Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 29 Number 9, June 2024 ISSN 1531-7714

A Culturally Enhanced Framework of Caring Assessments for Diverse Learners

Blair Lehman, *ETS*
Jesse R. Sparks, *ETS*
Diego Zapata-Rivera, *ETS*
Jonathan Steinberg, *EurkeaFacts, LLC*
Carol Forsyth, *ETS*

Most assessments adopt a one-size-fits-all approach to provide fair testing opportunities to all learners. However, this rigid approach to assessment may limit the ability for some learners to show what they know and can do. The Caring Assessments framework proposed a guide for the design and development of flexible, personalized, and adaptive assessments to provide each learner with the best opportunity to show what they know and can do. The original framework for caring assessments proposed that caring can be integrated into assessments by leveraging knowledge about learners' characteristics, behaviors, and learning context. Because we also recognize the critical role of acknowledging learners' cultures, identities, and social contexts to provide effective, caring support for all learners, in this paper we expand the framework to the Culturally Enhanced Caring Assessments framework to include personal, social, linguistic, and cultural aspects of learners and the contexts in which they learn. We discuss the culturally enhanced caring assessments framework and the need for further research to address the implementation challenges that can emerge when assessments are flexible, personalized, and adaptive.

Keywords: personalized assessment, formative assessment, culturally responsive assessment

Introduction

One-size-fits-all approaches to learning have long been dismissed as ineffective, especially by proponents of adaptive learning systems (ALS, VanLehn, 2011). However, one-size-fits-all approaches have largely persisted for educational assessments. These approaches have been adopted and maintained to promote fairness and accurate measurement through consistent, standardized tests and test administration conditions (i.e., marginal sense of fairness; Mislevy, 2018). However, recent examinations of standardized assessments through an equity-minded perspective have questioned the efficacy of standardization to promote fairness and accurate measurement (Sireci, 2020; Sireci & Randall, 2021). While recent calls to allow greater flexibility in assessment design and administration (Mislevy, 2018; Sireci, 2020) seem to suggest completely changing assessment practices, there already are administration practices that allow for personalization and adaptation in assessments. For example, there are modifications that provide learners with documented disabilities the appropriate accommodations to allow them to show what they know and can do (AERA et al., 2014), and there are computer adaptive tests (CATs) that adapt test features such as item difficulty based on the quality of learners' previous responses. Thus, recent calls to allow greater flexibility in assessments primarily seek to expand the ways in which personalization and adaptation are utilized in assessments (Bennett, 2023). This proposed expansion then raises the question: how should tests be designed to provide personalized, adaptive experiences for all learners?

Previously, we proposed caring assessments (Zapata-Rivera, 2017) as a framework to guide the design and development of assessments that are personalized and adaptive by leveraging learner characteristics (e.g., self-efficacy, interests), learner behaviors, and the affordances of technology-based assessments to integrate "caring" before, during, and after the assessment (Lehman et al., 2018). The support provided by the caring assessment (CA) framework is viewed as caring in that it considers the whole learner, rather than only their performance on the present or previous assessments. This framework leveraged prior research on ALSs that enable the delivery of personalized, adaptive experiences (du Boulay et al., 2010; Kay & McCalla, 2003; VanLehn, 2006; 2011; Weitekamp & Koedinger, 2023) and was consistent with the strengths of assessment identified in a recent analysis of the strengths, weaknesses, opportunities, and threats (SWOT) of assessment in ALSs (Zapata-Rivera & Hu, 2023). CA was developed in the context of formative, classroom-based assessments. However, we viewed it as applicable to all assessment contexts, with the caveat that not all methods of personalization and adaptation will be equally well-suited to all assessment contexts (Bennett, 2023).

The original CA framework sought to include many learner characteristics and the learning context as targets for adaptation, focusing on characteristics that were deemed more malleable (i.e., could be addressed with relatively brief interventions) such as prior knowledge, self-efficacy, and test anxiety (Abrahams et al., 2019; Duckworth & Yaeger, 2015). However, there was a noticeable gap with respect to consideration of learners' cultural identities and the social contexts in which they learn, both of which are critically important to understanding who learners are and how to deliver effective, caring support to them before, during, and after assessments (e.g., Gay, 2013; Ladson-Billings, 2014; Paris, 2012). A CA framework that does not include learners' cultural identities and social contexts may do a disservice to learners from marginalized groups and could be construed as pushing learners to conform to the standards set by the dominant culture, which in the U.S. privileges whiteness (Randall, 2021a). In the current paper, we describe a culturally enhanced caring assessments (CECA) framework that considers how learner characteristics and experiences intersect with their cultural identities and backgrounds. While not yet fully implemented, we present the research that inspired both the original and enhanced framework, the ongoing research efforts to develop the components necessary for full implementation, and the research that supports CECA's potential benefits for learners and their teachers. The remainder of this paper discusses relevant research for culturally relevant and responsive assessments, the components of CA and CECA, and challenges to implementing CECA.

Culturally Relevant and Responsive Assessments

Culturally relevant, responsive, and sustaining pedagogy seeks to create a learning environment in which all learners can feel respected, valued, and supported (Gay, 2010; Ladson-Billings, 2009, 2014; Paris, 2012). These pedagogical approaches provide learners from diverse groups the opportunity to leverage their deep cultural funds of knowledge throughout the learning process (González et al., 2005; Moll et al., 1992). Moll and colleagues characterize these funds of knowledge as "historically accumulated and culturally developed bodies of knowledge and skills essential for household or individual functioning," (p. 133) which may involve knowledge of a wide range of social, developmental, and labor-related practices (e.g., agricultural, scientific, economic). These funds have value in educational settings insofar as they can be leveraged to support instruction and assessment. For example, a learner who enjoys cooking at home may have experience investigating different variations of a recipe, which could be connected to the scientific process (Mills et al., 2019). Leveraging funds of knowledge has also been expanded to ALSs to increase the relevance of learning materials to promote deeper learning (Walkington & Bernacki, 2018). Increasing the relevance of learning materials for all learners enhances learning outcomes and can also enhance the learning experience. Creating a more welcoming learning experience for learners of diverse backgrounds is hypothesized to lead to greater engagement and feelings of belonging, both of which are critical for academic success (Fredricks et al., 2004).

Recently, there have been increasing calls to expand the benefits of culturally relevant, responsive, and sustaining pedagogy (e.g., Randall, 2021a) and personalized learning to assessments (Bennett, 2023; Zapata-Rivera, 2017). Researchers have suggested many approaches to creating culturally relevant and responsive assessments (e.g., Bennett, 2023; Hood, 1998; Lee, 1998; Qualls, 1998). Based on this prior research, we have identified four main areas that must be addressed to effectively design culturally relevant and responsive assessments that meet the needs of individual learners: involvement in the assessment development process, inclusion of context in assessment content, flexibility in the assessment experience, and framing of performance feedback. Next, we describe how each area contributes to the effective design of culturally relevant and responsive assessments.

Involvement in the Assessment Development Process

The assessment development process has often underrepresented or excluded the perspectives of marginalized groups, from defining what will be assessed to how the assessment will occur to how the test outcomes will be used (Randall, 2021a). Each of these decision points can influence whether an assessment is culturally relevant and responsive. The Hawaiian Language Immersion Program is an example of successfully involving diverse interest groups at each stage (Kūkea Shultz et al., 2019). Originally, state assessments that were translated into Hawaiian from English were utilized, but these assessments were viewed as unfair by community members. To address these issues, the Hawaii Department of Education worked with a team of community members to develop science standards that embraced the knowledge, skills, and understandings of Hawaiian immersion learners and then developed an assessment based on these culturally relevant standards. Inclusion of focal community members in the assessment development process can provide a more equitable sharing of power and confirms that all learners' social and cultural identities are legitimate and valued (Gutiérrez, 2012; O'Dwyer et al., 2023; Walker et al., 2023). It is important to note that learners should also be included in the assessment development process to support designs that are centered on learners' experiences with the assessment (Araneda & Sireci, 2021). Inclusion of diverse voices in the assessment development process can also positively impact learners' wider academic engagement by increasing their feelings of belonging as they see themselves represented on assessments.

Inclusion of Context in Assessment Content

A criticism of standardized assessments is that they attempt to create content that is context-free by removing any references to specific identities, cultures, or social perspectives to develop assessments that are fair for all learners, but in practice this may result in contexts that are most relevant to the dominant culture rather than truly context-free materials (e.g., Montenegro & Jankowski, 2017; Randall, 2021a). However, even when assessment development practices are successful at removing context from assessment content, it is likely this will lead to assessments that do not promote meaningful engagement for any learners due to decontextualization (Cordova & Lepper, 1996) and the difficulty for learners to connect with the assessment materials can lead to low engagement (Wise & Smith, 2011). Engagement is an important consideration for all assessments because low engagement has been found to negatively impact the validity of assessment outcomes by limiting learners' ability to demonstrate what they know and can do, which can result in lower test performance (Wise & Smith, 2016). This potential for low engagement across all learners suggests that context-free assessments may not be advantageous for supporting inferences about learners' knowledge, skills, and abilities (KSA).

However, even if context-free assessments could be designed to be sufficiently engaging, there is still the issue of whether or not context-free assessments are truly devoid of contextual information. As mentioned earlier, context-free assessments may in practice only remove cultural contexts from groups not aligned with the dominant culture (e.g., promotion of whiteness in the U.S.), as suggested by some critics of standardized assessments (e.g., Randall, 2021a). Context-rich items, on the other hand, embrace various cultural identities and practices by contextualizing items within explicit cultural practices and/or social justice issues and movements (Randall, 2021a). An example is an item designed to assess learners' abilities to draw conclusions from data presented in a bar chart (Randall, 2021b), in which learners are presented with real data showing the discrepancy in the proportion of teachers from different racial and ethnic groups compared to learners from those same groups. This item was designed to serve the dual purpose of assessing learners' mathematics KSAs and informing them about an important social issue, representation issues in the teacher workforce. So-called context-free content in assessments is at best limiting all learners' ability to perform their best and at worst is specifically disadvantaging learners from marginalized groups (O'Dwyer et al., 2023). It is therefore critical to include context-rich content that represents diverse lived experiences, cultural identities, and social contexts to engage learners and facilitate outcomes that more accurately represent what they know and can do.

Flexibility in the Assessment Experience

Another criticism of standardized assessments is that standardization does not provide all learners with their best opportunity to display their KSAs (Mislevy, 2018; Sireci, 2020). However, as mentioned previously, there are currently instances of flexibility in standardized assessments. Test accommodations represent "minor changes" that are employed for learners with documented disabilities that maintain the construct being assessed and allow for score comparability across learners (AERA et al., 2014; Steinberg et al., 2011). Some test accommodations are enacted to make the assessment content accessible to learners with documented disabilities, but this could be expanded by considering other ways that content can be less accessible for some learners than others. For example, items that reference specific cultural practices less familiar to some learners could reduce accessibility by increasing cognitive load. Personalizing the context, or in this example the cultural practices included in the item, could make the item more accessible to learners from different cultural backgrounds while maintaining the construct and score comparability. Maintaining the construct and score comparability are important points related to flexibility because culturally relevant and responsive assessments should maintain high expectations for all learners to counteract negative stereotypes of the academic abilities of learners from marginalized groups (Gay, 2010; Rojas & Liou, 2017; Walker et al., 2023) and to not inhibit learners' abilities to develop the KSAs believed to be necessary for success.

Flexibility also exists in current assessment practices in the form of CATs that adapt item administration based on learner performance (van der Linden & Glas, 2010). CATs are designed to present individual learners with items likely to maximize the measurement information collected from them based on performance. However, when learners are not fully engaged, their performance may not fully represent what they know and can do (Wise & Smith, 2016). To maximize the measurement information collected, adapting based on learner engagement to provide items more likely to elicit engaged responses from learners could be advantageous (e.g., adapting item context or format). By expanding how and when we apply the principles behind test accommodations and CATs to consider other learner characteristics (e.g., cultural identity, interests) and behaviors (e.g., engagement), a more personalized testing experience can be provided.

The flexibility discussed so far generally focuses on decisions made by assessment administrators and developers. However, it is also important to provide learners with the opportunity to make choices during assessments as choice can enhance the testing experience for learners (Bridgeman et al., 1997; Pitkin & Vispoel, 2001). The incorporation of learner choice into assessments may particularly benefit learners from marginalized groups by promoting feelings of autonomy (control over one's behavior) and competence (confidence when navigating one's environment), both of which are key determinants of motivation (Ryan & Deci, 2019), but can be negatively impacted by systemic inequalities, negative stereotypes, and prior negative experiences (Lewis & Hunt, 2019; Rojas & Liou, 2017). Prior research on the use of choice to promote feelings of autonomy and agency in ALSs has shown that the nature of choice should be personalized to maximize the benefits for each learner (Brod et al., 2023). This finding further highlights the need for flexibility even within tasks already personalized and for further research to better understand what choices will be most beneficial to individual learners.

While there are potential benefits of incorporating choice into assessments, it is important to note that learners do not always make the most advantageous choices and that choice does not guarantee improved testing experience or performance (Bennett, 2023; Pitkin & Vispoel, 2001; Powers & Bennett, 1999; Wainer & Thissen, 1994). This variation in outcomes can be partially attributed to the diversity in how choice can be implemented in an assessment. One example is providing choice for the context within which a scenario-based assessment (SBA) is presented based on a learner's more general interest in environmental issues, such as organic farming, hydropower, or wind power (Sabatini et al., 2014). In this example, each version of the SBA would assess the same underlying construct and the scores from each version would be comparable, while still allowing the learner to select a version that is most relevant to the environmental issues impacting their community. Thus, the choice would be at a more surface level because the choices would be viewed as interchangeable from a measurement perspective (Wainer & Thissen, 1994).

Another example can be seen in the Advanced Placement™ Art and Design Program where learners engage in year-long projects to develop a portfolio of artifacts and a portfolio documenting the process of creating those artifacts under the mentorship of a teacher (College Board, 2021). Learners are given complete choice over the artifacts they develop in terms of the materials used and the topics explored. This type of choice allows for learners to determine what is most relevant to them and how they want to incorporate (or not) their cultural backgrounds into the assessment (Bennett, 2023). Although the choice presented to learners in this example does not impact the measurement properties or the target construct of the assessment, this could be viewed as a deeper level of choice because learners generate their own unique testing experiences (Wainer & Thissen, 1994). Based on just these two examples, how choice is incorporated into assessments is a complex design decision and more research is needed to better understand the conditions under which choice is beneficial for learners from diverse backgrounds in different assessment contexts.

Framing of Performance Feedback

The nature of the language used to communicate performance outcomes is of the utmost importance. For example, many standardized assessments utilize labels for learners at different achievement levels that can employ deficit-based language (e.g., fail, below satisfactory; O'Donnell & Sireci, 2021), which can have a negative impact on learners' motivation to engage in academic activities and feelings of belonging (Ryan & Deci, 2019). In contrast, culturally relevant and responsive assessments should employ asset-based (or strengths-based) language (Gay, 2013) when providing performance feedback. For example, an asset-based approach would avoid terms such as "at-risk" or "vulnerable" when describing a learner's current or ongoing performance and could instead use the terms "at-promise" (Swadener, 2000) or "can do" (WIDA, 2020) to highlight the learner's potential for growth and progress. The asset-based approach to providing feedback would highlight what learners know and can do and recognize the variety of ways in which learners demonstrate their KSAs (Ramasubramanian et al., 2021).

Conceptual scoring involves adopting an asset-based approach where the meaning or content of a response is scored regardless of the language or dialect in which the response is provided, and results in performance feedback recognizing the knowledge that multilingual learners can display across languages (Guzman-Orth et al., 2019). It is important to note that the goal of asset-based feedback would not be to change the target construct of the assessment or to only provide positive feedback, rather the goal would be to provide feedback that includes both areas of strength and potential areas for improvement for learners. For example, if the target construct was providing a synthesis of the main ideas in a text in English and the learner provided a high-quality synthesis in Spanish, the asset-based feedback could highlight that they clearly understood and communicated the main ideas of the text (strength) and have room to improve on their communication in English. Underestimating the KSAs of multilingual learners, for example, can limit their ability to access appropriate resources and can also lead to an increased potential for academic failure (Hamre & Pianta, 2005). Thus, it is important to provide feedback that highlights the diversity of knowledge that learners bring to the classroom to support their current and future learning.

Framework for Caring Assessments

In the original CA framework, we provided guidance for the design and development of adaptive assessment systems (AAS) that provide a personalized and adaptive testing experience integrated into classroom teaching and learning through various types of assessments (Zapata-Rivera, 2017). The AAS would leverage information about learners and their interactions with computer-based assessments to provide caring support before, during, and after the assessments (Lehman et al., 2018). Next, we describe how caring can occur before, during, and after the assessment in our original CA framework, followed by a description of how "caring" has been enhanced in the CECA framework.

Before the Assessment

Caring support *before* the assessment would involve modifications to the assessment based on learner characteristics that are known or can be measured before an assessment. This component of CA is similar to how ALSs address the cold-start problem where ALSs are initially unable to provide personalized, adaptive support to learners in their first interaction due to insufficient in-system data but can be addressed through integrating information about learners from outside of the system (e.g., prior knowledge assessment; Grubišic´ et al., 2013). CA originally proposed to focus on learners' prior knowledge and experiences, social-emotional skills, and personal qualities (Abrahams et al., 2019; Duckworth & Yaeger, 2015) as these were identified as malleable factors that could be addressed through brief interventions (e.g., messages promoting a growth mindset; Samuel et al., 2022) and likely to impact performance. We conducted a series of experimental studies where a wide variety of learner characteristics were assessed via self-report measures administered to diverse learner populations to investigate the generalizability of learner profiles across assessments with different domains and formats (Sparks et al., 2019; Sparks et al., 2022). For example, our prior research involving interactions with conversation-based assessments identified opportunity to learn science, science self-efficacy, cognitive flexibility, growth mindset, and test anxiety as significant predictors of performance (Sparks et al., 2019). While we did include more stable background characteristics (e.g., race and ethnicity, gender, socioeconomic status) in our statistical analyses, they were previously not a focus for guiding adaptation.

We quickly encountered two challenges when considering a greater variety of learner characteristics to guide assessment development. First, more learner characteristics make it challenging to determine a clear direction for development. Second, it is impractical to consume multiple hours of class time to administer a large battery of measures. To address both challenges, we adopted a learner-centered clustering approach allowing for the creation of meaningful profiles of learner characteristics that co-occur (e.g., high test anxiety + low self-efficacy + high growth mindset), thereby reducing the number of measures needed for assigning profiles to future learners. This learner-centered clustering approach builds on so-called stereotype approaches leveraged in learner modeling. Although the name may be misleading, stereotype approaches are a class of learner modeling approaches that do not advocate societal stereotypes but rather rely on information collected prior to the first system interaction (which may include many variables beyond demographics) to assign learners to a particular group (Grubišic´ et al., 2013; Kay, 2000). Although stereotype approaches to modeling in adaptive systems are generally not limited to a single characteristic (Rich, 1979; Wilensky et al., 1988), such approaches often rely on single characteristics when applied in educational contexts (e.g., prior knowledge; Grubišic´ et al., 2013).

In CA we sought to build on these types of approaches by incorporating many characteristics (e.g., prior opportunities to learn, self-efficacy) and identifying those most relevant to learners' experience with and performance on assessments to create learner profiles. Learner profiles can then be leveraged to identify test characteristics likely to give each learner their best opportunity to show what they know and can do. Thus, the interaction between the learner and test characteristics must be considered to decide how best to provide caring support to each learner. These profiles can then provide initial guidance for modifying assessments to provide the best opportunity for each learner and can route learners in the AAS to the appropriate version of an assessment given their profile (Khayi & Rus, 2019). However, it is important to note that these profiles are malleable and given that CA has focused on more malleable learner characteristics, we would expect profiles to change over time and could be modified based on feedback from learners and teachers.

Modifications to the assessment based on learner characteristics can vary widely but should always aim to optimize the assessment experience for the individual learner. One type of modification could be providing tailored messages to learners at the beginning of the assessment that, for example, address test anxiety through motivational statements to maximize the cognitive resources that learners can devote to the assessment (Verschelden, 2017). For example, a recent investigation into how to better support first year college students utilized a similar learner-centered clustering approach and identified that stress management resources should be developed that resonate with Hispanic/Latinx learners at that college to better support this group of learners (Ludvik et al., 2022). Another type of modification would be to alter the item or assessment format. Prior research has shown that the same construct can be measured using a multiple-choice or constructed-response format (Rodriguez, 2003) and that item format can impact engagement (DeMars, 2000). This finding has been more recently replicated when comparing game-based and multiple-choice assessments measuring the same construct (Lehman et al., 2019). CA sought to promote both maximal performance and an engaging, positive experience as learners are typically able to perform to the best of their abilities when they are more engaged (Finn, 2015; Wise & Smith, 2016). All test modifications should be developed to target the needs of specific learner profiles, with input from learners and teachers to guide the design and evaluation of these modifications.

During the Assessment

Although caring support *before* the assessment would aim to optimize the experience of learners, it is likely that at least some learners will still experience struggles and disengagement. This could be due to either potential issues in learner profile assignment, which could result in the caring support before the assessment not being appropriate for a particular learner, or encountering items that learners are unsure how to solve due to the difficulty level, item presentation, or a combination of these two item characteristics. The goal of CA was not to remove all instances of challenge for learners, but rather to present appropriately rigorous content in an environment that provides appropriate supports to create and maintain a safe (students are respected and valued), supportive environment for learners (Gay, 2010). Instances of challenge can lead to unproductive states (Lehman & Zapata-Rivera, 2018), so on-demand support is needed *during* the assessment to maintain a safe, supportive, and engaging environment. CA would not guarantee all learners perform at the desired level on an assessment, but rather would provide learners with their best opportunity to show what they know and can do to understand the current state of their learning and identify areas of opportunity for future learning and improvement.

The two important considerations for on-demand support were when to provide it and how to deliver it. Both issues can be addressed through integrated learner modeling in the AAS (Zapata-Rivera et al., 2020). Learner models are employed by ALSs to maintain a representation of the learner as they progress through the system and guide the deployment of new tasks and supports by the recommendation system (Shute & Zapata-Rivera, 2012). Traditionally, learner models have focused on learners' mastery of target KSAs, but in CA we proposed enhancements to include more elements of the learners' experience and identity, such as cognition, metacognition, emotion, and learning context, resulting in a learner model that better integrates the learner's experiences and personal characteristics. Learner profiles would be used as initial information for the integrated learner model and then moment-by-moment interaction data collected during the assessment would be used to further refine the model.

The use of interaction data was guided by prior research on ALSs in which learner behaviors such as gaming the system (Baker et al., 2008), wheel-spinning (Beck & Gong, 2013), and emotions (Pardos et al., 2014) are detected within digital environments. ALSs that provide adaptive feedback based on a combination of learners' cognitive and emotional states have found positive impacts on learning, particularly for learners with lower prior knowledge (D'Mello et al., 2011; Forbes-Riley & Litman, 2011). The use of on-demand messages and proctor intervention when disengagement occurs in low-stakes assessments have also had positive impacts on engagement and assessment performance (Wise et al., 2006; Wise et al., 2019). CA could also expand on the types of interaction data that have been previously used in assessment and learning systems (Zapata-Rivera et al., 2023). For example, linguistic information from open-ended learner responses (written or spoken) could be utilized to better understand learners' experience with the assessment. The Linguistic Inquiry and Word Count (LIWC) tool utilizes a bag of words approach to analyze language (Boyd et al., 2022) and has been used to identify various learner emotions (e.g., boredom, confusion, frustration) during interactions with expert tutors (Lehman & D'Mello, 2010). The identification of these additional emotional experiences could then be incorporated into the integrated learner model. As with assessment modifications, on-demand support should vary in its implementation based on learner profile membership and the current state of the integrated learner model.

After the Assessment

Caring support provided to learners *after* the assessment focused on how performance feedback is delivered to learners and teachers. Learner profiles could be used to determine how to present feedback in a manner that is motivating to the learner, whereas interaction data could be used to contextualize feedback. For example, consider a learner who is building their math self-efficacy and completes a math assessment they did not find particularly interesting. This low interest could cause the learner to rush through the assessment as quickly as possible (i.e., low engagement), which could result in them earning a low score that does not accurately reflect their math knowledge. This example learner may not benefit from only receiving an overall score about their (potentially low scoring) performance. The feedback could be augmented to highlight areas in which the learner demonstrated strengths to support the development of their self-efficacy, indicate the ways in which the learner's behaviors (e.g., engagement) impacted their performance, and include the opportunity for the learner to provide feedback about their interest in the assessment content, which could be leveraged to guide future assessment development. Providing learners opportunities to offer feedback that can be leveraged for future development is consistent with an asset-based approach as it assumes that all learners are capable and want to succeed. As with caring support *before* and *during* the assessment, caring support *after* the assessment would be implemented in various ways to be personalized to the learner.

CA proposed to also provide teachers with enhanced feedback to support their instructional decisions and practices. This enhanced feedback would provide teachers with needed information to support learners' active engagement in learning as it is necessary for teachers to know each learner's current level of mastery as well as a variety of their characteristics well enough to identify their zones of proximal development to provide appropriately rigorous material (Vygotsky, 1978) and appropriate supports (Wellberg & Evans, 2022). Learner profiles can be utilized to provide teachers with information about learner characteristics that teachers may not know yet (e.g., at the beginning of the school year) or are perhaps not easily accessible in a typical school environment (Zapata-Rivera, 2021). Interaction data could be used to contextualize assessment performance (e.g., the learner had lower performance but was also highly disengaged, so this assessment result may not fully represent the learner's KSAs) or to highlight learners' struggles (e.g., spending too much or too little time on a certain question, text, or part of the task). This could be done *during* the assessment (e.g., Wise et al., 2019) or *after* the assessment to allow for the more nuanced support an AAS is not able to provide, but with which teachers are well versed. For example, teachers could recognize that a learner's disengagement was atypical and converse with the learner about what events in their life may be impacting their ability to concentrate on school tasks (Zapata-Rivera et al., 2018). An important consideration is how best to provide teachers with this enhanced feedback to support their instructional decision making, which will require collaboration with teachers to determine the best solution.

Culturally Enhanced Caring Assessments Framework

We recognize that the previous decision to emphasize more malleable characteristics (e.g., growth mindset, self-efficacy, test anxiety) over more stable learner characteristics (e.g., cultural identity, social context) when designing and developing caring supports in the original CA framework limits its ability to support all learners. Appropriate supports to create a safe, supportive environment for learners can only be achieved by acknowledging and celebrating learners' cultural backgrounds (Ramasubramanian et al., 2021) and leveraging those cultural funds of knowledge (González et al., 2005; Moll et al., 1992) to develop more effective caring supports. Thus, the culturally enhanced CA (CECA) framework more explicitly incorporates information about learners' cultural identities and social contexts into the development of learner profiles and more explicitly focuses on development of culturally relevant and responsive caring supports. We believe that the intersection of malleable and stable learner characteristics will allow for more effective, personalized support in CECA and help avoid using overly simplified or stereotypic representations of learners when developing assessments. Next, we reflect on the enhancements that have resulted in CECA to facilitate the development of culturally relevant and responsive assessments.

Involvement in the Assessment Development Process

It is important to explicitly outline the ways in which interest holders should be involved in the development of CECA. We view interest holders as the users of the assessment (whether it is those who are taking the assessment or using the assessment results to support learning), which can include learners, teachers, and community members from diverse backgrounds and perspectives, with a particular focus on those from marginalized groups previously underrepresented in or excluded from the assessment development process. Recruitment should particularly prioritize the recruitment of interest holders from Black, Brown, Indigenous, and People of Color (BBIPOC) communities as their voices, traditions, and values are often neglected in the assessment development process. It is also important to consider diversity within BBIPOC communities when recruiting interest holders to incorporate their rich and varied experiences. Recruitment of interest holders from marginalized groups should prioritize those with the lived and/or work experiences relevant to the assessment development process, which can include a variety of experiences beyond actual assessment development (e.g., parent of a child in the school system, community organizer). The particular focus on marginalized groups is critical to ensure all stages of the assessment development process promote equitable representation and cultural relevance (Hood, 1998).

CECA should include interest holders at each stage, not just evaluating materials already developed. Learner profiles, for example, are developed through interpreting profile analysis results, which is a subjective meaning-making activity. To ensure that the interpretations are as meaningful as possible, we propose reviewing relevant theories (Forsyth et al., 2014; 2020) and collaborating with diverse interest holder groups to identify context- or culture-specific interpretations. Interest holders can also be involved in developing assessment modifications, on-demand support, and feedback delivery through co-design practices in which interest holders and researchers collaborate in the design process (e.g., Penuel, 2019). To be most effective, such co-design practices should involve researchers and interest holders in sustained, long-term partnerships with an iterative design process that develops participants' capacities for engaging in continuous cycles of improvement (Penuel, 2019).

When engaging in sustained, long-term partnerships with diverse interest holders, it is also important to provide appropriate training and support to foster the mindsets necessary to develop assessments that are equitable for all learners. This training and support would be necessary for all development team members as we all exist within a society that has prioritized some experiences over others (e.g., promotion of whiteness in the U.S.; Lysicott, 2019). Training and ongoing support that raises awareness and facilitates the shift in mindsets across the entire team will be needed to maximize the benefits of diverse voices on the team and to maintain a justice-oriented approach to assessment development (Randall, 2021a).

Inclusion of Context in Assessment Content

CECA should explicitly call for context-rich assessment content that represents and celebrates diverse cultures, lived experiences, and perspectives and not adopt so-called context-free development practices (O'Dwyer et al., 2023). Assessment modifications can be expanded to include the context in which assessment items are embedded. Tasks viewed as congruent with learners' cultural identities, beliefs, and values can elicit positive outcomes similar to intrinsically motivating tasks (Ryan & Deci, 2000) and increase feelings of belonging in academic environments (Fredricks et al., 2004). The inclusion of learners throughout the development process can allow for centering learners and their experiences with the assessment (Araneda & Sireci, 2021) and leveraging learners' funds of knowledge and identities that can help identify meaningful and engaging contexts for learners (Esteban-Guitart, 2021).

Flexibility in the Assessment Experience

CECA should provide more learner agency in deploying different assessment versions (potential measurement issues are discussed in the Challenges section). In addition to increasing feelings of autonomy, this modification will also help address potential alignment issues between individual learners and their associated learner profiles. Learners will vary in the degree to which they align with a specific learner profile, which means that some learners will be very close to the prototypical learner for a given profile, whereas others may align less well to the profile (either due to general lack of fit among available profiles, or a potential misclassification). For learners who align less well to the profile, assignment to an assessment version designed for the prototypical learner within that profile may be less advantageous. CECA will provide learners with multiple recommended assessment versions with indicators of how these are likely to give them their best performance opportunity (e.g., learners with similar interests to you found this version to be 30% more engaging) and assessment versions that the AAS did not recommend for them. Providing this type of context for learner choice is one method to reduce issues related to learners making less advantageous choices (Pitkin & Vispoel, 2001; Wain & Thissen, 1994). Learner choice can also then be leveraged to refine the integrated learner model by understanding how choices that learners have made relate to their characteristics, experience with the assessment, and ultimate performance on the assessment.

CECA further seeks to optimize learner profile assignment and enhance the use of flexibility in assessments through the inclusion of learners' cultural identities and social context. This addition to the learner profile results in a focus on four categories of learner characteristics: personal (e.g., self-efficacy, test anxiety), social (e.g., feelings of belonging, interaction styles), linguistic (e.g., English language proficiency, preferred language), and cultural (e.g., cultural identity, cultural relevance of topic). Given the intended use of these learner profiles and that learners will be grouped based on a relatively limited number of characteristics, it is important these profiles embrace the complexity and intersectionality of learners' identities (Cole, 2009). For example, we would not want learner profiles only developed based on demographic information to guide assessment development, which could result in stereotypes that assume monolithic experiences, preferences, and identities within a particular demographic group. Rather, by including a variety of characteristics (personal, social, cultural, linguistic), we hope to identify profiles that represent the diversity and nuance of characteristics and prior experiences learners bring to the classroom to address the cold-start problem without relying on potentially overly simplified representations of learners.

Flexibility of response format is a principle of universal design for learning (Rose, 2000) that will be incorporated into CECA. Learners from different cultural backgrounds may be able to express their knowledge and ideas more confidently and effectively in different response formats (e.g., spoken, typed, home language; e.g., Abedi, 2010; Solano-Flores, 2008). Some assessments with constructed-response or essay items in the U.S. (implicitly or explicitly) require learners to respond in Mainstream White English to receive full credit for their responses. Providing Black learners, for example, with the option to respond in African American English would allow for a more accurate assessment of their KSAs while also promoting the value of their culture (Baker-Bell, 2020; Randall et al., 2021). This flexibility can allow for assessments to be more culturally responsive and provide learners with greater autonomy during the testing experience. However, it is important to note that this flexibility will also require revision of rubrics, recruitment of diverse item scorers, and enhanced training for item scorers to ensure diverse responses are scored in a fair and unbiased manner in line with conceptual scoring (Guzman-Orth et al., 2019). Although these updates to the current system may sound daunting, at their core, these updates would only require the focus to be placed on the content of learner responses and not on the format or structure (when those are not part of the standard or construct).

Framing of Performance Feedback

CA sought to personalize the framing of feedback and to contextualize performance feedback based on learner profiles and their interactions with the assessment (e.g., engagement), to provide feedback that is easy for learners to understand and motivates learners to continue their learning journeys. This approach aligns with the asset-based approach but can be improved in CECA by the inclusion of learners' cultural identities and social contexts to enhance the personalization of feedback.

Learners' cultural identities can also be integrated into the feedback reports provided to teachers to support their use of culturally responsive teaching practices in the classroom. In recent co-design sessions with middle and high school STEM teachers, we found that information about learners beyond their performance on educational activities (e.g., identity, interests, culture, values) and their perceptions of those activities (e.g., relevance to learners' lived experiences, difficulty, engagement) was viewed as helpful to support teachers to engage in more culturally responsive teaching practices (Lehman et al., 2023). Thus, the broader framing of feedback in the context of who learners are and how they interacted with the assessment can benefit both learners and teachers.

Challenges

The proposed CECA framework is a complex interaction of many components, which can result in many challenges for its implementation (Zapata-Rivera et al., 2020). We have categorized these challenges into four groups: design and evaluation of the AAS, unintended consequences of enhanced feedback, data privacy concerns, and current assessment development and measurement practices.

Design and Evaluation of the AAS

One challenge for the design of an AAS is the development of personalized content. An AAS will require the development of a large bank of items and tasks, similar to CATs but with the added development requirement of items and tasks that vary based on multiple learner characteristics and behaviors. Future CECA research will require explorations of how AI capabilities can be leveraged to reduce the time and cost of item and task development, while continuing to provide nuanced and personalized caring support to the learner. Another challenge for the design of an AAS is how the integrated learner model is used to deploy adaptive content. Given the nature of the integrated learner model, this will require coordinating multiple inputs while also regularly monitoring the AAS to avoid algorithmic bias (Noble, 2018). The use of open learner models (Bull & Kay, 2010) is one method to avoid algorithmic bias by making the learner model visible to learners and teachers.

Evaluation of an AAS will be challenging, as current guidelines for evaluating the effectiveness of assessments (e.g., AERA et al., 2014) were not created with personalized, adaptive testing scenarios, cultural responsiveness, or equity-minded perspectives in mind (Poe et al., 2023; Randall et al., 2024). To support the hopefully wider adoption of personalized and/or culturally relevant and responsive assessments, revisions to assessment guidelines and standards will be necessary to account for this paradigm shift in how assessments can be designed, developed, administered, and scored. Additional elements to be evaluated will include the appropriateness of learner profiles, accuracy of the integrated learner model, effectiveness of adaptive support, and impact of feedback provided. Thus, it will be necessary to not only build an AAS but to also determine the best methods to evaluate its effectiveness.

Unintended Consequences of Enhanced Feedback

The inclusion of learners' cultural identities, and the learner profiles more broadly, could lead to the unintended consequence of enhanced feedback being used to perpetuate negative stereotypes of learners from marginalized groups. In the context of K-12 education, approximately 79% of U.S. public school teachers identified as non-Hispanic White in 2017-18 (Taie & Goldring, 2020), whereas learner enrollment in fall 2018 was 1% American Indian/Alaska Native, 5% Asian, 15% Black, 27% Hispanic, less than 1% Pacific Islander, 47% White, and 4% of learners who reported two or more races (National Center for Education Statistics, 2022). Given these distributions, it is likely that teachers will be providing instruction to learners from cultural backgrounds different from their own and will need active and ongoing support to utilize knowledge of learners' cultural identities and social contexts in manners that promote learning and employ culturally relevant, responsive, and sustaining pedagogical practices. In addition to providing training and resources to support teacher implementation of culturally relevant, responsive, and sustaining pedagogical practices, it is also important to collaborate with teachers to cultivate their beliefs in the value of these practices to ultimately achieve successful implementation (McLaughlin, 1990; Wellberg & Evans, 2022).

Another method to address potential unintended consequences of enhanced feedback is to provide learners with the opportunity to decide not only what information they are comfortable sharing, but with whom they are comfortable sharing it. An example could be seen in asking learners to provide information about their pronouns (or gender identity). Some learners may be comfortable sharing this personal information with the AAS to receive a personalized experience (e.g., pedagogical agents within the assessment use the correct pronouns when referring to the learner) but may not want to have this shared with their teacher (e.g., due to school or district policies, lack of trust). Providing learners with the agency to decide what information to share with whom could also address some data privacy concerns that are discussed next.

Data Privacy Concerns

The requirement of collecting large amounts of learner data for the implementation of an AAS and the nature of that data (e.g., learner identities, social contexts) can lead to potential data privacy concerns. Issues of data security exist across many fields, and the challenge for CECA will be to determine (a) the infrastructure needed to securely store this potentially sensitive data, (b) who will have access to the data, and (c) how the data may or may not be used outside of the AAS. Efforts to develop ethical standards for AI and data use in ALSs can serve as a guide for AAS development (e.g., Knight et al., 2023; Regulation 2016/679). In addition to adhering to standards and guidelines, the utilization of a co-design methodology that involves teachers, learners, and other important interest holders can help develop trust and identify the specific needs and concerns that CECA will need to address during the development process. Addressing these challenges will be critically important, because if interest holders do not feel that they can trust the infrastructure and governance of the AAS and use of learner data, it is unlikely that CECA will be implemented or widely used.

Current Assessment Development and Measurement Practices

Lastly, CECA raises two issues that may conflict with current assessment development and measurement practices. First, current practices to ensure fairness in assessments seek to standardize the testing experience and limit the inclusion of context in assessment content (AERA et al., 2014), which is inconsistent with the need to meet learners where they are and to present culturally relevant content situated within meaningful contexts (Poe et al., 2023). Second, flexible assessments (e.g., assessment versioning, on-demand support) present challenges for current models of educational measurement to the extent that learners experience different versions of assessment tasks with differing degrees of support. For example, can the same construct be measured by assessments presented in different formats (e.g., game-based vs. multiple-choice items) and that provide different types of support based on needs of the individual learners? This raises issues around the validity of personalized, adaptive assessments that must be addressed. Conditional fairness (Mislevy, 2018) can provide an initial guide for addressing this issue of validity by emphasizing the importance of "equivalent evidence" being collected for the target construct across different assessment versions rather than equivalent surface features (e.g., item format, context; Mislevy et al., 2013). It is also important to note that when considering validity and other measurement issues, the assessment context must be considered. It may be that some elements of CECA will be better suited for different contexts. Formative, classroom-based assessments, for example, are likely to have different comparability or standardization needs than a large-scale, standardized assessment, and thus may have greater flexibility to implement personalization (Bennett, 2023).

Accurate measurement will remain critically important in CECA, but it is also critical to not force these new assessments into previous measurement models not developed for AASs. It may be necessary to reconceptualize (and re-operationalize) foundational educational measurement practices such as validity, reliability, and comparability in the context of AASs (Lederman, 2023; Randall et al., 2024; Sinharay & Johnson, 2024). Innovation will be needed in the design, development, and measurement models utilized for CECA to maximize the likelihood of providing all learners with their best opportunity to show what they know and can do.

Concluding Remarks

Several components of CECA are being actively explored, including efforts to understand how best to establish learner profiles (Sparks et al., 2019; 2022), leverage linguistic data to establish learner profiles (Forsyth et al., 2022), modify assessment versions (Lehman et al., 2019), provide on-demand support (Lehman & Zapata-Rivera, 2018), and provide supportive feedback (Zapata-Rivera et al., 2018; 2023). These ongoing research efforts will be guided by culturally relevant, responsive, and sustaining pedagogical strategies (Gay, 2010; Ladson-Billings, 2009, 2014; Paris, 2012), and will include regular engagement with interest holders from marginalized groups that have been underrepresented in or excluded from the assessment development process. Inclusion of diverse interest holders must be deliberate and ideally should be ongoing, long-term collaborations grounded in understanding and addressing the needs of practitioners and their learners.

References

Abedi, J. (2010). Performance assessments for English language learners. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Abrahams, L., Pancorbo, G., Santos, D., John, O. P., Primi, R., Kyllonen, P., & De Fruyt, F. (2019). Social-emotional skill assessment in children and adolescents: Advances and challenges in personality, clinical, and education contexts. *Psychological Assessment, 31*, 460-473.

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

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

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research*, *19*(2), 185-224.

Baker-Bell, A. (2020). We been knowin: Toward an antiracist language & literacy education. *Journal of Language and Literacy Education, 16*(1), 1-12.

Beck, J. E., & Gong, Y. (2013). Wheel-spinning: Students who fail to master a skill. In K. Yacef, C. Lane, J. Mostow, & P. Pavlik (Eds.), *Proceedings of 16th International Conference on Artificial Intelligence in Education (AIED2013)* (pp. 431-440). Berlin Heidelberg: Springer-Verlag.

Bennett, R. E. (2023). Toward a theory of socioculturally responsive assessment. *Educational Assessment*, *28*(2), 83-104.

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W., (2022). *The development and psychometric properties of LIWC-22*. The University of Texas at Austin.

Bridgeman, B., Morgan, R., & Wang, M. (1997). Choice Among Essay Topics: Impact on Performance and Validity. *Journal of Educational Measurement, 34*(3), 273-286.

Brod, G., Kucirkova, N., Shepherd, J., Jolles, D., & Molenaar, I. (2023). Agency in educational technology: Interdisciplinary perspectives and implications for learning design. *Educational Psychology Review*, *35*:25.

Bull, S., & Kay, J. (2010). Open learner models. In R. Nkambou, J. Bourdeau, & R. Mizoguchi (Eds.), *Advances in Intelligent Tutoring Systems* (pp. 301-322). Berlin Heidelberger: Springer.

Cole, E. R. (2009). Intersectionality and research in psychology. *American Psychologist*, *64*(3), 170-180.

College Board. (2021). *AP 2-D Art and Design, 3-D Art and Design, Drawing: Course and exam description.* https://apstudents.collegeboard.org/courses/ap-3-d-art-and-design/assessment

Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, *88*(4), 715-730.

D'Mello, S. K., Lehman, B. A., & Graesser, A. C. (2011). A motivationally supportive affect-sensitive AutoTutor. In R.A. Calvo & S.K. D'Mello (Eds.), *New perspectives on affect and learning technologies* (pp. 113-126). New York: Springer.

DeMars, C. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*, 55–77.

du Boulay, B., Avramides, K., Luckin, R., Martinuz-Miron, E., Rebolledo Mendez, G., & Carr, A. (2010). Towards systems that care: a conceptual framework based on Motivation, Metacognition and Affect. *International Journal of Artificial Intelligence in Education*, *20*, 197–229.

Duckworth, A. l., & Yaeger, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher, 44*(4), 237-251.

Esteban-Guitart, M. (2021). Advancing the funds of identity theory: a critical and unfinished dialogue. *Mind, Culture, and Activity, 28*(2), 169-179.

Finn, B. (2015). *Measuring motivation in low-stakes assessments.* (ETS Research Report No. RR-15-19). Princeton, NJ: Educational Testing Service.

Forbes-Riley, K., & Litman, D. (2011). Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication*, *53*, 1115-1136.

Forsyth, C. M. Andrews-Todd, J., & Steinberg, J. (2020). Are You Really a Team Player? Profiles of collaborative problem solvers in an online environment. In A. N. Rafferty, J. Whitehill, V. Cavalli-Sforza, & C. Romero (Eds.), *Proceedings of the 13th International Conference on Educational Data Mining* (*EDM2020*) (pp. 403-408). International Educational Data Mining Society.

Forsyth, C. M., Graesser, A. C., Pavlik, P., Millis, K., & Samei, B. (2014). Discovering theoretically grounded predictors of shallow vs. deep- level learning. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (*EDM2014*) (pp. 229-232). International Educational Data Mining Society.

Forsyth, C. M., Sparks, J. R., Steinberg, J. & McCulla, L. (2022). Linguistic profiles of students interacting with conversation-based assessment systems. In A. Mitrovic & N. Bosch (Eds.), *The Proceedings of the International Conference on Educational Data Mining* (*EDM2022*) (pp. 549-600). International Educational Data Mining Society.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59-109.

Gay, G. (2010). *Culturally responsive teaching: Theory, research, and practice* (2nd ed.). New York: Teachers College Press.

Gay, G. (2013) Teaching to and through cultural diversity. *Curriculum Inquiry, 43*(1), 48-70.

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

Grubis̆ic´, A., Stankov, S., Z̆itko, B. (2013). Stereotype student model for an adaptive e-learning system. *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, *7*(4), 440–447.

Gutiérrez, R. (2012). Context matters: How should we conceptualize equity in mathematics education? In B. Herbel-Eisenmann, J. Choppin, D. Wagner & D. Pimm (Eds.), *Equity in discourse for mathematics education: Theories, practices, and policies* (pp. 17-33). Springer.

Guzman-Orth, D., Lopez, A. A., & Tolentino, F. (2019). Exploring the use of a dual language assessment task to assess young English learners. *Language Assessment Quarterly, 16*(4–5), 447–463.

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development*, *76*(5), 949-967.

Hood, S. (1998). Culturally responsive performance-based assessment: Conceptual and psychometric considerations. *Journal of Negro Education, 67*(3), 187-196.

Kay, J. (2000). Stereotypes, student models and scrutability. In C. Cauthier & G. Van Frasson (Eds.), *Intelligent tutoring systems: 5th International Conference ITS 2000*, LNCS (Vol. 1839, pp. 19-30). Montreal, Canada: Springer.

Kay, J., & McCalla, G. (2003). The careful double vision of self. *International Journal of Artificial Intelligence and Education*, *13*, 1–18.

Khayi, N. A., & Rus, V. (2019). *Clustering students based on their prior knowledge.* Paper presented at the International Conference on Educational Data Mining (EDM). Montreal, Canada.

Knight, S., Shibani, A., & Buckingham Shum, S. (2023). A reflective design case of practical micro-ethics in learning analytics. *British Journal of Educational Technology*, *00*, 1-21.

Kūkea Shultz, P., Englert, K., Krug, K., Ruth, K., Ching, L., & Franco, L. (2019). *Context matters: The promise of cultural and community validity in assessment.* [Conference session]. Third Annual NCME Special Conference on Classroom Assessment, Boulder, CO.

Ladson-Billings, G. (2009). *The dreamkeepers: Successful teachers of African American children* (2nd ed.). San Francisco, CA: Jossey-Bass.

Ladson-Billings, G. (2014). Culturally relevant pedagogy 2.0: a.k.a. the Remix. *Harvard Educational Review, 84*(1), 74-84.

Lederman, J. (2023). Validity and racial justice in educational assessment. *Applied Measurement in Education*, *36*(3), 242-254.

Lee, C. D. (1998). Culturally responsive pedagogy and performance-based assessment. *Journal of Negro Education, 67*(3), 268-279.

Lehman, B., & D'Mello, S. K. (2010, August). *Predicting student affect through textual features during expert tutoring sessions.* Presented at the 20th Annual Meeting of the Society for Text and Discourse. Chicago, IL.

Lehman, B., Jackson, G. T., & Forsyth, C. (2019). A (mis)match analysis: Examining the alignment between test-taker performance in conventional and game-based assessments. *Journal of Applied Testing Technology, 20*, 17-34.

Lehman, B., Ober, T., Gooch, R., & Oluwalana, O. (2023). *Personalized learning as a method to achieve caring and culturally responsive assessments.* Paper presented at the annual meeting of the Center for Culturally Responsive Evaluation and Assessment (CREA). Chicago, IL.

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

Lehman, B., & Zapata-Rivera, D. (2018). Student emotions in conversation-based assessments. *IEEE Transactions on Learning Technologies, 11*(1), 1-13.

Lewis, E. L., & Hunt, B. (2019). High expectations: Increasing outcomes for Black students in urban schools. *Urban Education Research & Policy Annuals, 6*(2).

Ludvik, M. B., Zhang, S., Kahn, S., Potter, N., Richardson-Gates, L., Schellenberg, S. et al. (2022). Building intrapersonal competencies in the first-year experience: Utilizing random forest, cluster analysis, and linear regression to identify students' strengths and opportunities for institutional improvement. *Practical Assessment, Research, and Evaluation*, *27*, 21.

Lyiscott, J. (2019). *Black appetite. White food: Issues of race, voice, and justice within and beyond the classroom.* New York, NY: Routledge.

McLaughlin, M. W. (1990). The Rand Change Agent Study revisited: Macro perspectives and micro realities. *Educational Researcher*, *19*(9), 11-16.

Mills, K., Bonsignore, E., Clegg, T., Ahn, J., Yip, J., Pauw, D. et al. (2019). Connecting children's scientific funds of knowledge shared on social media to science concepts. *International Journal of Child-Computer Interaction*, *21*, 54-64.

Mislevy, R. J. (Ed.) (2018). A conditional sense of fairness. In *Sociocognitive Foundations of Educational Measurement* (pp. 218-244). New York, NY: Routledge.

Miselvy, R. J., Haertel, G., Cheng, B. H., Ructtinger, L., DeBarger, A., Murray, E. et al. (2013). A "conditional" sense of fairness in assessment. *Educational Research and Evaluation*, *19*(2-3), 121-140.

Moll, L. C., Amanti, C., Neff, D., & González, N. (1992). Funds of knowledge for teaching: Using a qualitative approach to connect homes and classrooms. *Theory Into Practice, 31*(2), 132-141.

Montenegro, E., & Jankowski, N. A. (2017). *Equity and assessment: Moving towards culturally responsive assessment.* (National Institute for Learning Outcomes Assessment Occasional paper #29).

National Center for Education Statistics. (2022). Racial/Ethnic Enrollment in Public Schools. *Condition of Education*. U.S. Department of Education, Institute of Education Sciences. Retrieved 05/06/2022.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York, NY: New York University Press.

O'Donnell, F., & Sireci, S. G. (2021). Language matters: Teacher and parent perceptions of achievement labels from educational tests. *Educational Assessment*, 1-26.

O'Dwyer, E., Sparks, J. R., & Nabors Oláh, L. (2023). Enacting a process for developing culturally relevant classroom assessments. *Applied Measurement in Education*, *3*, 286-303.

Poe, M., Oliveri, M. E., & Eilliot, N. (2023). The Standards will never be enough: A racial justice extension. *Applied Measurement in Education*, *3*, 193-215.

Pardos, Z. A., Baker, R. S., San Pedro, M. O. C. Z., Gowda, S. M., & Gowda, S. M. (2014). Affective states and state tests: Investigating how affect and engagement during the school year predict end of year learning outcomes. *Journal of Learning Analytics*, *1*(1), 107-128.

Paris, D. (2012). Culturally sustaining pedagogy: A needed change in stance, terminology, and practice. *Educational Researcher, 41*(3), 93-97.

Penuel, W. R. (2019). Co-design as infrastructuring with attention to power: Building collective capacity for equitable teaching and learning through design-based implementation research. In J. Pieters, J. Voogt, & N. P. Roblin (Eds.), *Collaborative Curriculum Design for Sustainable Innovation and Teacher Learning* (pp. 387-401). Cham, Switzerland: SpringerOpen.

Pitkin, A. K., & Vispoel, W. P. (2001). Differences between self-adapted and computerized adaptive tests: A meta-analysis. *Journal of Educational Measurement*, *38*, 235-247.

Powers, D. E., & Bennett, R. E. (1999). Effects of allowing examinees to select questions on a test of divergent thinking. *Applied Measurement in Education*, *12*, 257-279.

Qualls, A. L. (1998). Culturally responsive assessment: Development strategies and validity issues. *Journal of Negro Education, 67*(3), 296-301.

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

Randall, J. (2021a). "Color-neutral" is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice.*

Randall, J. (2021b). *The road to assessment hell is paved with color-blind intentions: Moving towards justice-oriented antiracist assessment.* Assessment for Learning Conference.

Randall, J., Poe, M., Oliveri, M. E., & Slomp, D. (2024). Justice-oriented, antiracist validation: Continuing to disrupt white supremacy in assessment practices. *Educational Assessment, 29*(1), 1-20.

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

Regulation 2016/679. *Regulation (EU) No 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).* https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

Rich, E. (1979). User modeling via stereotypes. *Cognitive Science*, *3*(4), 329-354.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement*, *40*(2), 163-184.

Rojas, L., & Liou, D. D. (2017). Social justice teaching through the sympathetic touch of caring and high expectations for students of color. *Journal of Teacher Education, 68*(1), 28-40.

Rose, D. (2000). Universal design for learning. *Journal of Special Education Technology*, *15*(3), 45-49.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivation: Classic definitions and new directions. *Contemporary Educational Psychology, 25*(1), 54-67.

Ryan, R. M., & Deci, E. L. (2019). Brick by brick: The origins, development, and future of self-determination theory. *Advances in Motivation Science, 6,* 111-156.

Sabatini, J. P., O'Reilly, T., Halderman, L. K., & Bruce, K. (2014). Integrating scenario-based and component reading skill measures to understand the reading behavior of struggling readers. *Learning Disabilities Research & Practice*, *29*(1), 36-43.

Samuel, T. S., Buttet, S., & Warner, J. (2022). "I can math, too!": Reducing math anxiety in STEM-related courses using a combined mindfulness and growth mindset approach (MAGMA) in the classroom. *Community College Journal of Research and Practice.*

Shute, V. J., & Zapata-Rivera, D. (2012). Adaptive educational systems. In P. Durlach (Ed.) *Adaptive Technologies for Training and Education* (pp. 7-27). New York, NY: Cambridge University.

Sinharay, S., & Johnson, M. S. (2024). Computation and accuracy evaluation of comparable scores on culturally responsive assessments. *Journal of Educational Measurement, 61*(1), 5-46.

Sireci, S. (2020). Standardization and UNDERSTANDardization in educational assessment. *Educational Measurement: Issues and Practice, 39*(3), 100-105.

Sireci, S., & Randall, J. (2021). Evolving notions of fairness in testing in the United States. In B. E. Clauser & M. B. Bunch (Eds.) *The History of Educational Measurement: Key Advancements in Theory, Policy, and Practice* (pp. 111-135). New York, NY: Routledge.

Solano-Flores, G. (2008). Who is given tests in what language by whom, when, and where? The need for probabilistic views of language in the testing of English language learners. *Educational Researcher, 37*(4), 189–199.

Sparks, J. R., Peters, S., Steinberg, J., James, K., Lehman, B. A., & Zapata-Rivera, D. (2019, April). *Individual Difference Measures that Predict Performance on Conversation-Based Assessments of Science Inquiry Skills.* Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada.

Sparks, J. R., Steinberg, J., Lehman, B., & Zapata-Rivera, D. (2022). *Leveraging students' background characteristics to predict performance on conversation-based assessments of mathematics.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Steinberg, J., Cline, F., & Sawaki, Y. (2011). *Examining the factor structure of a state standards-based science assessment for students with learning disabilities* (ETS Research Report No. RR-11-38). Princeton, NJ: Educational Testing Service.

Swadener, B. B. (2000). "At risk" or "at promise"? From deficit constructions of the "other childhood" to possibilities for authentic alliances with children and families. In L. Diaz-Soto (Ed.), *The Politics of Early Childhood Education.* NY: Peter Lang.

Taie, S., & Goldring, R. (2020). *Characteristics of public and private elementary and secondary school teachers in the United States: Results From the 2017–18 National Teacher and Principal Survey First Look* (NCES 2020-142rev). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved 05/04/2022.

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing.* New York: Springer.

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, *16*(3), 227-265.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197-221.

Verschelden, C. (2017). *Bandwidth recovery: Helping students reclaim cognitive resources lost to poverty, racism, and social marginalization.* Sterling, VA: Stylus Publishing, LLC.

Vygotsky, L. S. (1978). Interaction between learning and development. In M. Cole, V. John-Steiner, S. Scribner, & E. Souberman (Eds.), *Mind in Society: The Development of Higher Psychological Processes* (pp. 79-91). Cambridge, MA: Harvard University Press.

Wainer, H., & Thissen, D. (1994). On examinee choice in educational testing. *Review of Educational Research*, *64*, 159-195.

Walker, M. E., Olivera-Aguilar, M., Lehman, B., Laitusis, C., Guzman-Orth, D., & Gholson, M. (2023). *Culturally responsive assessment: Provisional principles* (ETS Research Report No. RR-23-11). Princeton, NJ: Educational Testing Service.

Walkington, C., & Bernacki, M. L. (2018). Personalization of instruction: Design dimensions and implications for cognition. *The Journal of Experimental Education*, *86*, 50-68.

Weitekamp, D., & Koedinger, K. (2023). Computational models of learning: Deepening care and carefulness in AI in education. In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, & O. C. Santos (Eds.), *Artificial Intelligence in Education: Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky (AIED 2023)*. Communications in Computer and Information Science, vol 1831. Springer, Cham.

Wellberg, S., & Evans, C. (2022). Assumptions underlying performance assessment reforms intended to improve instructional practices: A research-based framework. *Practical Assessment, Research, and Evaluation*, *27*, 23.

WIDA. (2020). *WIDA English language development standards framework, 2020 edition: Kindergarten–grade 12*. Board of Regents of the University of Wisconsin System. https://wida.wisc.edu/sites/default/files/resource/WIDA-ELD-Standards-Framework2020.pdf

Wilensky, R., Chin, D. N., Luria, M., Martin, J., Mayfield, J., & Wu, D. (1988). The Berkeley UNIX consultant project. *Computational Linguistics*, *14*(4), 35-84.

Wise, S. L., & Smith, L. F. (2011). A model of examinee test-taking effort. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High-stakes testing in education: science and practice in K-12 settings* (1st ed., pp. 139–153). Washington, DC: American Psychological Association.

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

Wise, S. L., Bhola, D., & Yang, S. (2006). Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT. *Educational Measurement: Issues and Practice*, *25*(2), 21-30.

Wise, S. L., Kuhfeld, M. R., & Soland, J. (2019). The effects of effort monitoring with proctor notification on test-taking engagement, test performance, and validity. *Applied Measurement in Education*, *32*(2), 183-192.

Zapata-Rivera, D. (2017). Toward caring assessment systems. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization* (UMAP '17). ACM, New York, NY, USA, 97-100.

Zapata-Rivera, D. (2021) Open student modeling research and its connections to educational assessment. *International Journal of Artificial Intelligence in Education*, *31*, 380-296.

Zapata-Rivera, D., & Hu, X. (2023). Assessment in intelligent tutoring systems SWOT analysis. In A. M. Sinatra, A. C. Graesser, X. Hu, G. Goodwin, & V. Rus (Eds)., *Design Recommendations for Intelligent Tutoring Systems: Strengths, Weaknesses, Opportunities and Threats (SWOT) Analysis of Intelligent Tutoring Systems* (Vol. 10) (pp. 83-90). Orlando, FL: US Army Combat Capabilities Development Command – Soldier Center.

Zapata-Rivera, D., Kannan, P., Forsyth, C., Peters, S., Bryant, A. D., Guo, E., & Long, R. (2018). Designing and evaluating reporting systems in the context of new assessments. In D. Schmorrow, & C. Fidopiastis (Eds.) *Proceedings of Augmented Cognition: Users and Contexts (AC2018)* (pp. 143-153). Cham: Springer.

Zapata-Rivera, D., Lehman, B., & Sparks, J. R. (2020). Learner modeling in the context of caring assessments. In R. A. Sottilare & J. Schwarz (Eds.), *Proceedings of the second international conference on Adaptive Instructional Systems, held as part of HCI International Conference 2020, LNCS 12214*, (pp. 422-431). Cham, Switzerland: Springer Nature.

Zapata-Rivera, D., Lehman, B., Sparks, J. R., Por, H., & James, K. (2018). *Identifying and dealing with unexpected responses in conversation-based assessments*

(Research Memorandum No. RM-18-13). Princeton, NJ: Educational Testing Service.

Zapata-Rivera, D., Sparks, J. R., Forsyth, C. M., & Lehman, B. (2023). Conversation-based assessment: current findings and future work. In R. J. Tierney, F. Rizvi, & K. Ercikan (Eds.), *International Encyclopedia of Education (Fourth Edition)*. Elsevier.504–518.

Citation:
Lehman, B., Sparks, J. R., Zapata-Rivera, D., Steinberg, J., & Forsyth, C. (2024). A Culturally Enhanced Framework of Caring Assessments for Diverse Learners. *Practical Assessment, Research, & Evaluation*, 29(9). Available online: https://doi.org/10.7275/pare.2102

Corresponding Author:
Blair Lehman
ETS
Email: blehman@ets.org