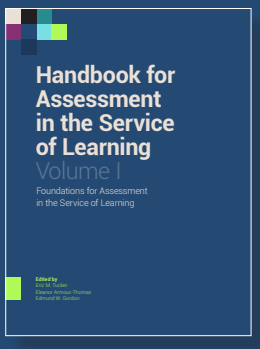
Perspectives on Socioculturally Responsive Assessment in Large-Scale Systems

Aneesha Badrinarayan, Randy E. Bennett, and Linda Darling-Hammond

UMassAmherst
University Libraries



Series Editors:
Edmund W. Gordon, Stephen G. Sireci, Eleanor Armour-Thomas, Eva L. Baker, Howard T. Everson, and Eric M. Tucker

ISBN: 978-1-945764-33-2

Suggested Citation:
Badrinarayan, A., Bennett, R. E., & Darling-Hammond, L. (2025). Perspectives on socioculturally responsive assessment in large-scale systems. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), H*andbook for assessment in the service of learning, Volume I: Foundations for assessment in the service of learning.* University of Massachusetts Amherst Libraries.

Perspectives on Socioculturally Responsive Assessment in Large-Scale Systems

Aneesha Badrinarayan, Randy E. Bennett, and Linda Darling-Hammond

Abstract

This chapter synthesizes key insights from Socioculturally Responsive Assessment: Implications for Theory, Measurement, and Systems-Level Policy (Bennett et al., 2025), offering a cross-cutting analysis of how large-scale assessments can become more valid, equitable, and educationally meaningful when designed through a socioculturally responsive lens. The authors highlight four central themes: aligning assessment content with students' lived experiences; increasing personalization of assessments via testing processes; broadening construct definitions to reflect diverse cultural, linguistic, and epistemological perspectives; and using frameworks and tools that guide inclusive development and interpretation. Drawing on examples such as the KĀʻEO Hawaiian language immersion assessment, AP Art and Design, Smarter Balanced, and adult education programs, the chapter illustrates how sociocultural responsiveness can be integrated into both test content and administration while maintaining technical quality. It also surfaces enduring challenges, including tensions between comparability and local validity, the need for assessments to reflect community values, and the evolving role of technology in supporting adaptive and culturally grounded assessments. The authors conclude by naming critical open questions—such as whose culture is centered, how assessment use cases influence design, and how to ensure assessments both reflect and serve diverse learners—that must be addressed to advance Socioculturally Responsive Assessment (SCRA) as a transformative and sustainable paradigm for large-scale educational assessment.

Introduction

At the heart of socioculturally responsive assessments (SCRA) are goals to intentionally account for the fact that learning—and demonstration of that learning—is inherently a social and cultural activity (Nasir et al., 2020; National Academies of Science, Engineering, and Medicine, 2018), and that assessments that do not account for how students develop knowledge cannot hope to accurately surface and communicate what a diverse range of students knows and can do. This fact influences both what we assess as well as how we surface evidence of learning. SCRA (Bennett 2023) can support culturally responsive, relevant, and sustaining teaching and learning through two focal strategies: (1) instruments that are themselves designed to be socioculturally responsive, and (2) instruments that incentivize and support culturally responsive, relevant, and sustaining pedagogical practices at various levels within the system (Badrinarayan et al., 2025). Identifying features of assessment systems that can meet one or both of these goals requires understanding major findings and advances across bodies of work focusing on the most effective teaching, learning, and assessment approaches for specific groups of students, and extending those learnings to assessment design, development, and implementation.

In compiling the recent edited volume *Socioculturally Responsive Assessment: Implications for Theory, Measurement, and Systems-Level Policy* (Bennett et al. 2025d), we sought to synthesize a range of leading perspectives on approaches, technical considerations, and instrument and process designs intended to serve a common purpose: the development of assessments that are responsive to the unique and shared social, cultural, and linguistic experiences that shape how people learn and how they make what they know and can do visible. The resulting volume intentionally explores the diversity of ideas related to the conceptualization of SCRA, technical approaches for contending with culture and language at scale, working examples and policy considerations of systems that have centered culture and language in their design and implementation, and ongoing research to provide empirical evidence for how the attention to sociocultural factors contributes to learning and performance so learners make what they know and can do visible on large-scale assessment instruments. This chapter offers insights drawn from across the range of scholarship present in the volume to better characterize SCRA at scale, focused on emergent themes related to:

1. Relevance of test content
2. Personalization via testing processes
3. Broadening of construct definition
4. Assessment development frameworks for SCRA

Taken together, the ideas reflected in this volume provide important considerations for large-scale systems intended to support learning as a primary goal. It should be noted that while many scholars and practitioners have considered how to make assessments more culturally responsive (see Bennett, 2025 for a detailed review), this chapter focuses on the specific insights offered by the authors within our volume.

Theme One: The Importance of Relevance of Test Content to Students' Lived Experiences

Across many different conceptions, socioculturally responsive assessment design is predicated on the idea that how examinees interact with an assessment is not a fixed feature of assessment design, and instead is inextricably linked to the social, cultural, and linguistic contexts within which learning and performance happen. A goal, then, of SCRA is to bring assessment instruments into better alignment with examinees' experiences through various approaches to increasing relevance and personalization (Bennett 2023, 2024, 2025). In assessments that are closer to the classroom, relevance and personalization are often achieved through deep relationships with the learners taking the assessment. For example, a teacher may account for their students' home languages and shared classroom experiences by adapting a unit assessment to better match their students' specific lived experiences; a different teacher may select particular texts she knows have relevance to interests and identities present in her classroom when designing an assessment of reading comprehension (Ebe, 2025). Such approaches to personalization and relevance are often considered best practice for responsive teaching when done in the classroom, but become infinitely more challenging when similar principles and expectations are extended to large-scale assessments that (1) operate across many different—and dynamic—learning contexts, student groups, and student experiences (2) are developed by assessment designers who do not have relationships with learners/examinees, and (3) often have fewer degrees of freedom for local adaptation by students and/or teachers. Authors in the volume

explore the relationship between students' funds of knowledge, test relevance, and performance through many different lenses, including disciplinary perspectives, approaches that attend to cultural ways of knowing, and ideas that account for relevance by broadening our conceptions of measurement targets.

Relevance Within Disciplinary Contexts: Insights From Reading Assessments

Reading assessments have been an area of focus for SCRA because of the considerable evidence that students' experience with and background knowledge of a given topic influences how they understand and make meaning of related texts. Text selection for large-scale assessments is a promising and important direction for content-based approaches to personalizing assessment instruments. Most large-scale reading assessments focus on common, provided texts as the basis for measuring students' abilities, and many qualitative and quantitative features of texts (e.g., text complexity, text types) can be accounted for while still allowing for more socioculturally grounded decisions about the nature of the texts students interact with on assessments. Ebe (2025) offers *The Cultural Relevance Rubric* as one way to identify whether texts are culturally relevant to the students engaging with them. The rubric, which is described in detail later in this chapter, asks students to rate texts on a series of factors related to age, geography, language use, and types of activities discussed, to determine how close the texts are to given students' experiences. The rubric gives students agency in determining the relevance of texts, and Ebe suggests that the rubric could be used either as a method for selecting the most appropriate text on a given instrument or be reported alongside reading scores to provide important context for interpreting student performance and planning next steps.

Similarly, Skerrett and colleagues (2025) discuss how large-scale reading assessments can better account for the social nature of reading, students' funds of knowledge, and the range of student experiences with texts and topics. Drawing on the development of the 2026 National Assessment of Educational Progress Reading Framework, Skerrett et al. suggest several strategies for socioculturally responsive reading assessments, including:

- Multimodal knowledge scaffolds that leverage video, audio, and visual supports to support background knowledge students may need to access and make sense of the texts present on the exam;
- Allowing students to respond in home languages; and
- Embedded contextual probes (similar in many ways to the questions Ebe [2025] proposes in the Cultural Relevance Rubric) that can help assess student interest, motivation, and familiarity with texts.

Wang and colleagues (2025) offer empirical evidence to support the idea that using culturally relevant texts affects students' reading comprehension performance on large-scale assessments. They explore how Black and non-Black student performance on reading assessments changed based on the relative familiarity and relevance of the texts used on each form. They found that Black students had significantly less background knowledge on two of the three topics—immigration and ecosystems—but had similar levels of background knowledge on the third topic, the Harlem Renaissance. Their performance mirrored this finding: Black students spent more time on task and performed better on the Harlem Renaissance form than on the other two forms. Importantly, using the form with higher relevance and familiarity to Black students did not adversely affect non-Black students' performance, nor did the high-relevance form demonstrate any threats to psychometric quality. Taken together, these findings are consistent with calls for more culturally attentive considerations when choosing texts, and suggests that text selection can be a critical way to personalize and make more relevant large-scale assessments when those selections are made with explicit considerations for the experiences of the diverse examinees intended to be taking the assessment.

Relevance to Specific Cultural and Linguistic Ways of Knowing.
A vital component of many conceptualizations of SCRA is the concept of cultural validity—the extent to which assessments account for the cultural, linguistic, and social repertoires students draw upon when interpreting and responding to disciplinary tasks (Solano-Flores & Ruiz-Primo, 2025; Nelson-Barber & Trumbull, 2025; Solano-Flores & Nelson-Barber, 2001). Cultural validity concerns how these repertoires shape meaning-making and influence not only performance but also the validity of inferences drawn about student understanding.

Nelson-Barber and Trumbull (2025), drawing on work with Indigenous communities, argue that cultural validity is as essential as construct validity: disregarding students' cultural and linguistic contexts during assessment development and score interpretation introduces threats to overall validity. For example, for Indigenous students—including American Indian, Alaska Native, Native Hawaiian, and other Pacific Islander groups—histories, governance structures, belief systems, heritage languages, cultural values, and community-based practices are distinct from Western, Eurocentric norms. When assessments ignore these lived realities, they may:

- Present information in unfamiliar or incongruent ways that create barriers to comprehension;
- Require response modes that conflict with students' typical ways of demonstrating understanding;
- Apply narrow success criteria that devalue culturally rooted ways of knowing and sensemaking; and
- Reinforce experiences of dehumanization, marginalization, or erasure, which can undermine motivation and perseverance.

In all such cases, the result is the same: student performance no longer reflects what students know and can do, but rather their ability to navigate unfamiliar or invalidating assessment structures.

While these issues have been examined carefully in the assessment of Indigenous students, they also extend to many other populations whose cultural and linguistic experiences are underrepresented in mainstream testing assumptions. These groups include multilingual learners; students with disabilities; students from urban, rural, or remote communities; students from low socioeconomic backgrounds; and immigrant or migrant youth, as well as students encountering assessments in countries other than the United States. For example, Raji and Baidoo-Anu (2025) examined the cultural relevance of test items on the West African Senior School Certificate Examination, a high-stakes university entrance exam used in Ghana and Nigeria. They found that some items were culturally problematic for the test-takers expected to engage in the exam: for example, stimulus information was sometimes culturally disconnected from West African students' experiences, and test language was in some instances unnecessarily complex. Based on their analysis, they

raise concerns that the exam may potentially unfairly limit or deny postsecondary opportunities for students.

Many assessment developers find balancing the fundamentally different world views required to enhance cultural validity nearly insurmountable in large-scale assessment design. However, Nelson-Barber and Trumbull (2025), Solano-Flores and Ruiz-Primo (2025), and Raji and Baidoo-Anu (2025) recommend some concrete steps that assessment developers can embed within assessment content development to strengthen cultural validity and ensure assessments are experienced as more relevant and valid measures of student understanding and ability:

- Define a clear theory of action connecting assessment content, student experiences, and intended outcomes;
- Consider how students' cultural experiences and ways of knowing influence interpretation of test content and response strategies—do not assume any particular understanding or premise is "universally" understood unless there is strong evidence to suggest all test-takers will have had sufficient experiences to have a shared understanding; this consideration might include some degree of population mapping to better understand the cultural and linguistic experiences and assets of the students to be engaged in the assessment;
- Right-size linguistic and semiotic complexity so that all students can access and engage deeply with assessment tasks, including through multimodal scaffolds and contextual cues that serve to enhance engagement without limiting opportunities for rich meaning-making;
- Broaden interpretations of what counts as a "correct" or valid response, accounting for diverse ways and modes of sensemaking;
- Include culturally and linguistically representative students throughout the assessment development process (e.g., through interest surveys, cognitive labs, co-design efforts); and
- Provide opportunities for open-ended responses that enable students to show their thinking in rich, authentic ways. These opportunities should include allowing for multiple modes of expression and authentic engagement with cultural funds of knowledge, community-based knowledge systems, and diverse perspectives and values (e.g., as part of selecting claims or making meaning of presented stimuli).

Creating relevant large-scale assessments fundamentally requires that test developers generate content by making some assumptions about students—assumptions can range from extrapolating insights from deep and intentional student engagement to a broader set of students, to making assumptions that certain aspects of students' backgrounds (e.g., race, ethnicity) will confer experiences or perspectives to all students who share those elements of identity. Authors in the volume repeatedly caution against essentializing communities—reducing complex, dynamic cultures to fixed traits or stereotypes. Instead, assessment design must be grounded in deep understanding of the specific histories, languages, and epistemologies of communities. For example, in the case of Indigenous students, this grounding might include recognizing their status as sovereign nations and respecting their self-determined educational and cultural frameworks (Nelson-Barber & Trumbull, 2025), as well as intentionally and explicitly contending with these factors when designing large-scale assessments for those students.

Theme Two: Right-Sizing Standardization of Testing Processes to Increase Personalization

Another approach to bringing assessment experiences into better alignment with students' social, cultural, and linguistic experiences is allowing for personalization or user-specific adaptations in ways students interact with assessments. Sireci and colleagues (2025) describe a framework for UNDERSTANDardization in which decisions about what aspects of test administration are standardized are driven by learner-centered approaches that seek to better understand how different groups of students might understand and interact with the test, rather than rigid models of standardization that emphasize uniform testing conditions as a primary way to ensure fairness and validity of score interpretation. Sireci et al. (2025) discuss UNDERSTANDardization in the context of two adult-centered assessments—the Adult Skills Assessment Program (ASAP) and the English Test for Adults (ETA)—both of which are intended to be administered to adults at scale. By leveraging literature reviews and focus groups to understand user experiences of both the examinees and the educators and employers who might use the resulting data, researchers identified several ways for being more flexible in the assessment administration. Such methods as translanguaging options and real-time translations of instructions, choice in task format, and adaptive pacing could potentially improve score validity and the utility of the assessments.

Sireci and colleagues' UNDERSTANDardization framework provides a useful model for examining administration-related approaches to SCRA. In the volume, authors describe two major approaches to this kind of personalization in K–12 assessment systems: (1) adaptations and accommodations that allow learners to better access the *same* content, and (2) allowing for student agency in deciding how students will demonstrate evidence of progress, proficiency, and mastery relative to common constructs via different content. For example, Michel and Shyyan (2025) describe the accessibility supports within the Smarter Balanced assessment system (SBAC) as a set of strategies for personalizing the testing experience such that all learners—particularly students with disabilities and multilingual learners—can engage meaningfully with the same test content. Smarter Balanced assessments are built around relatively narrow constructs tied to states' math and English-Language Arts standards and are designed to be administered across diverse state contexts and student backgrounds. The assessments' primary goal is to generate evidence of student progress toward academic standards for federal accountability purposes within each state, which requires a certain amount of rigidity in content and administration. Despite these requirements, the assessment system incorporates universal design features, such as language supports and alternative response modes, that allow students to access and demonstrate targeted knowledge and skills. This strategy of targeted personalization within a standardized system offers one potential pathway for making large-scale assessments more socioculturally responsive.

Smarter Balanced provides multiple levels of support for personalization, including:

- Universal tools that are available to all students (e.g., digital notepad), allowing them to customize their testing experience to better account for how they make meaning and access content;
- Designated supports that are available to students with educator-identified needs (e.g., glossaries in 13 languages) to support specific kinds of engagement; and
- Accommodations for individual students with documented IEP or 504 plans (e.g., braille versions).

These supports are designed to ensure that each student engaging with the assessment has the best opportunity to show what they know and can do. The intention is to *enhance* construct comparability, rather than detract from it by ensuring that differences in performance reflect differences in skill, not access. While these supports introduce some variability in test administration procedures, ranging from minor (e.g., notepad use) to more substantial (e.g., braille), they increase fairness by enabling students to more accurately demonstrate competency. This approach illustrates how standardized assessment systems can integrate flexibility to better serve a diverse population without compromising measurement integrity.

In a distinct disciplinary context and serving different purposes, Escoffery and colleagues (2025) present the Advanced Placement (AP) Art and Design examination as a compelling model for integrating SCRA principles into large-scale assessment design. This through-course portfolio examination operates at the intersection of formative and summative assessment, targeting a set of constructs that readily support flexible interpretation in terms of topic, content, and medium of demonstration, such as sustained inquiry; practice, experimentation, and revision; synthesis of materials, processes, and ideas; and creative artmaking. What distinguishes AP Art and Design is how it approaches standardization: Students are afforded significant autonomy to engage in artistic production across diverse media, enabling them to draw upon their cultural identities, personal experiences, and individual interests as integral components of demonstrating proficiency. This design feature explicitly positions the student's sociocultural background as a resource rather than a variable to be controlled. What is more rigorously standardized is the evaluation criteria and scoring process. A clearly articulated construct definition, operationalized through a detailed rubric and implemented by highly trained art educators, ensures construct comparability across an inherently diverse array of student submissions—from mat-board constructions to paintings to interactive installations. While the student work varies substantially in form and content, score comparability is maintained through robust rater training, the use of shared interpretive standards, and a calibrated scoring process. This approach illustrates a deliberate and productive tradeoff: by foregrounding construct clarity and scoring consistency, the assessment system supports valid and reliable credit and placement decisions while simultaneously honoring student agency in form, expression, and cultural perspective.

Indeed, principled design for student agency like that seen in AP Art and Design, in which students can directly play a role in appropriately personalizing assessments by making choices grounded in their own lived experiences, is a promising direction for assessments that seek to be authentically responsive to students without the harms of developer-based assumptions about what is most relevant or meaningful to learners and their performance. Many chapters in the volume touch on aspects of student agency—for example, Ebe's (2025) Cultural Relevance Rubric relies upon student-determined relevance, and Badrinarayan and Darling-Hammond (2025) describe several national and international assessment systems that include elements of student choice and agency in determining the tasks and topics with which individuals will engage (e.g., through AP and International Bacalaureate assessments).

Theme Three: Expanding Construct Definitions to Account for Social, Cultural, and Linguistic Contributions to Learning and Performance

Across arguments for increasing the relevance and personalization of assessments for specific learners, there has been an implicit—and at times explicit—call to rethink both what assessments are measuring and how those measurements are interpreted. This includes calls to:

1. Expand the range of assessment constructs to better reflect how learning and development actually occur;
2. Incorporate culturally and community-specific priorities and goals for student learning; and
3. Develop more trustworthy and inclusive measures of what students know and can do, even within narrowly defined domains.

Together, these shifts aim to ensure that assessments more accurately reflect the full breadth of student learning and experience.

Lee (2025) contends that the Science of Learning and Development (SoLD)—which synthesizes interdisciplinary research from human development, psychology, neuroscience, and the learning sciences to offer a comprehensive understanding of the diverse factors that shape learning across the lifespan—requires a fundamental rethinking of the aims and design of educational assessments. That rethinking has the potential to yield more actionable, equitable, and ecologically valid insights.

Lee argues that persistent disparities in assessment outcomes—by race/ethnicity, gender, and socioeconomic status—stem not only from inequitable opportunities to learn but also from the limitations of existing assessment systems. Current summative, interim, and formative approaches fail to capture the full range of influences on how people learn and demonstrate knowledge. These influences include cultural identity, perceptions of the task and setting, emotional salience, epistemological beliefs, mindsets, and self-efficacy.

Moreover, Lee critiques the dominance of narrow, Eurocentric definitions of disciplinary knowledge in U.S. education, which constrain both teaching and assessment. She argues that expanding our conceptions of what counts as knowledge—in ways that are more culturally and contextually responsive—could reduce disparities, foster a more holistic view of learners' capabilities, and better recognize the strengths and knowledge systems students from historically marginalized communities bring to school.

Lee's challenge to narrow definitions of disciplinary knowledge in current large-scale assessment systems is echoed in many other chapters in the volume. For example, Welch and Dunbar (2025) examine opportunities for integrating SCRA into federally mandated state assessments. Based on their analysis, they argue that current interpretations of alignment in both item development and item/test evaluation (e.g., for purposes of federal peer review) privilege overly narrow conceptions of what it looks like to demonstrate performance relative to established standards. They suggest that the underlying reason for these narrow conceptions are at least two-fold. First is that the standards themselves were often not developed with cultural relevance in mind—expectations for student performance are often imbued with White, Western-dominant perspectives and worldviews that then are operationalized in assessments accordingly (e.g., language standards that prioritize standard written English, use of terms like "effective" and "appropriate" which can marginalize students with diverse linguistic repertoires and serve to reinforce dominant structures in terms of whose culture is valued). Second, current conceptions of assessment alignment focus on very narrow interpretations of item-standard mapping and matching; if instead alignment were considered more holistically and inclusively, that consideration might allow for culture and lived experience to play a more meaningful role in how items surface—and students demonstrate—progress toward the goals being measured on an assessment, with

the result that large-scale math and ELA assessments would be better positioned to support student progress.

Kukea-Shultz and Englert (2025) describe how ideas related to reimagining what is assessed on large-scale, federally mandated assessments are factored into a culturally-specific operational assessment: the Kaiapuni Assessment of Educational Outcomes (KĀʻEO), the accountability assessment for students attending Hawaiian language immersion programs. Developing KĀʻEO required explicitly countering the standard, monocultural worldviews that govern most large-scale assessment development processes because central to KĀʻEO's purpose is the reclamation of Hawaiian culture and language; this reclamation required defining constructs in ways that are responsive to the local community's language, culture, and information needs while still meeting federal requirements for showing progress in mathematics, language arts, and science. Kukea-Shultz and Englert (2025) describe specific ways in which the measurement targets for KĀʻEO were culturally defined, including:

- Developing culturally and linguistically specific student learning outcomes that align with the intent and goals of Hawaiʻi's state standards in math, language arts, and science (Common Core State Standards and the Next Generation Science Standards) but reflect Hawaiian linguistic and epistemological priorities;
- Extending measurement goals beyond state standards to include culturally-specific forms of knowledge (e.g., use of metaphoric language) that are not generally prioritized in language arts assessments;
- Using the value of Hawaiian language, knowledge, and culture as a consistent lens throughout the assessment development process.

KĀʻEO also relies upon a deeply collaborative and relational approach, placing educators and community members within the Kaiapuni system as decision-makers throughout the assessment development and validation process. For example, teachers, families, and community members are directly engaged as part of the team establishing the student learning outcomes and goals of the assessment; teacher judgments of student performance are used as part of the process for validating test scores; and culturally relevant opportunity-to-learn measures are included as part of reporting efforts to facilitate collaborative, action-oriented meaning-making around student performance. Kukea-Shultz and Englert (2025) posit community

validity—reflective of the intentional processes governing KĀʻEO's development and validation—as a framework for assessment development that positions assessment as an activity that serves communities as a first-order principle.

Many scholars (e.g., Lee 2025, Nelson-Barber and Trumbull 2025, Kukea-Schultz and Englert 2025) emphasize the need to disrupt the dominance of White, Eurocentric cultural norms that shape not only how assessments are designed, but also how students experience schooling more broadly. Zandvakili and Gordon (2025) highlight what is often the elephant in the room—that European and American cultural frameworks continue to govern the operation of much of the developed world. They argue that while education must necessarily contend with these dominant norms, it must also intentionally accommodate the diverse cultural backgrounds, identities, and experiences of learners.

To do so, they propose designing assessments that cultivate, recognize, and incentivize a broader range of competencies—competencies that emerge through students' negotiation of their own cultural frameworks within an increasingly diverse and interconnected world. The authors describe five core competencies that would serve to equip learners to thrive in both dominant and marginalized systems while contributing to inclusive educational environments that value and validate a wide range of intelligence and knowledge:

- Accession: Embracing diverse perspectives and cultural knowledge;
- Accommodation: Adjusting behavior and mindset in response to different environments;
- Adaptation: Navigating new or evolving situations with flexibility;
- Adjustment: Fine-tuning strategies in response to feedback and challenges; and
- Agility: Thinking creatively and critically to solve complex problems.

Centering these competencies requires a new approach to assessment. Zandvakili and Gordon (2025) suggest that rather than relying on passive instruments rooted in a single cultural worldview, systems could center agentic assessments that are responsive to students' individual contexts while also actively inviting learners to apply their cultural and linguistic assets to engage with sociocultural perspectives different from their own. For example, culturally universal probes and agency-oriented probes might ask students to connect their learning to personal, cultural,

and community-based experiences and problem-solving of their own choosing, while critical thinking probes might encourage students to explicitly take multiple perspectives, practice empathy, and engage in deep analysis.

Across testing contexts, many of the book's chapter authors contend that transforming assessment requires a fundamental shift in what we value, how we define competence, and whose knowledge is recognized. By bridging definitions of what we measure—and what we value—with the full diversity of learner experiences, they suggest that assessment systems can become more inclusive, relevant, and empowering.

Theme Four: Assessment Development Processes that Account for Sociocultural Goals

A fourth theme that runs through the book is the need for and use of frameworks and tools to guide assessment development and score interpretation in ways that align with the conceptual and evidence-based recommendations for SCRA. Frameworks have long been used with standardized tests for such purposes. Classical test theory (Gulliksen, 1950), item response theory (IRT; Lord, 1980), evidence-centered design (ECD; Mislevy et al., 2003), and the argument-based approach to validation (Kane, 1992) are widely used examples. Frameworks are important because, in the best case, they offer principled approaches for thinking about a problem and taking action to address it.

SCRA presents no shortage of problems, including ones related to assessment design, development, analysis, and interpretation. Key to SCRA design is taking account of examinee sociocultural characteristics to allow individuals to demonstrate better what they know and can do. Sato (2025) offers an approach to design that focuses on deeper levels of culture, with the intention of accounting for those factors (e.g., values, norms, beliefs, language, social structure/dynamics, milieu) that affect students' meaning making and their representations of knowledge. The chapter presents a *sociocultural dimensions matrix* describing personal orientations. The matrix should be of use in designing more inclusive measurement targets, tasks, and scoring rules, as well as for guiding interpretations of diverse student performance.

The matrix is organized around three sociocultural dimensions that can influence an examinee's comprehension and, hence, their expression of knowledge: Social Relationships/Orientation; Epistemological Beliefs/Cognitive Patterns; and Communication Patterns. The social relationships/orientation dimension reflects such propensities as individualistic vs. collectivist and nurturing vs. challenging patterns of behavior, whereas the epistemological beliefs/cognitive patterns dimension concerns tendencies toward analytic vs. holistic and random vs. sequential thought. Five broad communication patterns are delineated: English, Romance (e.g., Spanish, French, Italian), Semitic (Hebrew, Arabic), Asian (Japanese, Vietnamese, Mandarin, Cantonese), and Russian. As noted in the matrix, the logical structure found in English tends toward the deductive presentation of information, with ideas related in an orderly sequence. In contrast, romance languages have a logical structure that is more likely to engender lines of thought that sometimes pursue complex digressions. Semitic logical structures may produce parallel lines of development, including tangential information, whereas Asian structures may encourage circular reasoning by indirection. Finally, Russian communication patterns typically entail one or more lines of development.

As Sato notes, the matrix suggests—and research supports—the contention that individuals with particular orientations process information differently. For example, the orientation of European Americans toward analytic ways of thinking may result in taxonomic reasoning whereby objects are categorized conceptually based on shared attributes. In contrast, Chinese who tend toward collectivist and holistic orientations, are more inclined toward relational reasoning, which may lead to grouping items on common functions. Glick (as cited in Greenfield, 1997) relates the consequences of such differences in orientation for ability test performance using a categorization task with Liberian subjects. Repeated trials across multiple examinees resulted in functional groupings (e.g., potato and knife) because that was what a "wise man" would do, rather than the expected separate conceptual groupings of foods and of tools. When asked how a fool would organize the objects, the participants quickly produced the expected conceptual groupings. Clearly, the examinees' and examiners' notions of intelligent behavior were culturally determined—and opposite one another. In short, sociocultural orientations matter. Sato's (2025) matrix offers a framework for understanding them better and acting upon that understanding, especially for purposes of assessment design and interpretation.

In her chapter, Ebe (2025) continues the concern with more effectively accounting for sociocultural orientation in assessment. As noted above, her focus is upon the cultural relevance of text in reading passages. In line with Sato's (2025) chapter, as well as with much reading research (Lee, 2025; Wang et al., 2025), Ebe's premise is that comprehension depends critically upon relevant prior knowledge. We more quickly and completely understand text that draws upon what we already know. Moreover, what we know is culturally shaped, as Glick's experience (cited in Greenfield, 1997) so memorably attests. Thus, knowing the cultural relevance of text should help in SCRA design and score interpretation. As described above, Ebe (2025) offers a tool to do just that. The cultural relevance rubric consists of eight questions, each of which is rated on a 1–4 Likert scale. The questions concern the characters, setting, and experiences described in any given text. The ratings, typically done by the student, are intended to index the proximity of the text to the student's lived experience.

Several previously published studies support the utility of the rubric. In one study, Ebe (as cited in Ebe, 2025) asked 3rd-grade emergent bilingual students to read and retell two stories from a commercial reading kit that identified the stories as being at the same level. After reading each story, students rated its relevance. The recordings of each retelling were then analyzed using miscue analysis to identify how well readers were using semantic, syntactic, and graphophonic cues, along with background knowledge to comprehend the text. All students were more proficient reading the story they rated as more relevant. Retellings were also more accurate, detailed, and complete for the more relevant story.

Ebe (2025) suggests two potential uses for the rubric. One use might be as a guide to text selection, with student ratings collected as part of a text-evaluation phase in examination development. The desired result would be to locate relevant texts for each numerically significant demographic group. A second use might be for scoring and interpretation. In this use, test-takers would rate the passages they read. Along with a conventional score, scores that weighted items according to passage relevance could also be computed, thereby giving test users a sense of how comprehension varied with perceived relevance for each student.

A general framework for guiding the development of SCRAs is offered in the chapter written by Badrinarayan and Darling-Hammond (2025). The framework was derived from a review of state and national attempts to account for sociocultural factors in large-scale assessment systems. The reviewed systems were of four types: (1) assessments constructed for a specific community's language and culture (e.g., students attending Hawaiian language immersion programs); (2) assessments that are embedded in High Quality Instructional Materials developed to be culturally responsive (e.g., Washington State's use of OpenSciEd); (3) assessments whose items should appeal to a wide array of cultural identities because they cross such factors as race/ethnicity, age, gender, language, immigration experience, disability, and geography (e.g., NAEP 2028 Science Framework); and (4) assessments that allow students to choose among tested subjects and/or problems (e.g., Finnish matriculation exam), as well as design their own problems (e.g., AP Research), enabling students to bring their interests, prior knowledge, and lived experience to bear.

Badrinarayan and Darling-Hammond's framework consists of five features intended to be used as "… a heuristic for defining the potential landscape of culturally conscious assessment systems at scale" (p. 362). The first feature is to *Emphasize authenticity, agency, and decision making.* Consistent with this feature are real-world tasks that call upon disciplinary modes of inquiry, allow students choice in what and how to inquire, and permit demonstration of competency through multiple avenues. The second feature is to *Create tasks that are relevant and meaningful to specific communities.* This feature is instantiated through assessment problems that are deemed important to communities, call upon students to generate solutions aimed at positive impact, and ask students to apply their disciplinary knowledge to address issues, keeping in mind social, economic, environmental, and political concerns. The third feature is to *Center asset-based narratives of minoritized people and communities, bringing diversity to those positioned as knowers and doers.* Assessment tasks consistent with this feature include ones that represent non-dominant people as role models in the discipline and world, avoiding superficial or stereotypic depictions. Fourth is to *Emphasize dynamic relationships to cultural relevance and perspective taking.* In keeping with this feature, tasks should be designed to simultaneously reflect the cultural experiences of some students, while acting as cultural learning opportunities for other students. Additionally, task design should encourage students to bring their own ideas, experiences, and perspectives to sense-making about the phenomenon under

study. The last framework feature is to *Engage positive, productive affect and effort.* The intention behind this feature is to build tasks around interesting, compelling phenomena that cause students to engage and persevere. Collaboration is cited as another mechanism for fostering a positive context for problem solving.

Sato's (2025), Ebe's (2025), and Badrinarayan and Darling-Hammond's (2025) chapters principally focus on frameworks and tools for SCRA development. In contrast, Moses (2025) deals primarily with issues of analysis and score interpretation. Moses contrasts and compares sociocultural theories of learning and development with the practices used in large-scale assessment for test scoring and linking. He notes that tests developed from a socioculturally responsive perspective may vary across examinees by, among other things, presenting different questions or posing them in ways that suit test takers' backgrounds. Large-scale assessment practices, in contrast, suggest that the maintenance of scores would ideally require that the tests be designed to the same constructs and specifications rather than adjusted to the backgrounds of different test takers. Moses depicts the challenge as a tradeoff between comparability across examinees vs. validity for a particular use and/or examinee group. He suggests a framework consisting of three ways in which this *local validity* vs. *broad comparability* tension might be resolved.

The first possibility is to standardize to the most appropriate group or construct, where the construct and assessment are engineered from the outset to be responsive to that group (e.g., KA'EO; Kūkea-Shultz & Englert, 2025). To facilitate score comparability, subsequent test forms are created, given, and scored following standardized procedures. Under this possibility, score interpretations are similar for all examinees, within and across test forms. Moses' second possibility is to keep scores inferences local, meaning particular to the individual or to that subset of individuals taking essentially the same assessment under similar conditions. Comparability across the entire group is restricted but validity for individuals or specific groups may be enhanced. The last possibility is to expand the construct definition to account for the variation resulting from a responsive test. For example, on an assessment that allows examinees to create problems (e.g., AP Research), their design choices will, implicitly or explicitly, become part of what is measured and, therefore, what examinees are being compared upon. By virtue of allowing problem creation, each examinee takes a different test, each test intended to measure the same high-level construct (e.g., the ability to design, conduct, and

defend a study). The expanded construct definition, a rubric to connect disparate examinee performances to that definition, and rater training/monitoring processes then become mechanisms for generating scores that are both roughly comparable and locally valid.

Like Moses, Mislevy et al. (2025) deal with the issues posed by SCRA for analysis and score interpretation, but additionally with the implications of SCRA for development. These authors propose a lens that connects the logical assessment-argument structures of Evidence-Centered Design (Mislevy et al., 2003) with the sociocultural aspects of tasks (e.g., aspects that are construct essential, related to sociocultural background, ancillary, enabling or restricting, genre specific). That combination allows for analyzing the relations among tasks, students, purposes, and inferences to facilitate assessment design decisions. The combination also allows for a better understanding of SCRA's potential effects on score validity vs. comparability.

Mislevy et al. make the useful distinction between *data comparability* and *construct comparability*. Data comparability results when all examinees take the same assessment under the same conditions such that the method of assessment is common. Data comparability does not necessarily imply construct comparability, although it may. In a reasonably homogenous population, using a common method will produce both data and construct comparability. However, when the examinee population is highly heterogenous, examinees will bring different understandings to the assessment that may introduce irrelevant difficulty, thereby compromising construct comparability (see Glick, as cited in Greenfield, 1997). By customizing the assessment to examinee groups or individuals, SCRA works to achieve construct comparability, though at the expense of lower data comparability.

Dealing with the consequence of lower data comparability calls for what Mislevy et al. deem a "rectification argument," in essence a mechanism for placing scores from disparate assessments on the same scale. Different types of rectification may be appropriate depending upon the situation, context, desired inference, populations, and differences across the assessments given to individuals. When parallel forms are administered with common anchors or common persons (e.g., as for the SAT), rectification can be achieved through standard equating procedures. When the assessments are not parallel but the respective populations are sufficiently similar that students can also take common anchors, concordances may be established that allow more limited types of inference (e.g., for groups

rather than individuals). In cases where both the assessments and the examinee populations are different, as is true for many SCRAs, score rectification can be approximated through other mechanisms, including expert judgment, the use of common higher-level rubrics, and theories that can be used to map different assessments to the same underlying framework (e.g., learning progressions).

Conclusion and Next Steps for SCRA

This chapter draws together a broad and evolving set of ideas, tools, and practices that are redefining what it means for large-scale assessments to be valid, fair, and educationally meaningful in increasingly diverse learning contexts. Across the many contributions to the source volume *Socioculturally Responsive Assessment: Implications for Theory, Measurement, and Systems-Level Policy* (Bennett, Darling-Hammond, & Badrinarayan, 2025b), one message emerges with clarity: SCRA requires more than technical improvements or inclusive messaging—it demands a transformational shift in how we conceptualize, design, implement, and interpret educational assessments at scale.

This shift begins by re-centering the foundational assumption that learning is inseparable from culture, identity, and context, and that assessment must reflect this reality. The chapters in the volume demonstrate how assessment systems grounded in sociocultural responsiveness can be designed to recognize multiple ways of knowing, support meaningful engagement with academic content, and honor the full range of students' cultural and linguistic repertoires. Whether by increasing the relevance of content, personalizing assessment processes, or broadening construct definitions, the work described here challenges dominant paradigms that have historically marginalized non-dominant learners.

Importantly, contributors show that technical quality and sociocultural responsiveness are not mutually exclusive. Through examples such as the KĀʻEO assessment in Hawaiʻi, the AP Art and Design portfolio, and Smarter Balanced's universal design features, the chapters illustrate how construct comparability can be preserved or even strengthened when design decisions are guided by community-informed theories of action, clearly articulated validity arguments, and inclusive development practices. This consideration includes not only attention to item and task development, but also to scoring practices, data use, and communication of results—all of which must be revisited to better serve students and communities.

To that end, the volume also highlights the value of new frameworks and analytical tools—including cultural relevance rubrics, sociocultural design matrices, and approaches to UNDERSTANDardization—that help operationalize SCRA in practical terms. These tools underscore the need to decenter monolithic conceptions of standardization and instead embrace models that prioritize construct validity over procedural uniformity, allowing for more nuanced, community-centered, and asset-based approaches to large-scale measurement.

While the field continues to wrestle with trade-offs between comparability and responsiveness, efficiency and authenticity, and alignment to standards and cultural relevance, the authors in this volume make clear that we must move beyond binary thinking. As Moses, Sato, Kukea-Shultz and Englert, and others argue, validity arguments must be both technically sound and socially credible, grounded not just in statistical evidence, but in the lived realities and values of the communities assessments are intended to serve.

Whereas this chapter highlights common themes across a wide range of ideas in the book, there are many open questions with regard to SCRA. Some of these questions include:

- **Whose culture should be centered and valued in assessment design?** This foundational question arises across scholarly, practical, and policy discussions surrounding SCRA. While some scholars argue that SCRA should create fairer assessments for all learners by expanding inclusivity (e.g., Lee, 2025; Zandvakili & Gordon, 2025), others advocate for explicitly centering the needs and experiences of historically marginalized student groups as a primary goal (e.g., Randall et al., 2022, Randall 2023). This tension becomes more complex as SCRA efforts increasingly seek to account for the nuanced, intersectional identities of learners. For example, although there is broad agreement on the dominance of White, Western/Eurocentric worldviews, there is less clarity about how to represent the experiences of specific subgroups within or adjacent to those dominant experiences—for example, students with particular religious affiliations, those living in rural areas, or individuals whose cultural identities are tied to underrepresented practices, activities, geographies, or communities. Advancing this work will require deeper inquiry into how cultural representation decisions are made within assessment systems—who is involved, what criteria are used, and how trade-offs are weighed. Such efforts are critical not only

for guiding the design and implementation of SCRA-aligned assessments but also for informing broader policy decisions about inclusion, accountability, and equity in education.

- **How do different use cases for large-scale assessment govern appropriate trade-offs for incorporating SCRA into assessment design?** There are many different reasons for including SCRA in assessment design, ranging from humanizing students' experiences with assessment to generating more trustworthy scores. When these reasons are further contextualized by the uses of assessments—both intended and actual—principled decisions about how SCRA is incorporated into different assessment designs can be more effectively made. For example, assessments that are used for advanced placement or admissions decisions may seek to reflect features of sociocultural responsiveness in different ways than assessments that are used primarily to make school- and district-level decisions. The appropriateness of uses might be further differentiated by the relationships among potential users: assessments employed by teachers—who know their students—to review progress with them and their families might attend to SCRA differently than external decision-makers who do not have regular touchpoints with the students being assessed. These distinctions among use cases become even more complex when assessments are used for multiple, sometimes unintended purposes. Nevertheless, making use cases and user relationships explicit can support more intentional, context-sensitive decisions about how to embed SCRA into assessment systems. It can also help identify which aspects of assessment design—such as item content, administration procedures, or reporting formats—require the most focused attention to ensure equitable and meaningful use.
- **How can emerging technologies support SCRA?** As Bennett et al. (2025a) describe, there are significant opportunities for emerging technologies, including artificial intelligence, to better support SCRA. For example, generative AI might support real-time personalization; more efficient generation of item pools that reflect greater linguistic, cultural, topical, and social diversity; the creation of immersive simulations that better capture social reasoning and authentic engagement to support more valid assessments of complex, deeper learning competencies; and the synthesis of evidence across multiple demonstrations of learning to produce more holistic representations of student capabilities. Additionally, AI-powered technologies show promise in

developing flexible scoring mechanisms capable of interpreting responses expressed through varied modalities—such as spoken language, prose, bullet points, graphics, or symbolic representations. These technologies also may help in producing responsive and interactive reporting systems that are better attuned to users' linguistic preferences, implementation settings, and immediate information needs (e.g., using natural language and semantic search to query reports to better understand student performance and next steps; highlighting elements of student performance that are aligned with productive next steps along meaningful learning progressions).

Ultimately, this chapter—and the volume as a whole—underscores that transforming large-scale assessment is both necessary and possible, and is already underway. It is a call to action for assessment developers, policy leaders, educators, and researchers to build systems that reflect a pluralistic vision of learning, one in which all students are seen, heard, and empowered. The future of large-scale assessment must not only measure what students know and can do, but also support who they are and who they aspire to become. If we are to create assessment systems that are truly equitable, valid, and educationally valuable, the work of SCRA cannot be peripheral—it must be central to our reimagining of assessment systems.

References

Badrinarayan, A., & Darling-Hammond, L. (2025). Enacting socioculturally responsive assessment design at scale: Approaches and policies for large-scale systems. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 336–366). Routledge.

Badrinarayan, A., Lyons, S., Miranda, A., & Klyachkina, A. (2025). Leveraging students' cultural and linguistic assets for assessment: A framework for culturally conscious assessment in the Chicago Public Schools. *Educational Assessment, 30*(3), 1–26. https://doi.org/10.1080/10627197.2025.2497775

Bennett, R. E. (2023). Toward a theory of socioculturally responsive assessment. *Educational Assessment, 28*(2), 88–104. http://dx.doi.org/10.1080/10627197.2023.2202312

Bennett, R. E. (2024). Personalizing assessment: Dream or nightmare? *Educational Measurement: Issues and Practice, 43*(4), 119–125. https://doi.org/10.1111/emip.12652

Bennett, R. E. (2025). A descriptive review of culturally responsive, socioculturally responsive, and related assessment conceptions. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 11–28). Routledge.

Bennett, R. E., Darling-Hammond, L., & Badrinarayan, A. (2025a). *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy*. Routledge.

Bennett, R. E., Darling-Hammond, L., & Badrinarayan, A. (2025b). Socioculturally responsive assessment: Present and future. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 434–442). Routledge.

Ebe, A. (2025). Examining the relationship between the cultural relevance of text and reading proficiency: Using a cultural relevance rubric in reading assessment. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 168–179). Routledge.

Escoffery, D. S., Fletcher, K. E., & Stone-Danahy, R. A. (2025). "A search for my voice": Socioculturally responsive assessment in AP Art and Design. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 262–282). Routledge.

Greenfield, P. M. (1997). You can't take it with you: Why ability assessments don't cross cultures. *American Psychologist, 52*(10), 1115–1124. https://doi.org/10.1037/0003-066X.52.10.1115

Gulliksen, H. (1950). *Theory of mental tests*. Wiley.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kūkea-Shultz, P., & Englert, K. (2025). A path to transforming the assessment landscape. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 306–335). Routledge.

Lee, C. D. (2025). Implications of the science of learning and development (SoLD) for assessments in education. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 29–49). Routledge.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum.

Michel, R., & Shyyan, V. (2025). Accessibility as a core value for locally responsive assessments. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 247–261). Routledge.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (Research Report No. RR-03–16). Educational Testing Service.

Mislevy, R. J., Oliveri, M. E., Slomp, D., Crop Eared Wolf, A., & Elliot, N. (2025). An evidentiary-reasoning lens for socioculturally responsive assessment. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 199–242). Routledge.

Moses, T. (2025). Linking responsive assessments: Challenges and possibilities. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 180–198). Routledge.

Nasir, N. S., Lee, C. D., Pea, R. D., & McKinney de Royston, M. (Eds.). (2020). *Handbook of the cultural foundations of learning*. Routledge.

National Academies of Sciences, Engineering, and Medicine. (2018). *How People Learn II: Learners, Contexts, and Cultures.* The National Academies Press. https://doi.org/10.17226/24783.

Nelson-Barber, S., & Trumbull, E. (2025). Assessment for social justice with a focus on American Indigenous students. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 76–98). Routledge.

Raji, M. O., & Baidoo-Anu, D. (2025). Socioculturally responsive post-secondary entrance examination: Implications for equitable assessment design in sub-Saharan Africa. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 399–414). Routledge.

Randall, J. (2023). It Ain't Near 'Bout Fair: Re-envisioning the bias and sensitivity review process from a justice-oriented, antiracist perspective. *Educational Assessment*. Advance online publication. https://doi.org/10.1080/10627197.2023.2223924

Randall, J., Slomp, D., Poe, M., & Oliveri, M. E. (2022). Disrupting White supremacy in assessment: Toward a justice-oriented, antiracist validity framework. *Educational Assessment, 27*(2), 170–178. https://doi.org/10.1080/10627197.2022.2042682

Sato, E. (2025). Born socioculturally responsive assessment: An approach to design and development. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 115–143). Routledge.

Sireci, S. G., Crespo Cruz, E., Suárez-Álvarez, J., & Rodríguez Matos, G. (2025). Understanding UNDERSTANDardization research. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 415–433). Routledge.

Skerrett, A. S., Pacheco, M., Hinchman, K. A., Cervetti, G. N., Pearson, P. D., & Greenleaf, C. (2025). Socioculturally responsive large-scale assessment: The case of the 2026 NAEP reading framework. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 283–305). Routledge.

Solano-Flores, G., & Ruiz-Primo, M. A. (2025). Cultural validity: Conceptual and methodological considerations for socioculturally responsive assessment. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 50–75). Routledge.

Wang, Z., Sparks, J. R., Walker, M., O'Reilly, T., & Bruce, K. (2025). Group differences across scenario-based reading assessments: Examining the effects of culturally relevant test content. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 369–398). Routledge.

Welch, C., & Dunbar, S. (2025). Designing socioculturally responsive assessments within the mandated bookends of K–12 assessment systems. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 144–167). Routledge.

Zandvakili, E., & Gordon, E. W. (2025). Transforming educational assessment: Implementing equitable and inclusive strategies. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 99–112). Routledge.