indian Education*, *59*(2 – 3), 7 – 32. https://doi.org/10.1353/jaie.2020.0008

Yuksel, D., & Inan, B. (2013). A review of major strands in discourse analysis in language teaching. In D. Yuksel & B. Inan (Eds.), *Discourse perspectives on second and/or foreign language teaching and learning* (pp. 1 – 12). Nova.

Zapata-Rivera, D., Lehman, B., & Sparks, J. R. (2020). Learner modeling in the context of caring assessments. In R. A. Sottilare & J. Schwarz (Eds.), *Adaptive instructional systems: HCII 2020* (pp. 422 – 431). Springer. https://doi.org/10.1007/978-3-030-50788-6_31

Suggested citation:

Walker, M. E., Olivera-Aguilar, M., Lehman, B., Laitusis, C., Guzman-Orth, D., & Gholson, M. (2023). *Culturally responsive assessment: Provisional principles* (Research Report No. RR-23-11). ETS. https://doi.org/10.1002/ets2.12374

Action Editor: Brent Bridgeman

Reviewers: Jesse Sparks and Carolyn Wylie

ETS, the ETS logo, and PRAXIS are registered trademarks of Educational Testing Service (ETS). ADVANCED PLACEMENT, AP, SAT is a registered trademark of the College Board. PSAT/NMSQT is a registered trademark of the College Board and the National Merit Scholarship Corporation. All other trademarks are property of their respective owners.

Find other ETS-published reports by searching the ETS ReSEARCHER database.

23308516, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ets2.12374, Wiley Online Library on [29/12/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License