A peer reviewed, open-access electronic journal: ISSN 1531-7714

Culturally Sustaining Content Development in Standardized Testing: Practical Considerations and Lessons Learned

Sihua Hu, *Educational Records Bureau*
Rebecca Meyer, *Educational Records Bureau*

Abstract: There is a pressing need to bridge the theoretical discussions and emerging empirical evidence of the benefits of culturally relevant, responsive, and sustaining assessments with the reality of its operational implementation. This paper details our efforts to operationalize culturally sustaining (CS) content development for a high-stakes, multiple-choice admission assessment for grades 2-12. Across two rounds of content development trials, we outline our design for different stages of content development, and discuss what works and what is less effective in our iterative processes. We collected and analyzed both qualitative and quantitative data to synthesize our lessons learned into three areas: organizing effective student focus groups for topic generation, supporting CS item authoring, and reconceptualizing item evaluation practices. We also provide practical strategies based on our iterative methods and lessons learned for other testing programs considering implementing CS content development. We then present our design of a comprehensive content development workflow joining an established conventional development flow with a proposed CS content development flow in operation to illustrate scalability. Finally, we discuss the limitations of our work and future avenues of research to accumulate additional validity evidence for CS assessment content.

Keywords: Culturally Sustaining Assessment, Assessment Content Development, High-stakes K-12 Testing, Student-Centered Assessment Design

Introduction

There is growing awareness in the educational measurement field of ways that large-scale standardized assessments perpetuate the marginalization of minoritized students. It has long been understood that a truly "culture-free" test cannot be constructed, given that tests are inherently "cultural devices" (Solano-Flores & Trumbull, 2003). Nonetheless, some current assessment development processes promote cultural "neutrality" in assessment items, often inadvertently introducing bias (Taylor & Ferrara, 2025). The efforts to decontextualize test content, intended to prevent bias, can instead center the dominant culture by avoiding references to other cultures (Randall, 2021). This, in turn, may reduce engagement and limit opportunities for students from underrepresented backgrounds to demonstrate their competence.

Consequently, many researchers and measurement experts are calling for a critical examination of the assumptions embedded in current assessment practices, particularly concerning test design, content development, review processes, item analysis, scoring, reporting, and policies on test uses (Dixon-Roman, 2020; Randall, 2021; Russel, 2023; Sireci, 2020). Although the field is collectively moving towards a culturally responsive/sustaining/anti-racist assessment system with accumulative research on guiding principles and conceptual frameworks, the Standards (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) that test developers often turn to for guidance have yet to reflect this shift. Current efforts to revise the Standards center on the reconceptualization of fairness and equity in validity arguments and aligning psychometrics practices with these new definitions ("Reconsidering Assessment Fairness," 2024). As the measurement field undergoes a paradigm shift, practitioners must continue to contribute empirical evidence on the practical implementation of these renewed conceptualizations of equity and fairness alongside ongoing theoretical discussions that may still be provisional.

In this paper, we extend these discussions to the context of a high-stakes, standardized, multiple-choice admission assessment used for entry into independent schools, providing practical considerations and lessons learned from our recent efforts to advance our understanding of approaches to operationalize culturally sustaining item development. This work provides an example of how testing organizations, especially those on the smaller side, can plan for the incorporation of culturally sustaining content into their existing assessment program and workflow in ways that are consistent with prior research on culturally responsive content (e.g., Steedle, et al., 2023; Lyons, et al., 2022) as well as more recent calls to action (Bennett, 2023; Lyons et al., 2021; Randall, 2021; Sireci, 2020; Walkers et al. 2023). Specifically, we share observations and recommendations based on two rounds of culturally sustaining assessment content development in four content areas—mathematics, quantitative reasoning, reading comprehension, and verbal reasoning, which are the four sections on our assessment for students in grades 8 to 11 applying for admission to high school. We discuss our experiences engaging with partner organizations for research and thought leadership, facilitating student focus groups, and working with educators as content writers and reviewers. We share what was effective and what was less effective in our trials and retrials, and offer lessons learned to support other testing programs in understanding processes and resource needs "where the rubber meets the road" when moving toward a culturally sustaining assessment system.

Background

This section reviews a set of empirical studies on incorporating culturally relevant content development practices in different aspects of the assessment content development process, including item writing, item reviewing, quantitative and qualitative item analysis, and the impact of the resulting content on student perception, engagement, and performance. By outlining the current state of the field, including potential gaps, we describe the theory of action upon which our works were grounded, the purpose of our study, and its contributions to the ongoing discourse around operationalizing culturally relevant assessment.

Existing K-12 assessment content development research discusses details on developing culturally relevant content, with students' qualitative feedback from focus group interviews as evidence of validity. These studies often focus on a specific content area or related content areas when engaging with details of content development. Steedle et al. (2023) examined the feasibility of developing culturally relevant (CR) math and science items for the ACT, a high-stakes college admissions test. Their research aimed to "create items that represented unique cultural aspects, raised awareness of social justice issues, promoted cultural learning, and represented people in positive, non-stereotypical ways. (p.3)." Their sources of topics, or context ideas, were gathered via internal survey from a wide range of their testing organization's employees that were not item developers, and were grouped into broad categories such as politics, social change, and traditions. Their study design involved pairing each culturally relevant item with a conventional (non-CR) item that assessed the same content and used as similar wording as possible. This approach, which resembles an A/B testing design, allows the direct comparison of the impact on students of two different versions of an item: one with and one without culturally relevant context.

Lyons et al. (2022) conducted another proof-of-concept study on content development, focusing on the development of reading comprehension passages for Boston Public Schools. This study aimed to explore how justice-oriented assessments influence the student experience. Their passage topic selection process was different from that used in the ACT study: researchers initially identified preliminary topics they believed reflected students' everyday lived experiences and interests, particularly within the Boston context. Engaging local teachers in refining those passages proved crucial. Teachers, reflecting on their daily interactions with students, provided nuanced insights into passage selection, advocating for a more direct approach to themes of racism and social justice.

Both studies employed qualitative data collection methods, utilizing focus group interviews to gather students' perceptions and reactions to the culturally relevant content. The findings suggest potential benefits to integrating cultural contexts into assessments, as students expressed interest in seeing culturally relevant items on tests. Students also noted that this content was more engaging, facilitated learning about different cultures, and was relevant to their own lives. The main concern raised was the length and complexity of passages or items, and the subsequent need for more time to complete them. Some students indicated their preference for shorter, traditional items despite the perceived benefits of culturally relevant content, especially in a high-stakes setting. Notably, in the ACT study, students from underrepresented groups showed a greater preference for culturally relevant items and reported that they anticipated performing equally well or better on these items.

Qualitative findings on the impact of CR content are corroborated by emerging quantitative studies and psychometric evaluations of CR items. In a study examining student performance on a sixth-grade math assessment using an A/B test design with control and experimental versions of CR items, Laine & Schellman (2023) reported in their preliminary results that while culturally relevant items tend not to negatively impact student performance by student subgroups, there are non-negligible variations across items. Further studies analyzing a large volume of student responses to CR mathematics items confirm that CR content did not change items' difficulty and discrimination; the items' psychometric properties were largely preserved (Anguiano-Carraco et al., 2025; Valdivia & Steedle, 2025). Although students generally spent more time on the CR items (Valdivia & Steedle, 2025), they also demonstrated increased engagement compared to standard items (Anguiano-Carraco et al., 2025). These findings both align with the smaller-scale qualitative student interview results discussed in the empirical studies above. The most surprising result is that, while there is evidence that some math items with CR context exhibited reduced Black-White Differential Item Functioning (DIF) (Anguiano-Carraco et al, 2025), the inclusion of CR items did not impact score gaps between different groups in NAEP (Sinharay & Johnson, 2025). In other words, despite the perceived benefits of CR items, they did not significantly reduce group differences in student performance nationally.

This may be explained by the long-held belief in testing that the performance gap is perpetuated by inequitable access in the overall educational system. Simply changing the content in an assessment measuring the outcomes of learning cannot fundamentally eradicate group differences in performance. An alternative perspective on the persistence of the achievement gap in CR items is that a culturally relevant/sustaining/responsive assessment system extends beyond mere test content; it encompasses a comprehensive system of test design, construction, administration, reporting, and interpretation (AERA, APA, & NCME., 2014; Randall, 2021; Sireci, 2020; Walker et al., 2023). Culturally relevant practices should not be limited to the selection of topics and addition of cultural contexts to items, but should also be implemented in other aspects of the assessment, such as test administration and scoring. For example, Sireci (2020) challenged the notion of standardization in large scale assessment and suggested flexible standardization strategies that draw from a broad definition of culture to accommodate all student subgroups while maintaining validity. White et al. (2025) argued that scoring methods and psychometric analyses of student performance on constructed response tasks must incorporate ways to ensure that students' ideas and potentially multi-linguistic patterns of expression are valued. Gradual improvements in the overall CR assessment system in terms of practice and policy may eventually lead to a narrowing score gap among different student subgroups.

Considering both qualitative and quantitative data regarding CR assessment, a crucial question for any testing organization is: "What is the assessment's goal in achieving equity and fairness by incorporating CR content?" For our high-stakes admission assessment, the primary and immediate goal is not to close score gaps among subgroups or to offer deep cultural learning experiences within a timed test. Instead, our goal is to enhance student engagement and sense of belonging during the test, delivering a more positive testing experience without compromising validity and reliability.

Our organization's positionality emphasizes engaging stakeholders and rightsholders from beginning to end, redistributing power to students and teachers, particularly those from marginalized groups and "access organizations" (Center for Measurement Justice & Educational Records Bureau, 2023), which are non-profit educational organizations that aim to prepare socioeconomically disadvantaged students for competitive independent schools. We have implemented policies and recommendations across assessment processes, including offering flexible test administration times, locations, and modalities, support for students with accommodations and fee waivers to provide financial and test preparation support for students from traditionally under-resourced backgrounds. Content development and the subsequent item analysis are the largest remaining undertaking in moving toward a culturally relevant/sustaining/responsive assessment system. Scalability is the primary consideration, as this is the bulk of ongoing work in the annual development cycle.

To reconcile the inherent tension between large-scale standardized multiple-choice assessments and the goal of connecting test items to every student's culture and lived experiences (Evans, 2021), we adopt a broad definition of culture in our culturally sustaining content development. This includes a wide range of cultural representations beyond those of our specific test-taking population of students entering the independent school system. We use the phrase "culturally sustaining" (CS) deliberately to acknowledge that a "culturally sustaining" approach to content development places the rightsholders–those who are most impacted by the assessment–at the core of the design, development, and implementation of an assessment. Additionally, as Evans (2021) put it: "an assessment cannot be culturally sustaining without first being culturally sensitive, relevant and responsive". Though the culturally contextualized test content that we develop may fall somewhere along the stepladder continuum from culturally sensitive to relevant, responsive, and sustaining, we aspire ultimately to operate within a culturally sustaining assessment framework.

As we aim to elevate student assets within the framework of culturally sustaining assessment, we expand the definition of culture to encompass elements of youth culture, including digital nativity, engagement with social media and online communities, and involvement in social movements (Duncan-Andrade, 2010; Faverio & Sidoti, 2024; UNICEF Innocenti, 2024). The intersection of contemporary youth culture and cultural diversity distinguishes our approach from typical portrayals of diversity-infused items in the narrow sense of culture rooted in race and ethnicity, and away from which topics are deemed acceptable in traditional bias and sensitivity review. This better positions students as both producers and consumers of culture

(Evans, 2021), aligning with the goals of culturally sustaining practices by highlighting student voices in assessment.

Our content development decisions in the pilot project stem directly from our commitment to a student-centered approach. Students are not always the originators of content topics in existing empirical studies (e.g., Lyons, et al., 2022; Steedle, et al., 2003). When students did contribute to the topics as direct sources, such as in the work of Laine and Schellman (2023), the details regarding their engagement and the methods used to effectively solicit their ideas for content development are often not elaborated. Consequently, the practical utility of these approaches for scaling remains unclear. Our model aligns more closely with the approach adopted by Laine and Schellman (2023) who developed questionnaires and focus groups to prioritize the solicitation of student interests as a direct source of culture in the items. We aim to have students to be the primary source of cultural elements integrated into our content. This includes, and may even over-represent, students from access organizations.

Empirical research on culturally relevant education has linked culturally responsive pedagogies to student outcomes across various content areas (Aronson & Laughter, 2016). In contrast, existing studies on culturally relevant assessment often focus on processes for topic generation and item writing for specific subject areas. These studies provide valuable insights within those contexts, but often remain isolated and do not contribute to the development of a broader theoretical framework that connects culturally relevant assessment practices across multiple content areas. Addressing this gap requires not only accumulating concrete examples within individual disciplines, but also understanding how subject-specific knowledge and practices interact with culturally sustaining practices to shape the approaches used in content development.

Lastly, the empirical studies discussed above tend to focus solely on content development without referencing the full operational test development and test publishing cycle. As the body of qualitative and quantitative validity evidence grows rapidly in support of the benefits of incorporating culturally relevant content in assessments, there is a pressing need to bridge academic discourse and the realities of implementation. Theoretical discussion must be paired with practical guidance grounded in real-world challenges and solutions to advance the field collectively toward a more culturally sustaining assessment system. To translate research insights into actionable strategies for testing organizations, it is essential to develop a deep understanding of the content development processes, resource requirements, dynamics between internal and external stakeholders, and logistical considerations that underpin a successful implementation of CS content development at scale. Accordingly, this paper seeks to answer the following research question as our contribution to the field: Informed by our organizational positionality and the existing research on CR and CS assessment, what challenges and best practices can we identify through practical applications of CS content development?

Creating Culturally Sustaining Processes: A Trial and Retrial Approach

In this section, we describe the two inter-related rounds of culturally sustaining content development activities that we conducted, first in a standalone pilot, and then intersecting with our program's operational content development. The process of the initial pilot resembled the activities in the empirical studies discussed in the literature review, and was designed to provide proof of concept for the operational aspects and scalability of a CS content development workflow based on the guiding principles and research questions above. The specific goals of the initial pilot were: 1) Learn the process of creating these items; 2) produce a set of items across the four sections of our assessment: mathematics, quantitative reasoning, reading comprehension, and verbal reasoning; 3) collect qualitative and quantitative data on these items; and 4) make an informed decision with respect to the next steps towards an operational test development workflow.

With reflections on and lessons learned from the first pilot round, we oversaw another round of content development for incorporation in our live operational workflow for pretesting in scored exams. In this round we focused on the practical aspects of involving content development vendors while exploring methods of engaging teachers and students that may be more effective than those we used in the initial pilot. Scaling up brought challenges and improvements, with more stakeholders and new considerations for different content areas and grade levels. As we discuss our processes below, we also highlight modifications we made to our approach across two rounds of content development, and the protocols and deliverables we created to support the development process.

External Partnerships

Giving power to stakeholders beyond assessment professionals and centering the assets of Black, Latinx, and Indigenous students and teachers meant actively decentering ourselves. Staff on our testing program lack both expertise in culturally sustaining content development and lived experiences as Black, Latinx, or Indigenous people. We therefore sought to partner with consulting and research service organizations specializing in the intersection of justice-oriented theoretical frameworks and practical educational measurement. An organization with expertise in these areas was the most appropriate source of, first, the foundational justice-oriented testing principles, and then the culturally sustaining content development design.

Our initial pilot began by establishing guiding principles for culturally sustaining assessment across all aspects of assessment and all of the testing products in our organization's portfolio, defining expectations and appropriate stakeholder engagement from form construction through test implementation and score use. These guidelines, co-developed with an external partner, laid the foundation for moving toward a culturally sustaining assessment system at our organization and greatly informed the subsequent content development workflow design.

We then partnered with another educational measurement research organization to conduct a pilot round of content development and data collection. This was an opportunity to observe their researchers' expertise and learn about best practices in the processes unique to developing culturally sustaining assessment content. We then experimented with managing a round of culturally sustaining content development ourselves from start to finish, drawing directly from lessons learned during the pilot. We had three primary goals in this extension of our study: 1) to flesh out the quantitative and qualitative data collected during the pilot to support more robust results; 2) to adapt the pilot's CS content development processes to fit our organization's operational flow; and 3) to carry out the recommendations of our partner organization in experimenting with alternate approaches to some of the processes used in the pilot.

Recruiting and Engaging Stakeholders and Rightsholders

To achieve our goal of developing assessment items that are relevant, meaningful, and empowering for stakeholders and rightsholders, we engaged students and educators from historically marginalized backgrounds throughout our content development process. Recruitment was conducted through school administrators and professional contacts at institutions serving our target populations using a convenient and snowball sampling initially. We also leverage our existing relationships and partnership with individuals and organizations for specific data collection trials and retrials.

We faced challenges recruiting both teachers and students in the initial pilot, in part because many schools' spring breaks fell during our data collection window of March and April. We found that in recruiting students in particular, we were heavily dependent on the efforts of contacts within schools. In addition, Black, Latinx, and/or Indigenous teachers are underrepresented in the high school teacher workforce overall, especially in the subject area of mathematics (National Center for Education Statistics, 2020–21).

These issues slowed the pace of our recruitment, delaying when we reached a desirable volume of interested participants from different geographical locations.

After reflecting on this recruiting experience, we modified our outreach plan in the subsequent round of content development to identify strategies that could be more efficiently scaled up. To recruit educators with diverse backgrounds as item writers, we leveraged existing relationships with Black, Latinx, and/or Indigenous teachers who had previously consulted for our organization, asking them for referrals to any colleagues of theirs who would be a good fit for the project. We went on to work with one math teacher and one ELA teacher in the second round of content development, both of whom self-identified as Black. Similarly, for support in recruiting students for the second focus group, we sought Black, Latinx, and/or Indigenous teachers who also worked with students in afterschool activities. Educators who have built trusting relationships with students outside the classroom may be best positioned to recruit them for "extracurricular" projects like ours. With the help of the ELA educator we recruited to facilitate the second student focus group, for example, we were able to quickly recruit 4 students with minoritized backgrounds as participants. Table 1 below summarizes details of the participating rightsholders and stakeholders and the types of engagement through which they contributed.

Table 1. Participants' Details and Engagement Type

Engagement Type	Number & Role	Gender	Characteristics	Grade(s)	Location
Focus Group Round 1	5 students	NA	Black, Latinx	8, 10, 11	GA, CA, NJ, AZ
Cognitive Interview	4 students	2 Female, 2 Male	Latina, Southeast Asian, Asian, White	8, 9	NJ, NC
Open-ended Survey	64 students	NA	Fee waiver eligible	8	Nationwide
Focus Group Round 2	4 students	NA	Fee waiver eligible	8	NJ
Focus Group Round 2	1 educator	Female	Black	Middle School	NJ
Item Authoring Round 1	4 educators	2 Female, 2 Male	Black, Latinx	8, 9	NJ
Item Authoring Round 2	3 educators	2 Female, 1 male	Black, Latinx	Middle School	NJ, NY

Student Focus Group for Topic Selection

In the initial pilot, our interview protocol was developed to foster student autonomy with permission to "think big" in reimagining assessments (Lee, Taylor, Patterson, & Hazelwood-Cameron, 2024b). The primary focus was to solicit input on topics that are relevant to students' daily lives and intersectional identities, with the aim of using these topics as the basis of culturally sustaining items. We included a secondary focus of gathering a baseline of student impressions of current standardized test content, especially regarding the representation of their perspectives, experiences, and interests. Using the protocol, facilitators aimed first to build rapport by asking students about what interests them in their coursework. Students were then asked to connect to their experiences with standardized testing. Finally, students were asked to imagine themselves as test developers who can choose topics for test questions, with prompting to consider topics related to culture, communities, relationships, home life, interests, and views. They were invited to imagine what culturally sustaining test questions might look like.

Despite the intentional design of the focus group protocol and the skilled facilitators' efforts to establish rapport, the students contributed to the discussion only with a great deal of prompting, and the topic list collected was too brief and shallow to support an operational volume of content development. Just as we reflected on what did and did not work effectively and what would or would not be scalable as recruitment strategies, so did we design the second trial student focus group based on our experiences during and the outcomes of the first.

Researchers from the organization that partnered with us on the pilot suggested verbal games and visual stimuli as focus group strategies that might engage students in more fruitful discussion (Lee, Taylor, Patterson, & Hazelwood-Cameron, 2024a). We researched within the literature on focus groups and developed comprehensive facilitator onboarding materials to support co-design of a revised focus group format with facilitators themselves. The revised materials begin with a list of general strategies for effective focus group interviews (Guthrie, 2020), such as opening with warm-up activities. These guidelines are followed by a "menu" of facilitation strategies and details on their implementation from which a facilitator can choose.

We recruited one of the pilot's ELA item writers to facilitate the second student focus group, aiming to address the challenges we encountered during the pilot's focus group. She examined the materials that we provided, evaluated each strategy on the "menu" with written feedback, and then selected three strategies that she judged to be best suited to the students she recruited for the focus group: "Let's Bake a Cake," which uses visual and metaphorical devices as stimuli (Nind & Vinha, 2016), "Would You Rather (WYR)" (Simko, et al., 2021), and a semi-structured discussion. This educator also created specific prompts aligned with the strategies she chose, and organized the prompts using visual PowerPoint slides. She used that slide deck to facilitate the focus group.

The facilitator used "Let's Bake a Cake" as the opening activity. It used an extended visual metaphor of baking and decorating a cake to encourage creative thinking; students shared the "ingredients" that would make a test more relevant to their lived experience.

The facilitator introduced the WYR strategy as the next activity. To generate topics on culturally sustaining themes, the game included pairs such as "WYR move to a new country as an elementary school student or as a high school student?", "WYR perform in the Super Bowl halftime show or play football in the Super Bowl?", and "WYR be friends with [youth climate activist A] or [youth climate activist B]"? The facilitator developed deeper dive questions tied each pair to themes such as immigration experiences, personal abilities and interests, and local activism.

Semi-structured discussion was the final strategy. The questions prepared were adapted from our initial pilot and the work of Lyons et al (2022), and ranged in topic from how students might perceive connections to their real lives in standardized test content to whether students would be more likely to appreciate or skip over culturally sustaining context embedded in a math item.

We developed a template for the educator-turned-facilitator to share her post-focus group reflections and recommendations on each strategy, which will help us evaluate this approach to focus groups and design future groups more effectively. Table 2 below summarizes the differences and similarities between the two focus group designs.

Table 2. Comparison of Initial Focus Group and Revised Focus Group Design

	Initial Focus Group	Revised Focus Group
Role of Facilitator	● Facilitator was a researcher who identified as Black, Latinx, and/or Indigenous. ● Facilitator did not have prior relationships with students in the focus group.	● Facilitator was an educator who identified as Black, Latinx, and/or Indigenous. ● Facilitator had an existing relationship with students in the focus group.
Relationship among Students	● Students do not necessarily know each other.	● Students know each other well and have established socio-cultural norms from prior group interactions.
Focus Group Strategies	● Facilitator followed a semi-structured interview protocol to ask the group questions.	● Facilitator tried different focus group strategies in this order: ○ Visual metaphor: "Let's Bake a Cake" ○ Verbal game: "Would You Rather?" (WYR) ○ Semi-structured discussion based on prepared questions about standardized tests

Item Writing and Reviewing

The item development process was structured to maintain cultural relevance by hewing to the students' input in both the pilot and the subsequent round of development. We also centered historically marginalized stakeholders by working with Black, Latinx, and/or Indigenous educators with ELA or math subject matter expertise as item writers. The item writers were encouraged to draw on their own lived experiences and insight into student engagement to embed the student-generated culturally sustaining topic list seamlessly into valid, level-appropriate items. They applied their pedagogical expertise in crafting contexts based on the list as part of test questions that were authentic, engaging, and comparable in cognitive load to conventionally-developed test questions.

This student- and teacher-centered process produced 21 CS items across four content areas– mathematics, quantitative reasoning, reading comprehension, and verbal reasoning. After evaluating the qualitative and quantitative data collected on these piloted items and reflecting on the extensive revisions required to prepare them for field testing, we decided to make several changes to the item development workflow for use in the subsequent round feeding into our operational forms. Enhancing the item writer onboarding materials was one key change. We recognized the need for more structural support and practical guidance, in addition to culturally sustaining theories and concepts.

A key element of this structural support for item writers was providing "item shells" that were pre-leveled and aligned to our tests' standards, enabling item writers to work from a structured template rather than authoring items from scratch. We chose fully-developed items intended for field testing in our program's conventional development flow that seemed well-suited to culturally sustaining contexts. In practice, this meant that we chose math and quantitative reasoning items that lent themselves to a brief narrative and incorporated concrete elements such as shapes. We chose vocabulary items whose keyed terms lent themselves to a broad range of contexts. To create the item shells, we removed all aspects of the items except the keys (and, in the case of some of the math items, the distracters) and any elements of the stem or stimulus–such as graphics, labels, and the basic math problem formerly embedded in a word problem–that were critical in testing the intended skill. We included notes to the item writers on which aspects of the remaining content could or could not be changed, as well as the operations the item should or should not require students to perform. With the keys/options and basic elements of the stem and/or stimulus from the developed items retained, the result was a pre-aligned, pre-leveled item shell that the item writers could flesh out to create a culturally sustaining item. Table 3 below shows a detailed comparison of our two rounds of item development.

Table 3. *Comparison of Two Rounds of Item Development*

	Initial Round	Second Round
Onboarding	Item writers received: ● An overview of concepts within culturally sustaining pedagogy ● Technical documentation of our test's constructs and content standards for alignment ● Item samples ● A detailed explanation of the intended concept and skill measured in each item type. ● A list of topics generated by student participants in our earlier focus groups	Item writers received: ● An overview of concepts within culturally sustaining pedagogy ● An overview of item development and item quality best practices, based on issues observed in the items developed during the pilot ● "Item shells" that were pre-leveled and aligned to standards ● A detailed explanation of the intended concept and skill measured in that specific item shell. ● Lists of the invariant elements of the item shell and those variables that could be manipulated (the context, certain numbers, the order of conditions, etc.) ● A list of topics generated by students from other empirical studies, and linked resources for the topics ● Note: the reading comprehension writer received all of the above except for the item shells, since the items would draw from passages that had yet to be written.
Authoring Stage	● Item writers developed CS items from scratch to align to any one of the content standards, drawing on topics of their choice from the	● Item writers developed detailed culturally sustaining scenarios to frame the content in item shells.

	Initial Round	Second Round
	list generated by the student focus group.	● Note: the reading comprehension writer drafted passages and revised them based on assessment specialist feedback.
Review	● Item writers reviewed each other's work and provided written feedback on the cultural components embedded in the item. ● In-house assessment specialists reviewed and revised for concision, key security, level, and alignment for two to four rounds, conducting a fresh eye review of each other's work at the end.	● Item writers provided a written rationale on how the context they constructed is culturally sustaining, and to whom. ● In-house assessment specialists reviewed for fidelity to the item shell content, concision, key security, level, and alignment. ● External assessment specialists from the testing program's content development vendor reviewed for bias and sensitivity, editorial issues, and cueing/overlap within operational forms in which the items would be embedded for field testing.
Deliverables	21 culturally sustaining items: These included: ● Four vocabulary items, consisting of two one-blank items and two two-blank sentence completion items; ● Five quantitative reasoning items consisting of three word problems and two logical comparison items; ● Six reading comprehension items associated with a reading passage also authored by a teacher as part of the pilot; and ● Six mathematics items that covered the range of content domains measured by our test.	51 culturally sustaining items: These included: ● Four vocabulary items, consisting of three one-blank items and one two-blank sentence completion items; ● Four quantitative reasoning items consisting of three word problems and one logical comparison item; ● 40 reading comprehension items associated with two reading passages, all authored by a former independent school teacher who had also worked as an assessment specialist. 20 items per passage includes the amount of overage always developed in our program. ● Three mathematics items, each aligned to a different skill measured on our test.
Pretest Channel	● Third-party test preparation delivery platform	● Live operational test forms

Design of Qualitative Data Collection

To collect students' reactions and feedback on the CS items developed in the pilot, we designed a cognitive interview protocol for use in individual student think-aloud interviews. Each student interacted with five items, including at least one item from each of four domains to which the pilot items were aligned. After reading and choosing an answer for an item, narrating their problem-solving thought process aloud, the interviewee responded to open-ended questions such as (but not limited to):

- What do you generally like or not like about this question?
- What made this question easy or not easy to answer?
- Do you feel the context in the question makes it easier or harder for you to answer and demonstrate your skills in [content domain]?
- This question highlights [name the specific aspect of lived experience]. Have you seen that highlighted on a test before? How would you feel about seeing several questions on a test that highlight such contexts?

The wording of questions in the protocol is specific to the item's content area and the particular lived experience reflected in the item (generally either cultural heritage or youth culture), as we posited that students' perceived engagement and cognitive processes would vary based on the specific content and context types that they interacted with.

While the cognitive interviews allowed for deep analysis of a few students' engagement with the pilot items, we also wanted to collect a higher volume of basic student reactions. We administered a brief survey to a group of access organization summer program students–representing the rightsholders who the pilot was primarily intended to serve–who had encountered the pilot items as part of a practice test. Our survey questions presented students with pairs of items in a table following the A/B design we used elsewhere in the pilot: a CS math item paired with a conventional math item, and a CS verbal reasoning item paired with a conventional verbal reasoning item. The CS items were labeled "Type 1" while the conventional items were labeled "Type 2," to avoid activating any associations with the phrase "culturally sustaining." The survey itself consisted of two open-ended questions:

- In what ways do you like and/or dislike the Type 1 test questions in the left-hand column of the table? Please share specific details.
- In what ways do you like and/or dislike the Type 2 test questions in the right-hand column of the table? Please share specific details.

Both sets of qualitative results were used to inform the next round of content development and to provide validity evidence on the CS items.

Design of Quantitative Data Collection

To collect quantitative data on items and passages developed in the initial pilot, we field tested the items using an A/B test design. Our aim was to compare the performance of the culturally sustaining pilot items to the performance of items resembling them as closely as possible with the culturally sustaining contexts removed as proof of concept.

To accelerate this stage of the initial pilot, we conducted field testing on an online test preparation platform managed by an organization with which our organization has a longstanding partnership. The platform is widely used to practice for the standardized test that we manage, and as such, offers several advantages in the context of our pilot. A decent number of students nationally routinely interact with practice tests on the platform, obviating the need to recruit schools and students for standalone field testing events.

We were able to receive a substantial number of responses in a much shorter timeframe than our usual operational flow. The test preparation content on the platform was designed to resemble the same test as the culturally sustaining items developed in our pilot, allowing the pilot items to be seamlessly embedded in practice tests. Finally, our preexisting working relationship with the test preparation organization facilitated the rapid and seamless setup of forms and data transfers for the field testing.

In "Part A" of the A/B design, the pilot's 21 culturally sustaining items were embedded in existing practice tests for eighth to eleventh grade students. The items remained on the platform for 12 weeks. We also administered the practice test containing the CS items to a group of access organization students participating in a summer program administered by our partner organization, collecting additional data on responses from these students.

To prepare for "Part B" of the A/B design, in-house assessment specialists developed versions of the teacher-authored items that used the same keys and, to the degree possible, the same distracters, key support, and general item structure, while removing the context that was based on the student focus groups' and the item writers' input. See Appendix A, Table A1 for a paired example. In the same vein, an external consultant also familiar with our standardized test developed a version of the reading comprehension passage and items "scrubbed" of the themes generated by the student focus group. The in-house assessment specialists served as each other's and the consultant's reviewers in this process. The "Form B" items replaced the culturally sustaining items on the test preparation practice tests, and remained available for student responses for 20 weeks. The discrepancy between the time frames for which each set of items was exposed to students was based on the test preparation organization's available resources for managing data transfers.

We collected 1,259 student responses to the CS items on Form A, and 10,681 responses to the "scrubbed" versions on Form B. Our assessment is structured by sections, with one content area per section. The number of responses varied by section. Sections that appeared later in the test, such as mathematics and quantitative reasoning, generally had lower completion rates, resulting in smaller sample sizes for those content areas. Even after removing salient non-effortful results, however, the sample sizes for all sections across both forms remain sufficient.

Three main statistics were computed for each pilot item with and without a culturally sustaining context: proportion correct (*p*-value), point-biserial correlation, and average response time. Given the large sample sizes of both groups, we used independent samples t-tests to compare average response time on paired items, applying Welch's correction when variances were unequal. In addition, thanks to the access organization summer students' participation, we were able to conduct Differential Item Functioning (DIF) analysis on the CS items embedded in Form A using students' fee waiver status as a proxy for under-resourced subgroup membership. We used the Mantel-Haenszel (MH) procedure with purifications for DIF analysis (Holland & Thayer, 1988; Mantel & Haenszel, 1959) and interpreted the results using the ETS DIF classification system (Dorans & Holland, 1993), which categorizes items into three levels: 1) A-level: Negligible DIF; 2) B-level: Slight to moderate DIF; and 3) C-level: Large DIF. Items categorized as 'C' suggest a high level of differential functioning, raising concerns about potential bias or unfairness in the item's performance across these subgroups.

All of the statistics we computed mirrored some of the item analysis we conducted during our ongoing operational content development workflow. These quantitative results were used to inform our second round of content development pilot as well as our future psychometric evaluation practices of operational CS items.

Lessons Learned and Strategies for Operational Implementation

Build Facilitator Rapport into Student Focus Group Makeup

Based on our observations during the pilot's student focus group, our revised focus group design included aiming to work with students who already knew each other, and with a focus group facilitator who had a preexisting, trusting relationship with the students in the group. Our goal in situating the focus group within a network of established relationships was to help students feel more comfortable expressing their authentic interests.

A facilitator with existing rapport with a group of students is particularly important if the design relies on one focus group rather than repeated sessions. In a repeated focus group design, the facilitator may take time to build trust with the participants and gradually establish group norms such as the acceptability of respectful debate and building on others' ideas to encourage deeper thinking. In a design calling for a single focus group, it is beneficial to have established trust and group norms in place.

Our revised focus group formulation approach proved highly effective in fostering discussion. It also streamlined both recruitment and scheduling. Moving forward, we plan to adopt a similar strategy for our annual student focus groups, which will help us update our list of topics for CS content development. We will leverage our professional connections within assessment organizations and continue to prioritize focus groups with under-resourced students to elevate their voice within the work.

Use Non-Traditional Focus Group Strategies for Fruitful Discussion

Our analysis of focus group transcripts and the facilitator's written reflections show that, of the three facilitation strategies used, the Would You Rather (WYR) activity elicited the most eager student participation and was the most fruitful in terms of topic generation. Students engaged not only with the facilitator but also with each other as they explained their reasons for choosing the option that they selected in each pair. During the WYR game, students often participated in friendly debate with each other, which is the participation pattern the strategies were designed to foster and the pattern most likely to spark deep thinking beyond the prompts.

The level of abstraction involved in the "Let's Bake a Cake" visual metaphor proved to be challenging and may have been inhibiting to some–even older students such as 8th graders. Nonetheless, this activity served as a good warm-up to orient students to the focus group's general topic and to activate the group dynamic and norms. This echoes the general advice on having brief introductory warm-up activities for focus group interviews (Guthrie, 2020).

Finally, although the students engaged in the semi-structured discussion, they tended to share their feelings about the testing experience and their preferences for certain test formats, rather than topics and themes that they would want to see on tests. The semi-structured discussion strategy might be better suited to cognitive interviews, and for the purpose of soliciting student feedback on standardized testing in general or on individual items.

Tailor Focus Group Designs to Subject Area and Grade Level

As we analyzed the focus group transcript and the educator's written reflections, we noticed that her expertise as a humanities educator (as opposed to mathematics) may have shaped her approach to facilitating the focus group. As we debriefed with the educator on her interactions with the students, both parties came to the realization that she had likely deployed specific aspects of disciplinary thinking in creating visual stimuli for discussion as well as in composing follow-up questions. Furthermore, since the students perceived her as the ELA expert in the room, they may have offered examples rooted in ELA item types even when the questions were phrased generically to inquire about topics they wished to see across various test sections.

This interaction between the facilitator's content area of expertise, student perception, and focus group direction was evidenced by the semi-structured discussion at the end of the focus group. The ELA teacher's prompts asked students about math items. With this explicit reference to a subject area, students' conceptual image of a math item was activated; they offered their preference for context-free math items, contradicting the desire to see particular topics that they'd expressed previously. For example, students had mentioned earlier in the focus group that they were interested in current presidential policies as a general topic on an assessment. When students expressed their preference for context-free math items, there was an opportunity for the facilitator to illustrate a hypothetical math item using the topic students supplied–for example, a statistics item on calling a presidential race using sampling and predictions, or calculating the new price of an item based on tariff policies. Without math-related pedagogical knowledge, there may be missed opportunities to scaffold student thinking within a math content area; content experts may be best prepared to solicit deep thinking from students on CS topics and themes within that content domain.

On a related note, the facilitator also noted that the strategies' suitability is age-dependent. She selected focus group strategies based on her understanding of 8th-grade students, and she recommended that we apply any takeaways only to focus groups with students at that grade level. In general, lower grades may require more visual stimuli and more gamified activities to encourage meaningful engagement. For example, students could vote on the games or activities they would like to do during the focus group. This would require the facilitator to prepare in advance different versions of the prompts adapted for each strategy the students might choose.

In summary, the second focus group, while effective, highlighted important considerations for future scaling. We discovered that recruiting facilitators with preexisting relationships with students was a more effective way to both assemble groups of participants and to tailor the session to those students. Additionally, the facilitator's subject matter expertise proved beneficial for expanding on students' ideas and fostering deeper discussions. Therefore, we recommend tailoring a focus group design to specific subject areas by framing the discussion with explicit references to a subject area and/or item types. In practice, this might look like an ELA teacher and a math teacher facilitating separate focus groups in parallel. Similarly, focus group designs should be adapted for different grade levels. Finally, facilitators should be trained with concrete examples of soliciting and building on students' ideas. Taken together, these measures will support tailoring focus group strategies to different contexts.

Provide Guidance and Scaffolding for Item Authoring

One of the principles in this work is to decenter assessment specialists and embrace direction from stakeholders and rightsholders such as students and teachers to shape the content. During the initial pilot, we saw that the content produced by teachers with no formal assessment development experience was especially raw. Assessment specialists' expertise in identifying salient testing points, fitting item type to testing point, aligning items to standards, eliminating overlap and cueing within a set, ensuring key security, and writing with/editing for concision and clarity all proved to be indispensable complements to student and teacher insight in preparing content for a high-stakes standardized test.

Therefore, we aimed to provide more structured support for the teachers as item writers in the second round of content development. We developed instructions for content development based on our analysis of the qualitative and quantitative data collected in the pilot. These included guidance on clarity and concision, a preference for narratives rather than informational texts, and options that are distinct from each other and from the stem. Perhaps the most differentiating aspect of the second round of content development was that, instead of asking teachers to take on the role of an assessment specialist by developing items from scratch, we sought to "bake" leveling and alignment to standards into the content from the beginning, so that the item writers could focus on contributing via their expertise in student engagement and the pedagogical applications of real-life scenarios.

We did this by creating what we came to call "item shells" as described in the Item Writing and Reviewing section above. The scaffolding of the item shells obviated the need for extensive revisions that we saw during the pilot, leading to a process efficient enough for operational content development. In an operationalized culturally sustaining content development flow, assessment specialists would create item shells for this purpose rather than retroactively turning fully-developed field test items into shells.

The assessment specialists reviewing and revising the pilot items were also given specific instructions. They were urged to consciously separate the aspects of item development best practices outlined above from the conventional bias and fairness guidelines in which they had also been trained and had applied in their work for many years. Under the principle that standardized test items have always contained (invisible to some) framing irrelevant to the construct, and that the culturally sustaining contexts in the pilot items simply contained framing that was more visible to content developers in that they represented lived experiences other than those of members of the dominant culture, assessment specialists' and editors' revisions are instructed to retain as much as possible the context and details supplied by the item writers. Nonetheless, heavy manipulation of the content beyond the authoring stage risks compromising the culturally sustaining elements–so, with the item shells and instructions to assessment specialist reviewers, we aimed to minimize in-house revisions.

Diversify Item Writers and Reviewers as Organizational Strategy

Researchers advocating for culturally sustaining assessment frequently recommend broadening the diversity of item writer pools (e.g., Lyons et al., 2021). In establishing a culturally sustaining content development workflow, we found that recruiting as large and diverse a group of item writers as possible would be valuable. Culturally sustaining item writers are encouraged to bring their personal lens into the work–because of this, specific perspectives may be overrepresented in the content produced by a smaller group of item writers.

Diversifying the item writer pool and also the item reviewer pool, however, is a long-term effort at the organizational level. For testing organizations that depend on vendors for content development, the asynchronous nature of different organizational strategies may at times create misalignment on priorities and resource allocation. Our current strategy involves engaging educators from Black, Latinx, and/or Indigenous communities as item writers to harness their funds of knowledge and their insights on their students. This approach not only allows us to foster deeper and more meaningful engagement with stakeholders, but also enables us to depart temporarily from our existing content development stream without overhauling the entire process in a short time frame. We discovered that it was more effective not to expect teachers to be assessment specialists knowledgeable in all item writing best practices (e.g., cueing, key security), but rather to support them in focusing on creating authentic narrative contexts that would resonate with and engage students.

This strategy may need to be reconsidered in the future. Some may envision a future assessment practitioner workforce in which specialists bring both the technical expertise of conventional content development and a deep understanding of culturally sustaining perspectives and students' everyday lived experiences. In our second round of culturally sustaining content development, we worked with a reading comprehension content author with expertise both as an in-house assessment specialist and as a high school ELA teacher, whose authoring drew upon her own lived experiences as a Latina and first-generation American. This content developer's work underscores the profound impact that a diverse workforce, rich in varied perspectives and experiences, can have on shaping more equitable and impactful assessment content development.

Pilot Item Statistics and Implications for Practice

Using student responses from the pilot items on Form A and Form B, we calculated statistics for both the CS items and their counterparts that had been "scrubbed" of culturally sustaining elements to establish a baseline evaluation of the CS items' performance. Table 4 and Table 5 summarize the statistics from the paired items in all four content areas.

Using the same parameters that we do in our operational item analysis for classical item statistics, we found that item difficulty (*p*-value) and item discrimination (point-biserial correlation) were within the acceptable range for all of the piloted CS items. This suggests that the teachers who served as item writers successfully targeted their items to the appropriate grade range. The average response time was also acceptable across all content domains.

When comparing the statistics of CS items to their non-CS counterparts, we observed that the *p*-value and point-biserial statistics were largely consistent within each pair of items. For the few pairs that showed a disparity between the CS and non-CS versions, the results were mixed. Regarding average response time, although the time spent on the CS items was within the threshold range of 60 seconds per item, students tended to spend significantly more time on 70% of the CS versions. One exception is the reading comprehension passage set with six items. Despite slightly more time spent on the passage and item 1 (the passage and the first item are displayed side-by-side in a split-screen view), students on average spent less time answering most of the remaining questions on the culturally sustaining passage. Furthermore, DIF analysis conducted using fee waiver registrants as a focal group also flagged five out of the 21 piloted items as exhibiting C DIF, requiring additional review and potential removal from the pool based on our operational business rules for DIF.

Our program generally loses about 10% of field-tested operational items due to failure to fall within all parameters; C DIF across all subgroups of gender, race, and ethnicity make up only a very small percentage of those flagged items. In contrast, a much higher percentage of the pilot items would have to be reviewed, re-field-tested, and/or rejected based on their psychometric performance. The bulk of the items that demonstrated C DIF were within the math and quantitative reasoning sections. This poses considerations for operational planning; we must anticipate CS items having higher failure rate, which would require developing additional overage at earlier stages.

The CS items' tendency towards more flagging for poor statistics also suggested lines of inquiry into item quality. We have several hypotheses about these statistics that informed our second round of item development. First, mathematics may be one area where the pilot's tight content development timeline and the limited involvement of professional assessment specialists and editors most strongly impacted item performance. Second, the culturally sustaining context of some of the items may have been too detailed and therefore carried a high cognitive load. This may have affected math and quantitative reasoning in particular, as most students do not expect to have to read a great deal to answer questions in those domains and lack motivation to do so.

These two hypotheses led us to adopt the approach described above in our second round of item development: we provided item writers with item shells that were pre-leveled and aligned to standards. Furthermore, we paid particular attention to the word count in math and quantitative reasoning items. This involved at times extensive editorial review for concision while maintaining the narrative intended by the item writers and alignment to the mathematics concepts and skills needed to solve the problems. Lastly, we began to explore the logistics involved in changing the item types in our assessment design. Currently, the mathematics and quantitative reasoning sections consist solely of discrete multiple-choice items. Integrating item sets–comprising two to three multiple choice questions linked to a brief stimulus with a real-world, culturally sustaining context—could reduce the cognitive load per item while achieving the goal of a culturally sustaining assessment. This approach would also maintain alignment with the test sections' emphasis on problem-solving through reasoning within a real-world context. Such item sets hold significant promise for culturally sustaining high-quality mathematics assessment content, provided the stimulus's context is relevant to both students' lived experiences and to the necessary mathematical concepts and skills.

Table 4. Test Statistics and C DIF Flags for Paired CS Items and Their Non-CS Counterparts

	Item Difficulty (P-Value)[a]		Item Discrimination (Point-Biserial Correlation)[b]		C DIF Flag[c]
Item Type	CS	Non-CS	CS	Non-CS	CS
MA 1	0.61	0.55	0.48	0.47	C+
MA 2	0.46	0.44	0.43	0.37	C-
MA 3	0.68	0.59	0.34	0.41	C+
MA 4	0.91	0.90	**0.29**	**0.34**	No C DIF
MA 5	0.37	0.37	0.34	0.34	No C DIF
MA 6	0.31	0.26	0.37	0.37	No C DIF
QR 1	0.60	0.62	0.48	0.43	No C DIF
QR 2	**0.52**	**0.69**	0.43	0.44	No C DIF
QR 3	0.45	0.39	0.46	0.44	No C DIF
QR 4	0.70	0.67	0.37	0.39	C+
QR 5	0.46	0.43	**0.38**	**0.29**	No C DIF
RC 1	0.94	0.94	0.42	0.41	No C DIF
RC 2	0.76	0.74	0.40	0.42	No C DIF
RC 3	0.80	0.82	0.47	0.49	No C DIF
RC 4	**0.44**	**0.25**	**0.32**	**0.23**	No C DIF
RC 5	**0.51**	**0.40**	0.29	0.26	No C DIF
RC 6	0.91	0.92	**0.31**	**0.24**	No C DIF
VR 1	0.85	0.83	0.50	0.44	C+
VR 2	**0.68**	**0.57**	0.49	0.48	No C DIF
VR 3	0.62	0.61	0.33	0.34	No C DIF
VR 4	0.39	0.38	0.43	0.46	No C DIF

Notes: [a]Item Difficulty (P-Values): Paired items with a disparity between them greater than 0.1 are highlighted in bold.

[b]Item Discrimination (Point-Biserial Correlation): As a general rule of thumb (Ebel & Frisbie, 1991), point-biserial correlation values below 0.3 suggest either marginal (between 0.20 to 0.29) or poor discrimination (<0.2). Paired items with a disparity between them within the range of good and marginal discrimination are highlighted in bold.

[c]C DIF Flag indicates the result of a DIF analysis for the fee-waiver group (focal group) using MH approach and ETS' classification of DIF. "C+" indicates the item demonstrates C DIF, and was differentially easier for the fee waiver group, and "C-" indicates the item was differentially more difficult for the fee waiver group.

Table 5. Independent Samples t-Tests Comparing Time Spent (in Seconds) on Paired CS Items and Their Non-CS Counterparts

	CS Version		Non-CS Version		
Item	Mean	SD	Mean	SD	t
MA 1	52.7	35.6	52.9	36.9	-0.10
MA 2	70.2	55.1	58.3	45.1	4.79**
MA 3	80.9	54.5	81.6	54.4	-0.27
MA 4	28.5	21.9	26.8	23.4	1.59
MA 5	48.8	35.6	42.7	31.7	3.74**
MA 6	65.9	61.7	57.3	49.2	3.09*
QR 1	72.2	54.9	63.4	45.4	4.01**
QR 2	88.6	69.5	67.9	45.7	7.53**
QR 3	95.8	74.5	91.7	68.5	1.36
QR 4	71.6	52.7	66.0	45.2	2.67*
QR 5	65.3	61.8	60.0	49.5	2.14*
RC 1	131.6	94.1	122.5	86.3	2.42*
RC 2	41.6	34.6	45.8	34.8	-2.94
RC 3	50.3	43.0	23.1	20.9	16.47**
RC 4	44.5	32.0	43.0	29.2	1.22
RC 5	34.8	29.8	42.4	35.6	-6.10**
RC 6	22.1	28.3	28.4	22.7	-5.65**
VR 1	22.7	18.4	20.5	16.8	3.88**
VR 2	32.0	24.0	27.2	21.1	6.46**
VR 3	40.4	32.4	39.1	26.6	1.51
VR 4	38.9	26.2	36.7	27.2	2.56*

*Note: * $p<0.05$, ** $p<0.001$.*

Reconceptualize DIF Analysis Within Our Current Practices

Educational measurement researchers emphasize the importance of collecting racial, ethnic, and gender demographic data to ensure fairness as part of an assessment's validity program ("Reconsidering Assessment Fairness," 2024). Comparing results across various demographic subgroups can guide test developers in preventing potential bias in assessment content. A common approach to using racial, ethnic, and gender demographic information to prevent bias during content development is through DIF analysis on items. This analysis helps ensure that assessment items do not unintentionally favor specific subgroups, particularly those from minoritized populations.

As an admission assessment, we give all registrants for our test the option to share demographic information pertaining to their gender, race and ethnicity. Disclosing this information is strictly voluntary so as to avoid evoking any experience of stereotype threat that may impact registrant's test performance (e.g., Spencer, Steel, & Quinn, 1999). The optional nature of these demographic fields on the registration page, however, poses challenges in DIF analysis using subgroup categories. About one third of registrants for our test decline to report their racial demographic information.

As part of the pilot, as well as an ongoing larger initiative to reconceptualize DIF analysis using meaningful subgroups to strengthen the validity of our assessments (Hu, 2025), we experimented with using students' need-based fee waiver status to define the groups in DIF analysis within the pilot as well as in our operational test data analysis. Using fee waiver eligibility in DIF analysis presents several key advantages. Firstly, fee waiver status offers a more direct and accurate reflection of socioeconomic status (SES) and the concept of underserved groups within the specific population of our test takers. Unlike other potential indicators which may correlate with but not directly signify financial need, fee waiver programs are specifically designed to identify students whose families meet defined financial thresholds. This direct link enhances the validity of fee waiver data as a proxy for SES to examine differential group performance, if any.

Secondly, data on fee waiver eligibility is comprehensive, with virtually no missing information. Because registrants who use a fee waiver enter a voucher code at the point of registration, our organization possesses a complete dataset for all applicants who seek this form of financial assistance. This absence of missing data ensures a more robust and reliable basis for analysis compared to other data points that may suffer from incomplete reporting.

Supplemental DIF analysis using fee waiver status in addition to traditional gender, racial, and ethnic demographic information may allow testing organizations to gather data on CS content quickly, since fee waiver status information tends to be more readily available to the testing organization. Our partnership with a third-party test preparation provider offered a significantly shorter turnaround time for establishing some, albeit not comprehensive, validity evidence for the fairness of the CS items using in-house fee waiver data. Those initial DIF analyses on our piloted items then informed our reflections on item development and review processes, leading to revisions in our onboarding approach for item writers before deployment to operation. For CS items from the second round of content development, which were put onto operational forms for live pretesting, DIF analysis using fee waiver eligibility as a subgroup has been incorporated into our routine psychometric item analysis. This allows us to evaluate student performance on more CS content, accumulating validity evidence for our program.

Use Qualitative Data from Rightsholders to Supplement Quantitative Data

The pilot's qualitative data, both supported and added nuance to its quantitative data, which we took as further evidence in favor of continuing to explore culturally sustaining content development. In particular, the access organization students' survey responses spoke directly to the validity issues suggested by the math items' tendency towards poor statistics. Students reported higher engagement with items using real-life, culturally sustaining contexts; several noted, however, that the wordier contexts stood out as irrelevant to the skills measured in math items. Another actionable insight from the survey responses is that the students felt narrative-style contexts were better vehicles for culturally sustaining topics than those resembling an informational text. Finally, in a finding that also speaks to item quality concerns, student respondents noted that more straightforward syntax and better differentiation between answer choices would have made the items more effective.

Results from the pilot's cognitive interviews revealed a great deal of consistency between the interviewees' interactions with the CS items and those of the survey respondents (Lee, Taylor, Patterson, &

Hazelwood-Cameron, 2024b). We took this as an indication that the students' collective feedback is reliable as guidance in future development efforts. The interviewees expressed greater engagement with items that were personally interesting to them, including contexts such as online life, hobbies, and cultural heritage; in fact, the theme of engagement with personally relevant topics was the one most strongly represented in the interview transcripts, brought up by each student and in relation to each content domain. The interviewees also noted increased engagement when encountering less sanitized versions of history on a test, including references to racism.

The students noted that relatability was important in positioning the culturally sustaining contexts as helping rather than hindering them in demonstrating the measured skills. Nonetheless, as in the survey responses, the interviewees shared that they were more likely to ignore the culturally sustaining context or even feel negatively towards it if they felt that the context was incongruent with the skill being measured. This came up in relation to math items in particular, and this corroborates our recommendations above on the need for an item's context to be relevant not just to students' lived experiences but also to the content domain.

Overall, we highly value this qualitative student feedback, which we believe complements psychometric evaluations of item quality. To calibrate our understanding of CS content development and keep students' voice at the forefront of the work, we have developed plans to incorporate periodic student cognitive interviews on sampled CS items into our operational content development flow.

Design A Comprehensive Content Development Workflow

Considering how to apply our lessons learned in our program's operational content development, we designed the following workflow. Figure 1 outlines the basic steps in both a conventional and culturally sustaining content development workflow, and the points at which our experiences suggest the culturally sustaining content workflow might depart from and then rejoin the conventional content workflow.

The initial steps in the culturally sustaining stream establish the input of students and teachers from historically marginalized communities as the foundation of the content. The process begins with one or more student focus groups, including an overrepresentation of students attending access organizations, if possible, to develop a list of topics and themes that would offer them affirmation during the testing experience. The results of these series of focus groups become the basis of both culturally sustaining passage topic selection/passage authoring and item writing for the different content areas. The authoring stage is a chance for more granular ideation based on the student focus group output, carried out by education professionals who are themselves members of minoritized demographic groups and who work directly with students. Finally, after tweaks by assessment specialists as needed for alignment to standards, length, level, house voice, and so forth, a final authenticity review by a group similar to the item authors will ensure that the culturally sustaining perspective remains. We also plan to supplement our routine psychometric evaluation of items with student cognitive interviews on sampled CS items to gather additional qualitative data to monitor the item quality.

There are specific details regarding the authoring stage for different content areas. Figure 2 zooms in for a detailed view of the authoring stage where item shells are provided for item writers to develop items for mathematics, quantitative reasoning and verbal reasoning. The authenticity review also happened at this stage for item reviews on top of the passage review.

Overall, this comprehensive workflow demonstrates how culturally sustaining content development can be thoughtfully integrated into our long-established annual content development cycle. By strategically planning points of divergence and convergence with the conventional workflow, we are able to address the need for culturally sustaining assessment while largely maintaining the existing operational calendar, and managing dependencies on vendors, costs, timelines, and other critical resources.

Figure 1. Assessment Content Development Flow with Adjoining Culturally Sustaining Content Development Stream

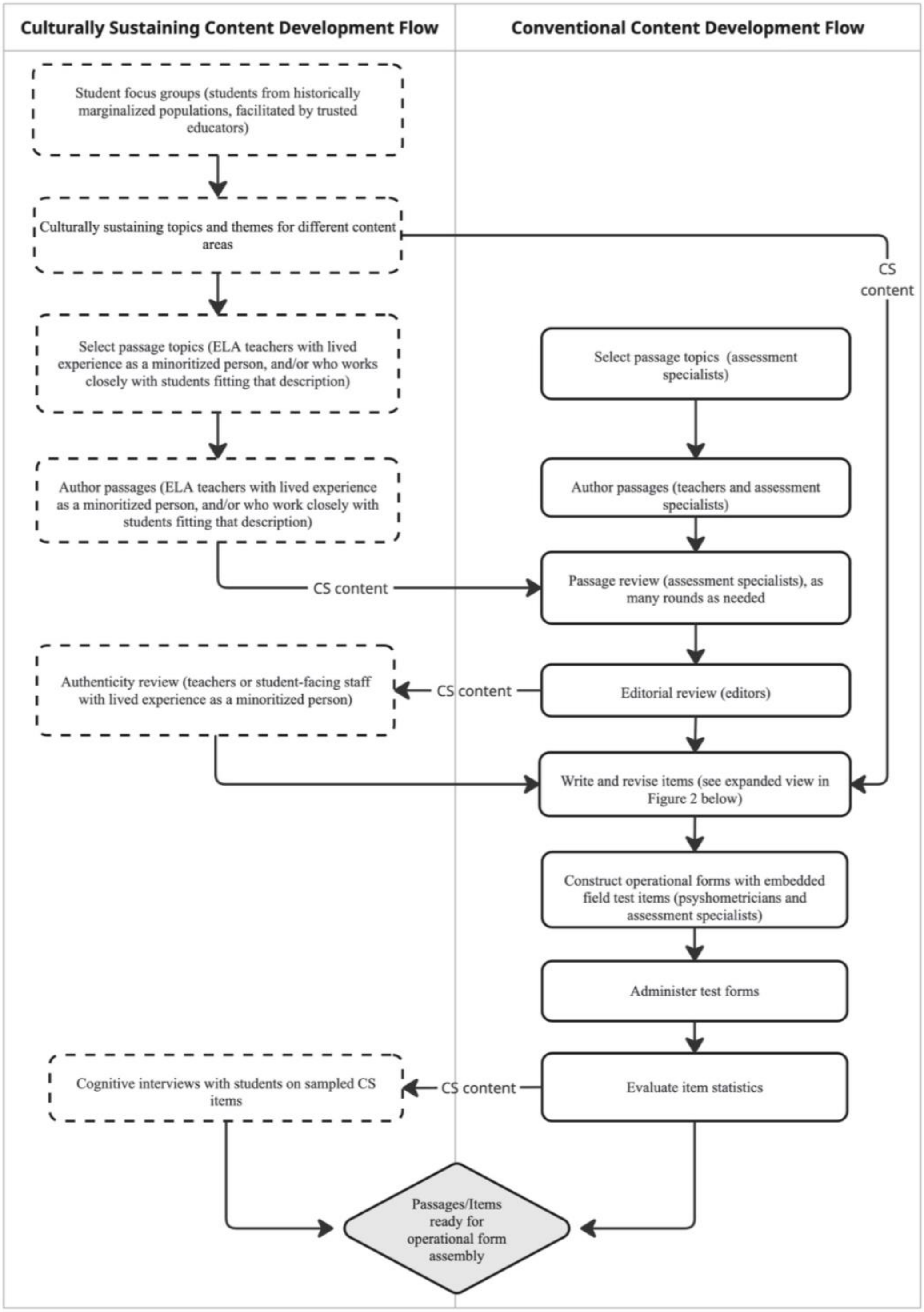
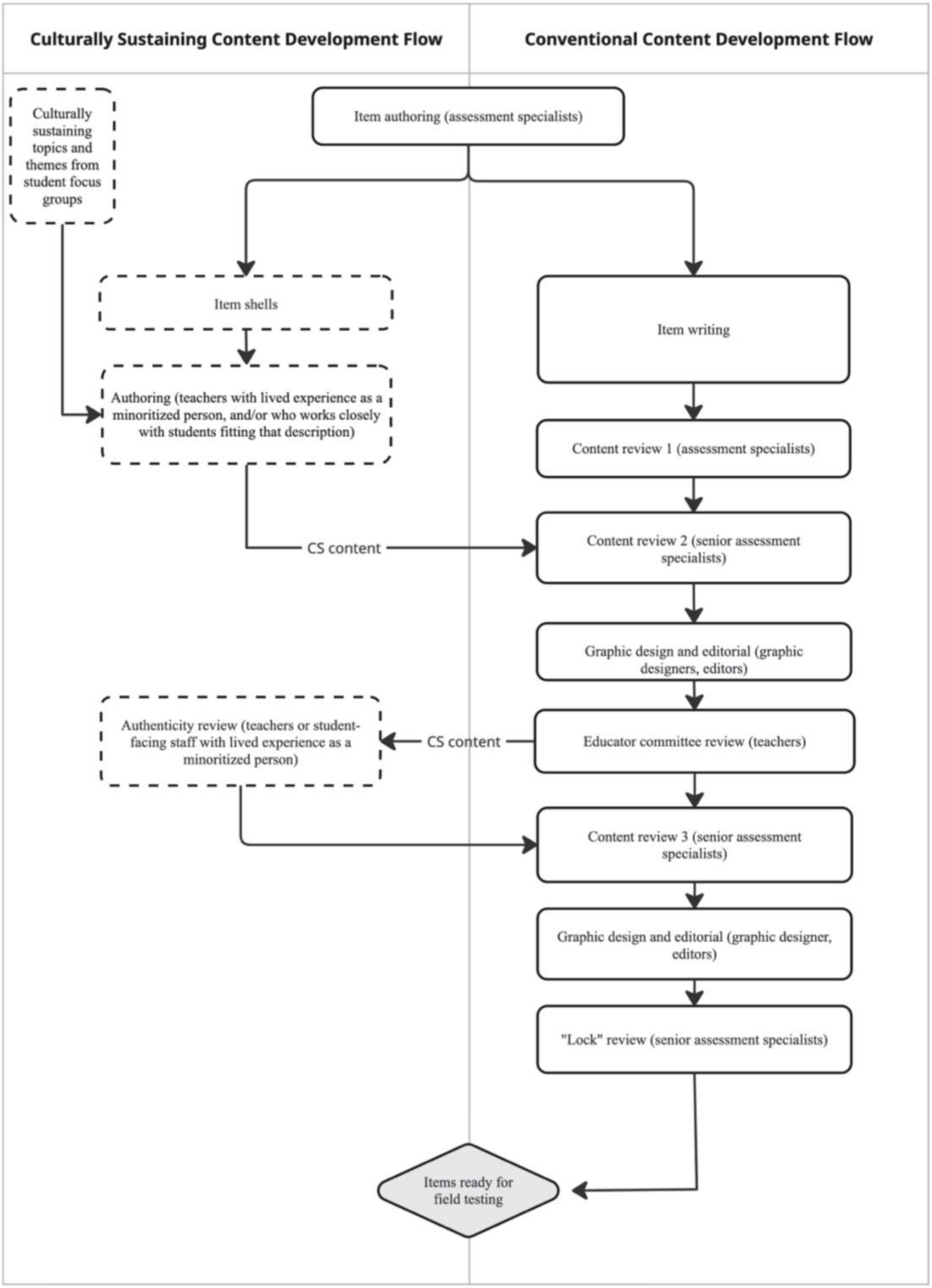

Culturally Sustaining Content Development Flow
Conventional Content Development Flow
Student focus groups (students from historically marginalized populations, facilitated by trusted educators)
Culturally sustaining topics and themes for different content areas
CS content
Select passage topics (ELA teachers with lived experience as a minoritized person, and/or who works closely with students fitting that description)
Select passage topics (assessment specialists)
Author passages (ELA teachers with lived experience as a minoritized person, and/or who work closely with students fitting that description)
Author passages (teachers and assessment specialists)
CS content
Passage review (assessment specialists), as many rounds as needed
Authenticity review (teachers or student-facing staff with lived experience as a minoritized person)
CS content
Editorial review (editors)
Write and revise items (see expanded view in Figure 2 below)
Construct operational forms with embedded field test items (psyshometricians and assessment specialists)
Administer test forms
Cognitive interviews with students on sampled CS items
CS content
Evaluate item statistics
Passages/Items ready for operational form assembly

Figure 2. Write and Revise items: Expanded View of Workflows


Culturally Sustaining Content Development Flow
Conventional Content Development Flow
Culturally sustaining topics and themes from student focus groups
Item authoring (assessment specialists)
Item shells
Item writing
Authoring (teachers with lived experience as a minoritized person, and/or who works closely with students fitting that description)
Content review 1 (assessment specialists)
CS content
Content review 2 (senior assessment specialists)
Graphic design and editorial (graphic designers, editors)
Authenticity review (teachers or student-facing staff with lived experience as a minoritized person)
CS content
Educator committee review (teachers)
Content review 3 (senior assessment specialists)
Graphic design and editorial (graphic designer, editors)
"Lock" review (senior assessment specialists)
Items ready for field testing

Discussion and Limitations

Our approach to developing culturally sustaining assessment content was shaped by both the resources available to our organization and the considerations of integrating a new content development stream into an established, ongoing content development workflow. The primary goal of this paper is to describe our trial and retrial of processes based on empirical evidence, knowledge from the field, and lessons learned as we experimented. We grounded our proposed best practices on these considerations as well as real-world parameters. Accordingly, several limitations should be noted that are inherent to our proof-of-concept focus.

First, a comprehensive psychometric evaluation of item performance and student performance requires an extended timeline and a large volume of test-taker data. Our initial validity evidence is necessarily limited: it does not include full-scale psychometric analyses of CS item performance in live, operational settings. While the pilot items were field tested on our partner organization's test preparation platform, which was designed to approximate an authentic testing experience, these data cannot fully capture the dynamics of actual test administrations nor represent the student experience during a high-stakes admission test. This is evidenced in the quantitative data; some students finished only some sections of the practice test, resulting in the exclusion of results from discarded sessions as well as from non-effortful test-taking behaviors. In addition, we analyzed only basic classical item statistics for the piloted items; this is the initial step in our operational item analysis, used to discard items that fall outside of certain parameters before further evaluating the remaining items. Item statistics from the Rasch model under the Item Response Theory framework determine the actual item estimates, including the difficulty and differentiation statistics that we use to assemble the test forms based on targets. The CS items developed in the second round are currently being field tested on operational forms, and future analyses will be needed to assess their psychometric properties in operational contexts.

Another related limitation is the strength of our psychometric evidence for proof of concept, particularly our DIF calculations, which is limited by the absence of student gender and racial demographic information. We used fee waiver status–that is, whether or not the student registered with a need-based fee waiver–as a DIF focal group serving as a proxy for socioeconomic disadvantage. This also had limited efficacy due to low numbers in this group of interest. Our internal evaluation (Hu, 2025) indicates that while using fee waiver groups for item analysis is more time-efficient for gathering sufficient responses for DIF analysis compared to certain other minority groups among our test-takers (including Black, Hispanic, and Indigenous students), students registering with fee waivers constitute an average of only 12% of our annual test-takers. This inherent imbalance in the distribution of reference and focal groups necessitates a long period of accumulation to reach the threshold numbers for sample sizes in operational DIF analysis.

Determining the ideal proportion of CS items within an assessment is another empirical question that extends beyond the scope of our initial development effort. We have proposed an initial range of 10% to 20% CS items on each section of our test (that is, 3 to 10 items). However, it takes time to develop and pretest a sufficient number of items given the resources involved in CS content development, limited slots on forms available for pretesting, and the timeline of collecting sufficient pretest responses. Determining the ideal proportion also requires further research to understand the impact of varying proportions of CS items on testing time, accommodation policies, and other factors related to test validity as well as the student experience.

Finally, our CS content development focused on only one level of our assessment: the upper level, targeting students in grades 8-11. This is based on both the assumption that older students would be best able to express themselves freely in focus groups and cognitive interviews, and because this grade range represents the highest volume of students who take our test. The applicability of our development process and findings to younger grades remains to be explored and may present distinct challenges or opportunities. For example, lessons learned from our student focus group suggest that engaging younger students effectively may require different interview strategies. Additionally, students of varying age groups interact with cultural contexts differently and may be drawn to different topics than older students, given their distinct stages of social, cognitive, and affective development. These developmental differences necessitate specific considerations in the item development process. Item writers must adapt cultural contexts to be relevant to the particular age group of students and their everyday lived experiences. Exploring adjustments to the content development process in the lower and middle levels of our assessment is an ongoing experiment, but beyond the discussion of this paper.

Received: 7/7/2025. **Accepted:** 9/27/2025. **Published:** 12/2/2025.

Citation: Hu, S., & Meyer, R. (2025). Culturally sustaining content development in standardized testing: Practical considerations and lessons learned. *Practical Assessment, Research, & Evaluation*, *30*(2), Article 3. Available online: https://doi.org/10.7275/pare.3281

Corresponding Author: Sihua Hu, Educational Records Bureau. Email: ahu@erblearn.org

References

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

Anguiano-Carrasco, C., McVey, J., & Steedle, J. (2025, April). Psychometric evaluation of culturally relevant items in a college admissions test. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Aronson, B., & Laughter, J. (2016). The theory and practice of culturally relevant education: A synthesis of research across content areas. *Review of Educational Research, 86(1), 163–206.* https://doi.org/10.3102/0034654315582066.

Bennett, R. E. (2023). Toward a theory of socioculturally responsive assessment. *Educational Assessment, 1–22.*

Center for Measurement Justice., & Educational Records Bureau. (2023). *ERB's Principles for the Development of Justice-Oriented Assessments* [Unpublished internal report]. Educational Records Bureau.

Dixon-Román, E. J. (2020). A haunting logic of psychometrics: Toward the speculative and indeterminacy of Blackness in measurement. *Educational Measurement: Issues and Practice, 39(3), 94–96*.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning (pp. 35–66).* Lawrence Erlbaum Associates.

Duncan-Andrade, J. M. R. (2004). Your Best Friend or Your Worst Enemy: Youth Popular Culture, Pedagogy, and Curriculum in Urban Classrooms. *Review of Education, Pedagogy, and Cultural Studies, 26(4), 313–337.*

Educational Records Bureau. (2023). *The 2023 Admission Survey and ERB's Next Generation Admission Initiative* [Report]. Educational Records Bureau. https://cdn.erblearn.org/www/20231114_ERB_ISEE_NextGen_Executive-Summary-Survey.pdf

Evans, C. (2021, November). Culturally sensitive, relevant, responsive and sustaining assessment. CenterLine Blog. *https://www.nciea.org/blog/culturally-sensitive-relevant-responsive-andsustaining-assessment/*.

Faverio, M. & Sidoti, O. (2024, December). Teens, Social Media and Technology 2024. Pew Research Center. Retrieved from: *www.pewresearch.org/internet/2024/12/12/teens-social-media-and-technology-2024/*.

González, N., Moll, L. C., & Amanti, C. (Eds.). (2005). *Funds of knowledge: Theorizing practices in households, communities, and classrooms*. Lawrence Erlbaum Associates.

Guthrie, K. H. (2020). Qualitative inquiry with adolescents: Strategies for fostering rich meaning making in group interviews. *American Journal of Qualitative Research, 4(3), 92–110. https://doi.org/10.29333/ajqr/8586*.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity (pp. 129–145)*. Lawrence Erlbaum Associates.

Laine, B., & Schellman, M. A. (2023, April). Experimental examination of the impact of culturally responsive assessment practices. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Chicago, IL.

Lee, J. C., Taylor, D., Patterson, C., & Hazelwood-Cameron, A. (2024). *A culturally sustaining item development process* [Unpublished internal report]. JCRG USA Inc.

Lee, J. C., Taylor, D., Patterson, C., & Hazelwood-Cameron, A. (2024). *Culturally sustaining item development in high stakes admissions testing: Learning from the Independent School Entrance Exam Pilot Study* [Unpublished internal report]. JCRG USA Inc.

Lyons, S., Hinds, B. F., Denker, H., & Student, S. (2022). Impact of a Justice-Oriented Assessment on Student Experiences [Research report]. Lyons Assessment Consulting.

Lyons, S., Johnson, M., & Hinds, F. (2021). A call to action: Confronting Inequity in Assessment. Lyons Assessment Consulting.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22(4), 719–748*.

Nind, M., & Vinha, H. (2016). Creative interactions with data: Using visual and metaphorical devices in repeated focus groups. *Qualitative Research, 16(1), 9–26*. https://doi.org/10.1177/1468794114557993.

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.

Reconsidering assessment fairness: Extending beyond the 2014 Standards for Educational and Psychological Testing Invited conference session. (2024, April). Annual meeting of the American Educational Research Association, Philadelphia, PA, United States.

Russell, M. (2023). *Systemic racism and educational measurement: Confronting injustice in testing, assessment, and beyond.* Routledge.

Hu, S. (2025). *The Utility of Differential Item Functioning (DIF) Analysis for Students Registering with a Fee Waiver in a K-12 Admission Assessment Program* [Unpublished internal report]. Educational Records Bureau.

Sinharay, S., & Johnson, M. (2025, April). Findings from culturally responsive assessments comprising original and adapted NAEP items. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Sireci, S. G. (2020). Standardization and UNDERSTANDardization in educational assessment. *Educational Measurement: Issues and Practice, 39(3), 29–36*. https://doi.org/10.1111/emip.12377.

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher, 32(2), 3–13.* https://doi.org/10.3102/0013189X032002003.

Steedle, J. T., Anguiano-Carrasco, C., Lewin, N., & McVey, J. (2023). Developing Culturally Relevant Math and Science Items (Research Report). ACT, Inc.

Taylor, C. S., & Ferrara, S. (2025). A measurement argument for culturally responsive assessment. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy (pp. 135–155).* Routledge. https://doi.org/10.4324/9781003392217-11.

U.S. Department of Education, National Center for Education Statistics. (2022). *2020-21 National Teacher and Principal Survey (NTPS): Public school teacher & private school teacher data files* [Data set]. U.S. Department of Education. https://nces.ed.gov/surveys/ntps.

UNICEF Innocenti (2024). Youth, protests and the polycrisis. Research Report.

Valdivia Medinaceli, M., & Steedle, J. (2025, April). Psychometric impacts of enhancing cultural representation in math items. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

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.

White, L., Nesbitt, J., Roeters-Solano, H., Quesen, S., Lottridge, S., & Lochbaum, K. (2025). Antiracist approaches to scoring large-scale assessments. In C. M. Evans & C. S. Taylor (Eds.), *Culturally responsive assessment in classrooms and large-scale contexts: Theory, research, and practice (pp. 202–221).* Routledge.

Appendix A
Example of Paired Form A and Form B Items

Table A1 below presents a pair of items aligned to a standard covering Numbers and Operations in the Mathematics Achievement section of the test. The item on the left is culturally sustaining (CS), including elements of youth culture, one of the themes generated by the student focus group. The item on the right is a 'scrubbed' version of the item measuring the same skill.

Table A1. Paired Item Example: Culturally Sustaining (CS) vs. "Scrubbed" Items

Form A CS Item	Form B "Scrubbed" Counterpart
Local R&B artist Samir has been working hard to get their original music noticed online. Their first single was used in many short videos on social media and has been viewed 500 million times. Samir wants to increase the total view number by 2% next month. Which of the following choices correctly represents their target total number of views by the end of next month? A. 5.01×10^{8} B. 5.01×10^{6} C. 5.1×10^{8} D. 5.1×10^{6}	Consider the quantity below: 500 million increased by 2%. Which of the following options correctly represents this quantity in scientific notation? A. 5.01×10^{8} B. 5.01×10^{6} C. 5.1×10^{8} D. 5.1×10^{6}