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

Application of Justice-Oriented Content Validation to a Classroom Assessment Literacy Measure

Valeria Zunino-Edelsberg, *University of California, Davis*
Megan E. Welsh, *University of California, Davis*
María Veronica Santelices, *Pontificia Universidad Católica de Chile*
Tony Albano, *University of California, Davis*

Abstract: Content validation is critical to instrument development (McCoach, 2013; Sireci & Benitez, 2023). However, recent scholarship suggests that instrument validation efforts may unintentionally contribute to educational inequities and proposes frameworks for justice-oriented test validation (Chang & Cochran-Smith, 2022; Randall et al., 2022). This study is among the first to report on efforts to apply justice-oriented validation to an affective measure. Our scenario-based instrument measures Chilean teacher candidate (TC) assessment literacy. Two content alignment studies assess the instrument's alignment with Chile's Teacher Preparation Standards and its ability to capture a variety of classroom assessment approaches—assessment as, of, and for learning and supremacist, culturally responsive, and culturally sustaining assessment. Results highlight that justice-oriented content validation methods improved the measure. It helped refine conceptions of equitable classroom assessment and modify the instrument to address instances where negative stereotypes about minoritized groups were inadvertently reinforced. As such, the study offers a model for applying justice-oriented content validation frameworks to affective instruments.

Keywords: Assessment literacy, Teacher education, Justice-oriented validation, Content validation, Classroom assessment

Introduction

Test validity refers to the degree to which evidence supports a specific measure use or interpretation, with multiple sources of validity evidence contributing to comprehensive evaluations (American Educational Research Association et al., 2014). The review of evidence based on test content is particularly important during the early stage of instrument development (McCoach, 2013; Sireci & Benitez, 2023). Recent scholarship criticizes such content alignment approaches because they too often rely on numerical data (Polit & Beck, 2006) and neglect justice-oriented principles (Chang & Cochran Smith, 2022; Randall et al., 2022), risking the production of instruments that perpetuate educational inequity. Justice-oriented content

validation addresses these limitations by explicitly examining whether item content, construct definitions, and content validation processes advance—or undermine—equity goals.

Despite these contributions, few studies have examined evidence with respect to instrument content using justice-oriented approaches. The current study integrates Chang and Cochran-Smith's (2022) and Randall et al.'s (2022) recommendations to inform the development of a scenario-based instrument measuring Chilean teacher candidate classroom assessment literacy.

The measure, which we call, the *Literacidad en Evaluación para Educadores Chilenos* (LEEC), is based on Chile's *Standards for the Teaching Profession for Elementary Education Careers* (Standards), which identify the competencies that teacher preparation programs are expected to develop in elementary teacher candidates. The LEEC also measures the extent to which practices are consistent with classroom assessment approaches—assessment *for* learning, assessment *as* learning, assessment *of* learning (Black & Wiliam, 1998; Earl, 2013, Heritage & Wiley, 2018); and culturally responsive, culturally sustaining, and dominant-group oriented classroom assessment practices (Bennett, 2023; Nortvedt, 2020; Randall et.al, 2022, 2024a, 2024b; Solano-Flores & Nelson-Barber, 1999; Wolf et al., 2025). It is a scenario-based instrument, as it presents respondents with a set of vignettes, intended to address different Standards, followed by items designed to represent different assessment theories or approaches. Vignettes serve as anchors describing concrete realistic situations and items prompt participants to apply their professional judgment to those specific contexts. Contextualized, scenario-based instruments are particularly useful for capturing complex, practice-oriented constructs (Banitalebi et al. 2025; Bonner & Chen, 2021). Consequently, content alignment must jointly consider a vignette and its associated items. However, content alignment studies of education-related scenario-based instruments have typically addressed either the vignettes or items in isolation (St. Marie et al., 2020; Spoto et al. 2025).

We applied justice-oriented content validation practices in examining the instrument to answer two research questions: *(1) How well does the instrument reflect both the Standards and the classroom assessment approaches that we intended to measure, and (2) To what extent, and in what ways, did application of justice-oriented content validation improve the instrument?*

We highlight the literature informing how we conceptualize justice-oriented content validation as applied to scenario-based measures next. Then we present the methods used in our content validation study and the results of these analyses. The paper concludes with reflections on the practical challenges and affordances of justice-oriented content validation, offering guidance for researchers who wish to apply these approaches to affective measures.

Review of Literature

This literature review explores previous efforts to evaluate evidence with respect to test content and contextualizes this study within other work focused on scenario-based measures and justice-oriented approaches to instrument development. We first provide an overview of the role of content validation in instrument development. Then, we examine specific methods used for evaluating the content alignment of scenario-based instruments. Finally, the paper explores justice-oriented frameworks for collecting content validity evidence.

Common Approaches to Content Validation

Studies that gather evidence with respect to test content evaluate the extent to which an instrument represents the intended construct and is consistent with its purpose (American Educational Research Association et al., 2014). They are essential during the initial stages of instrument development as they ensure that items accurately reflect the construct (Bandalos, 2018). Several authors have emphasized the importance of content validation as an initial step in instrument development and have argued that these processes

should be thoroughly documented, including details about construct definition, item development procedures, expert qualifications, and the methods used to analyze feedback (Gerst et al., 2025; McCoach et al., 2013; Sireci, 1998; Sireci & Benitez, 2023).

Most content validity studies involve systematic review by subject matter experts (SME)—individuals with expertise in the areas the measure captures—to determine whether each item is representative of the construct being measured (McCoach et al., 2013). Collecting both quantitative and qualitative feedback is essential as qualitative feedback allows for critical reflection about how the wording of items affects interpretation, preventing a focus on surface-level correctness (Almanasreh et al., 2022; Banitabeli et al., 2025; McCoach et al., 2013; Sireci & Benitez, 2023; Spoto et al., 2025; St. Marie et al., 2020).

McCoach et al. (2013) discuss content validation methods for affective measures and recommend the systematic collection of information about the certainty with which experts match each item to its intended construct and about the relevance of the item to the construct. They assert that this quantitative information should be compared with qualitative data, collected via open-ended responses at the end of the review form, to make decisions as to which items should be retained, need to be modified (and how to modify them), or should be eliminated. In a similar vein, Spoto et al. (2025) argue that, "experts are typically asked to judge the relevance of the presented items to the construct, but not to evaluate whether the construct is fully covered by the items" (p. 205). As such, making decisions about eliminating an item solely based on quantitative results could worsen challenges with construct underrepresentation. In addition, experts might agree on the relevance of an item to the larger construct but disagree on the specific dimension that the item targets. For this reason, Spoto and colleagues suggest that qualitative feedback should play a significant role in the validation process.

A significant gap exists between these recommendations and the ways in which content alignment studies have been implemented. Early critiques by Polit and Beck (2006) report concerns with underreporting and inconsistent implementation of content validation procedures across studies. More recently, Spoto et al. (2025) highlight the underutilization of qualitative feedback, limited rater diversity, and failure to use content alignment evidence to engage in iterative refinement of instruments. Gerst et al. (2005) report a more significant problem; many scholars fail to examine content validity evidence altogether despite its foundational importance. While these methodological concerns are significant, they represent only part of a larger problem with traditional validation approaches.

Content Validation of Scenario-Based Instruments. Scenario-based instruments—comprised of vignettes that present real-world challenges followed by a set of items describing how one might respond—are increasingly ubiquitous and are particularly useful for assessing affective or practice-oriented constructs (Banitalebi et al., 2025; Bonner & Chen, 2021; Chang et al., 2019; DeLuca et al., 2019). We use the term scenario to refer to the combination of vignettes and items that participants review in formulating a response. Little guidance exists on content validation of these complex measures, especially with respect to evaluating the vignettes, the items, and how they jointly function. These distinct components add extra complexity when vignettes and items represent different purposes, requiring adaptation of procedures to account for the interdependence of different components.

Most validation studies of scenario-based instruments have addressed either the vignettes or items in isolation (Bonner & Chen, 2021; DeLuca et al., 2016; St. Marie et al., 2020). We first describe content validation of one scenario-based instrument in education that examines content validation evidence for items and then describe an approach used in nursing education focusing on vignettes.

DeLuca and colleagues (2016) conducted a content validation study for the Approaches to Classroom Assessment Inventory (ACAI). The first part of the instrument consisted of five vignettes, each followed by four items with three response options per item. The vignettes, items and response options were each

designed to measure distinct aspects of classroom assessment, addressing 21 distinct characteristics in total—five topics addressed by vignettes, four themes addressed by items, and 12 assessment priorities captured by response options. Content validation focused on the four themes and 12 assessment priorities.

The authors recruited 10 North American assessment experts, including 10 classroom teachers and 10 specialists in educational assessment. SME used a five-point scale to rate the degree of alignment between each item/response option combination and: (a) the four assessment literacy themes that items were intended to capture, and (b) the twelve assessment priorities captured by response options. They asked SME who provided a low alignment rating to explain their rationale, to suggest revisions, and to provide general feedback on the vignettes. DeLuca et al. (2016) repeated this process three times, following iterative revisions to the instrument, until reaching acceptable content validation indices.

St. Marie et al. (2020) proposed a three-stage process for developing vignettes for scenario-based measures and apply it to an instrument designed to guide nurses in providing pain management support. Phase one involves drafting vignettes based on clinical practice and creating: (a) a content validation data collection form and (b) a qualitative interview guide that asks about missing or unnecessary information in, and the authenticity of, the vignettes. In the second phase, SME reviewed vignettes and rated their clarity, relevance, and importance to the measured dimensions. SME also participated in semi-structured interviews to provide more nuanced feedback about the instrument. Researchers used both the quantitative ratings and interview results to revise the instrument. In the third phase, revised vignettes were re-evaluated with a second round of quantitative ratings from the same group of SMEs, allowing for comparison of ratings between phases. St. Marie and colleagues (2020) observed improvements in two ratings of their pain management instrument (average relevance increasing from 0.93 to 0.96, importance from 0.68 to 0.93) and argue that this demonstrates the importance of using an iterative process.

These examples reinforce the value of using content validation approaches that integrate quantitative and qualitative information, as well as the importance of iterative scenario-based measure development. While innovative, these papers do not address how to fully capture the distinct, yet interconnected purposes that vignettes and items play in scenario-based measures. As such, content validation procedures must be adapted to account for both the distinct purposes of each component—in our case, alignment with the Standards and with assessment approaches—while also attending to their interdependence.

Justice-Oriented Validity Approaches

Content validation efforts have often centered on technical adequacy and psychometric rigor, overlooking the cultural, racial, and political dynamics that shape both the development and use of assessments. As such, they risk inadvertently reinforcing systematic barriers to equitable educational opportunities leading to negative social consequences (Chang & Cochran-Smith, 2022; Randall et al. 2022). Scholars have documented this equity gap in most instrument validation efforts (Chang & Cochran-Smith, 2022; Randall et al. 2022). A growing body of scholarship calls for justice-oriented frameworks moving test validation from a potential tool of inequity to an instrument of educational justice (Randall et al. 2022). Two frameworks guide our justice-oriented approach to content validation: (a) Randall et al.'s (2022, 2024a, 2024b) justice-oriented antiracist validity framework (JAV) and (b) Chang and Cochran-Smith's (2022) multicultural validity framework (MVF; adapted from Kirkhart 2013 and 2015). The JAV asserts that test validation practices reproduce racism through the uncritical promotion of hegemonic practices and proposes a set of critical questions to consider in evaluating whether items reinforce or disrupt stereotypes, incorporate anti-racist content, and challenge linguistic or cultural biases. As such, it calls for a fundamental rethinking of how instruments are validated—asking whose knowledge is legitimized, who is included in the validation process, and what broader social consequences arise from assessment use—and offering specific questions to guide the collection of different forms of validity evidence (Randall et al., 2022).

Chang and Cochran-Smith (2022) reviewed 45 assessment tools designed to capture culturally responsive teaching and found that the content validation and instrument development processes used for these instruments neglected cultural responsiveness during validation efforts, thereby perpetuating inequities and limiting their potential to advance equity in teacher preparation. They discuss how future validation efforts could be improved across five dimensions—theoretical, methodological, relational, experiential and consequential—asserting that all five dimensions are necessary to critically interrogate assessment tools and prevent situations in which future measures of teaching for equity and social justice inadvertently reinforce White hegemony.

We focus on the theoretical, methodological, and relational dimensions here because they directly relate to content validation. Chang and Cochran-Smith (2022) define these dimensions as follows. The theoretical dimension is concerned with the way in which the assessed constructs relate to understandings of equity and justice, including a recognition that culture is never neutral and that theories always carry cultural assumptions. The methodological dimension addresses the extent to which the methods used to develop assessments are culturally appropriate and consider participants' perspectives and interests, including ensuring that diverse stakeholders have a voice throughout the instrument development and validation process. And the relational dimension focuses on "the quality of the relationship between assessment developers and the participants" (Chang & Cochran-Smith, 2022, p.11), including respecting cultural norms and practices, engaging participants during the work, and reflecting on their own cultural biases.

Taken together, these frameworks challenge conventional approaches to content validation, offering critical questions that emphasize that validation is not just a technical process, but one that is also deeply cultural and ethical. Table 1 synthesizes how the MVF and the JAV work together to inform content alignment.

Method

The instrument examined here measures teacher candidate (TC) dispositions toward different classroom assessment moves by asking them how they might respond to 18 vignettes depicting various assessment challenges related to the Standards. Six items, each targeting a distinct assessment approach, follow each vignette. Items ask TCs to rate the likelihood with which they would implement specific assessment moves (1 = very improbable to 6 = very probable). We used two studies to gather content validity evidence: one that examined the alignment of vignettes with the Standards (Study 1, presented first), and another that examined the alignment of the items with classroom assessment approaches: (a) equity-orientation—supremacist, culturally responsive, or culturally sustaining approaches, and (b) assessment purposes—focusing on assessment of, for, or as learning (Study 2).

Study 1: Vignette Content Alignment

Participants. We contacted eight SME and all agreed to participate. SME were either current Chilean teacher educators working in educational assessment (6 people), or professionals directly involved in the development of the Standards (2 people). See Table 2 for SME information.

By design, our panel included Chilean experts from a range of higher education institutions serving diverse student populations. Six participants were based in Santiago and two lived in smaller cities in central southern and southern Chile. It is widely believed that socioeconomic status is the main driver of inequity in Chile (Valenzuela et al. 2014). For the purposes of selecting SME for Study 1, and considering the country's educational landscape, we defined diversity in terms of the socioeconomic status of the students served and of the community surrounding the university. While experts had a shared nationality, several brought cross-cultural perspectives from international study or work experiences.

Table 1. Implications of the MVF and JAV Framework for Content Validation

	Multicultural Validity Framework (Chang & Cochran-Smith 2022)	**Justice-Oriented, Antiracist Validity Framework (Randall, et al. 2022, p. 175)**
Operationalizing constructs	Theoretical dimension focuses on two aspects of defining constructs: *Selection of theories:* Are the theories cited in the assessment framework focused on criticality and issues of social justice? *Interpretation of theories:* To what extent does the assessment framework operationalization of theory center criticality and justice?	Concerned with the extent to which constructs are operationalized in ways that reinforce racism or are antiracist. Asks: ● How well understood is the construct being measured for all, including minoritized learners? ● Whose values, perspectives, ways of knowing, and experiences does the construct reflect, normalize, or marginalize? ● How stable is the construct across social, cultural, and racial contexts? ● Is the construct explicitly antiracist? Does it articulate the specific false and oppressive narratives it seeks to disrupt?
Content validation methods	Methodological dimension, concerned with all validation methods. *Methods used*: To what extent were conceptual frameworks fully elaborated? Did methods result in a deeper understanding of constructs? *Inclusivity and cultural responsiveness of methods:* Were stakeholders (e.g., educators and community members, especially from diverse backgrounds) involved in conceptualizing constructs, creating assessment tasks, and analyzing results?	Explores the extent to which content validation efforts purport to be colorblind and therefore reinforce systematic racism. Asks: ● Do the test items reflect/reify negative stereotypes of minoritized populations? ● Are there test items that actively disrupt negative stereotypes about minoritized populations? ● Has antiracist content been integrated into items explicitly? ● Does the content/language of the items privilege a particular linguistic or cultural way of thinking/making sense of the world?
Overarching approach to justice	Relational dimension, deals with the quality of relationships between researchers and participants. *Engagement of assessment stakeholders:* Were assessment stakeholders engaged throughout the assessment development process and given ample opportunities to provide feedback? *Cultural assumptions of assessment developers:* To what extent were cultural assumptions embedded in assessment frameworks and assessment items explored or questioned?	Asserts that we must go beyond ensuring representation of minoritized groups. Instrument developers must engage in self-reflection and analysis of personal identities and relationships to address the impact of identity on assessment development practices. Asks: ● Are marginalized stakeholders involved at every stage of the construct definition and refinement stage?

Table 2. Characteristics of Vignette Subject Matter Experts

ID	Nationality	Area of Expertise	Sex	Language	Degree
SSME 1	Chile	Classroom assessment; Educational policy	Female	Spanish	Ph.D.
SSME 2	Chile	Classroom assessment	Female	Spanish	Masters
SSME 3	Chile	Classroom assessment	Female	Spanish	Ph.D.
SSME 4	Chile	Classroom assessment	Female	Spanish	Ph.D.
SSME 5	Chile	Classroom assessment	Female	Spanish	Ph.D.
SSME 6	Chile	Classroom assessment	Female	Spanish	Ph.D.
SSME 7	Chile	Classroom assessment; Educational policy	Female	Spanish	Masters
SSME 8	Chile	Classroom assessment; Educational policy	Female	Spanish	Masters

Materials. SMEs received a copy of two sections of the Standards: *Assessment Planning* (Standard 4) and *Assessment and Feedback for Learning* (Standard 9), written in Spanish but translated here. Each standard contained more detailed focus areas (e.g., "Construction and collection of learning evidence") and descriptors (e.g., "Build, select and adapt evaluation criteria consistent with the learning objectives to guide your observation") as presented in Appendix A. Standard 4 contained two focus areas and eight descriptors—four descriptors associated with the focus area *Construction and collection of learning evidence* (AP1) and four descriptors associated with the focus area *Analysis of learning evidence and feedback* (AP2). Standard 9 contained three focus areas and seven descriptors. Three descriptors were associated with focus area *Criteria for assessment and monitoring of learning* (FFL1), three descriptors were associated with focus area F*eedback* (FFL2), and one was associated with the focus area *Self-assessment of learning* (FFL3).

We collected data via a Spanish language Qualtrics form that began with a brief written explanation of the study's purpose and detailed instructions for completing the task. This included information about the purpose of the survey, including the assessment approaches (equity-oriented assessment and assessment purposes), Standards, focus areas and the descriptors we intended to measure. SME reviewed 18 vignettes, presented above their six corresponding items. They then selected the descriptor(s) from the Standards document that best matched the vignette and its items taken as a whole. They could select up to three descriptors but were asked to indicate which descriptors matched best (Priority 1), second best (Priority 2), and which matched worst (Priority 3). SME could also indicate that no descriptors matched the vignette, or that only one or two descriptors matched. At the end of each vignette, an open-ended comment box collected additional feedback.

Procedure. We asked SME to participate in data collection over Zoom, selecting between two formats: (1) two 90-minute meetings, or (2) one three-hour session. Data collection spanned five sessions: three three-hour sessions with two SME each, and one set of 90-minute meetings with two SME. SME received a $100 gift card as an honorarium.

One week prior to the data collection sessions, experts received an email asking them to print the excerpt of the Standards and to bring it with them. Each session began with the lead researcher presenting a slide deck that outlined the purpose of the activity and provided an overview of the structure of the Standards. The group reviewed slides, including completing two practice scenarios (vignettes and items) that were unrelated to assessment but associated with other Standards. Following these practice exercises, we asked SME to complete the online data collection form while consulting the printed copy of the Standards. They

turned off their cameras and worked independently. At the end of the session, SME were asked to provide additional verbal feedback on the instrument.

Data Analysis. Analyses holistically considered both SME ratings and qualitative feedback. Following D'Agostino et al. (2008), who also evaluated the match of items to state content standards by asking SME to identify up to three standards that a given item captures, we assigned one point to SME matches to the intended descriptor as Priority 1, 0.5 points when they assigned the intended descriptor to Priority 2 or 3, and 0 points when the intended descriptor was not matched.

Alignment indices were calculated as the mean match rate with the intended descriptor. Vignettes adequately matched descriptors when the mean match rate reached 0.7 or greater. Because descriptors are more specific than focus areas, we also calculated mean match to focus areas. Finally, we holistically reviewed SME feedback along with these statistics, both to try to make sense of low ratings and to address any other issues that experts identified, especially those relating to issues of language, culture, and justice.

Analyses identified ways to improve vignettes and/or to inform decisions to eliminate them. Our goal was to create a final instrument with six to nine scenarios. However, no final decisions were made at this stage. Results from both the vignette and item validation studies informed final judgments.

Item-level Content Validation

Participants. A total of 16 experts with diverse cultural, racial, and linguistic backgrounds participated in the study and brought a range of perspectives to the analysis. Five SME had expertise in culturally responsive assessment and eleven were experts in classroom assessment. While we designed the survey for Spanish speakers, some SME did not speak Spanish. We therefore collected data in Spanish (five SME) and in English (11 SME) to get feedback from a variety of experts in justice-oriented assessment (see Table 3).

Table 3. Characteristics of Item Subject Matter

ID	Nationality	Area of Expertise	Sex	Degree	Form
ISME 1	USA	Justice-oriented assessment	Female	Ph.D.	A: English
ISME 2	USA	Justice-oriented assessment	Female	Ph.D.	A: English
ISME 3	USA	Classroom assessment	Female	Ph.D.	A: English
ISME 4	USA	Classroom assessment	Female	Ph.D.	A: English
ISME 5	USA	Classroom assessment	Female	Ph.D.	A: English
ISME 6	USA	Justice-oriented assessment	Female	Ph.D.	A: English
ISME 7	USA	Justice-oriented assessment	Female	Ph.D.	B: English
ISME 8	USA	Classroom assessment	Female	Ph.D.	B: English
ISME 9	Canada	Classroom assessment	Male	Ph.D.(c)	B: English
ISME 10	Italy	Classroom assessment	Female	Ph.D.	B: English
ISME 11	Chile	Classroom assessment	Female	Masters	B: English
ISME 12	Chile	Classroom assessment	Female	Ph.D.	A: Spanish
ISME 13	Chile	Justice-oriented assessment	Female	Ph.D.	A: Spanish
ISME 14	Chile	Classroom assessment	Female	Ph.D.	B: Spanish
ISME 15	Chile	Justice-oriented assessment	Female	Ph.D.	B: Spanish
ISME 16	USA	Justice-oriented assessment	Male	Ph.D.	B: Spanish

Materials. Item-level content validation materials consisted of four Qualtrics forms (Forms A and B, each in both English and Spanish). SME rated the match of individual items to the six assessment approaches—assessment as learning, assessment for learning, assessment of learning, culturally responsive approach, culturally sustaining approach, and supremacist approach. We defined assessment of, for and as

learning following Earl (2013). Assessment of learning refers to any use of classroom assessment intended to certify performance through grading, including promoting understanding of grading processes and making decisions about the fairness of grades. In contrast, assessment for learning helps teachers make instructional decisions, including analysis of student work and gathering feedback to inform teaching. Finally, assessment as learning helps students reflect on their learning to improve their work by encouraging use of self-regulation strategies, including peer- and self-assessment.

Supremacist uses of assessment include prioritizing the ways of knowing and expressing most often used by dominant groups, including those that set low expectations for students belonging to minoritized groups (Randall et al., 2022). Multiple authors (Bennett 2023; Nortvedt et al., 2020; Randall et al., 2022; Solano-Flores & Nelson-Barber, 1999; Wolf et al., 2025) informed our definition of culturally responsive classroom assessment to include assessment efforts that value student identity and diverse ways of knowing and expressing ideas. We drew on the same authors to define culturally sustaining assessment as assessment activities that take active steps to preserve, promote, and deepen understanding of: (a) the linguistic and cultural practices of historically marginalized communities and (b) issues of justice.

Study 2 evaluated 108 items (6 items x 18 vignettes). To reduce SME burden, we divided the scenarios into two different data collection forms and asked SME to review 54 items, six each presented under nine vignettes. We provided data collection forms written in SME's preferred language, taking care to balance the number of SME with varying expertise and linguistic preferences across forms, as shown in Table 3.

Consistent with McCoach et al. (2013), forms began with instructions outlining the task, followed by brief definitions of the six assessment categories. We displayed these definitions alongside each scenario. For each item, three adjacent columns captured different types of expert judgments: (a) selecting the assessment approach that best represented the item, or indicating "none of the above" if no category was appropriate; (b) rating the level of confidence in the match to the assessment approach, from 1 (*completely unsure*) to 4 (*very sure)*; and (c) rating the perceived relevance of the item to the assessment approach, from 1 (*completely irrelevant*) to 3 (*highly relevant)*. An open-ended prompt, "Please feel free to add any comments to the items and/or vignettes," followed each scenario.

Procedure. We sent 24 email invitations to potential SME. Emails explained the purpose of the study, the procedures involved in reviewing the instrument, and offered an honorarium in the form of a $100 gift card. Sixteen of the 24 SME accepted the invitation: seven with expertise in justice-oriented assessment and nine with expertise in classroom assessment. We asked SME to complete the review within one month by completing a Qualtrics form.

Data Analysis. We jointly analyzed responses from both the Spanish and English versions. Both quantitative and qualitative analysis informed item revisions. Quantitative analysis included calculating validation indices across 16 raters and the nine items each rater examined per assessment approach (16*9 = 144 ratings). We first calculated the proportion of SME who matched each item to its intended assessment approach (match rate), the mean relevance for items that were correctly matched, the proportion of SME that were "pretty sure" and "very sure" of their classification for items that were correctly matched, and the item level content validity index (CVI; or proportion of SME that correctly matched and rated the items as "highly relevant" to the designated assessment approach) to examine how specific items functioned. And then considered how well different assessment approaches were being measured by aggregating across items sharing the same approach. We also calculated these statistics counting culturally responsive and culturally sustaining items as correct if they matched either category, recategorizing items as using an equity-oriented approach to assessment.

We organized all results, including qualitative feedback, in a spreadsheet containing item-level information. This allowed for a holistic review that gave equal weight to qualitative and quantitative

indicators and informed decisions to re-word or eliminate specific items. Items with match rates less than 0.70 received extra attention. Final decisions about how to modify or eliminate vignettes were based on the results of both vignette-level and item-level analyses.

Qualitative analyses critically interrogated themes related to justice-oriented assessment by intentionally applying the content validity recommendations included in Table 1. These MV and JAV frameworks call for explicit attention to racial and ethnic dynamics, including examination of whose values and perspectives are reflected or marginalized and examining the extent to which negative stereotypes are endorsed. We could (and should) have taken the extra step of explicitly asking SME to rate and provide feedback on issues of equity, justice, and systematic bias in the instrument and in our approach to content validation using the questions posed by each framework.

Our approach was more nuanced. The research team collaboratively discussed the literature on justice-oriented classroom assessment and its application in the Chilean context throughout each stage of the study. Chile is quite different from the United States where much of the literature on justice-oriented assessment and validation originated. This ongoing reflection helped us iterate on our conceptions of what it means to be justice-oriented.

We also recruited SME who represented as wide an array of lived experiences and kinds of expertise as possible. The justice-oriented goals of the study were highlighted during recruitment, and many SME likely participated because of this goal. Alignment study materials also reflected this focus. These factors may have encouraged SME to provide feedback on the instrument in ways that align with the MV and JAV frameworks. Both the results section and Table 4 summarize how the MV and JAV frameworks did and did not influence study methods.

Results

We present the results of our content validity study in four main sections. First, we report the quantitative findings at the vignette level. Next, we present item-level quantitative findings. This is followed by the qualitative results for both vignettes and items, organized into thematic categories. Finally, we summarize the decisions derived from a holistic analysis that integrated quantitative and qualitative evidence.

Vignette Alignment

We examined the extent to which the vignettes included in the scenario-based measure aligned with the specific descriptors and focus areas outlined in the Standards, taking vignettes and items holistically into account. We report results for both the match to descriptor and the match to focus area. Results varied considerably across vignettes, with descriptor match rate ranging from 0.00 to 0.88 and focus area match rate ranging from 0.25 to 1.00. Seven vignettes met the matching criteria (match rate >= 0.70) for descriptors and four more matched the broader focal area but not the descriptor. Therefore, enough vignettes worked to develop a final instrument of the intended length based on quantitative analyses.

The four vignettes that sufficiently matched only the focus area seemed to capture an assessment skill in the Standards, but not the intended descriptor. Vignettes tended to match to descriptors falling within the *Criteria for assessment and monitoring of learning* and *Analysis of learning evidence and feedback* focus areas more than vignettes measuring other descriptors, allowing the study team to select between vignettes capturing these competencies on the final instrument. We only included one vignette designed to address *Self-assessment*. It met match rate criteria at both the descriptor and focus area levels (see Table 5 and Table B1 in Appendix B for detailed results).

Table 4. Application of the MVF and JAV Framework in this Study

Multicultural Validity Framework (Chang & Cochran-Smith 2022)	**Justice-Oriented, Antiracist Validity Framework (Randall, et al. 2022)**
Operationalizing constructs	
Selection of theories: Assessment framework draws from seminal works in justice-oriented, sociocultural, and culturally responsive classroom assessment.	● SME advice that the difference between culturally responsive assessment and culturally sustaining assessment is not well understood by the field at large and is therefore difficult to operationalize led to a decision to collapse these into equity-oriented assessment approaches.
Interpretation of theories: SME provided extremely helpful feedback on revising supremacist, culturally responsive and culturally sustaining classroom assessment items to ensure that they were fully consistent with theory.	● Discussion of what it means to be justice-oriented in Chile, where socioeconomic status is seen as the main driver of inequity. Yet Chile also has a large, marginalized indigenous population and explosive immigration growth from across Latin America. ● SME provided feedback on areas in which the instrument failed to disrupt supremacist ideologies.
Content validation methods	
Methods used: SME qualitative feedback, paired with examination of quantitative alignment ratings, supported reflection about assessment approaches and equity-oriented assessment items, leading to a deeper understanding of the constructs and better items.	● SME were not asked to provide feedback on the extent to which test items: (a) reflected/reified or (b) actively disrupted negative stereotypes, but they did so anyway.
Inclusivity and cultural responsiveness of methods: Stakeholders provided feedback on items and vignettes. Stakeholders represent many nationalities, including Chileans from a wide array of communities. Data collection was conducted in both Spanish and English, depending on SME preference.	● SME automatically provided feedback about the content/language of items that privilege linguistic or cultural ways of thinking, some reflection by researchers about when it might be necessary to include items that describe privileging certain groups to identify when TC use this kind of thinking.
Overarching approach to justice	
Engagement of assessment stakeholders. The research team was comprised of two Chilean nationals and one US national, all from privileged backgrounds. SME only served as content alignment participants.	● Researchers repeatedly discussed their personal identities and relationship to power. Assumptions about sources of inequity in Chile and what it means to be supremacist or justice-oriented in the Chilean context were also discussed.
Cultural assumptions of assessment developers: Researchers repeatedly discussed the cultural assumptions embedded in construct definitions and in the assessment itself, but we did not explicitly ask SME to provide feedback on this.	● Stakeholders represent many nationalities, including Chileans from diverse communities. Data collection was conducted in both Spanish and English, depending on SME preference. The research team was comprised of two Chilean nationals and one US national, all from privileged backgrounds.

Table 5. Number of Vignettes with Average Match ≥ 0.7 to Descriptors and Focus Areas

Focus Area	# of Scenarios Matched to Descriptors (≥0.7)	# of Scenarios Matched to Focus Areas (≥0.7)
Construction and collection of learning evidence (n=4)	0	2
Analysis of learning evidence and feedback (n=5)	2	3
Criteria for assessment and monitoring of learning (n=5)	3	3
Feedback (n=3)	1	2
Self-assessment (n=1)	1	1
Total (n=18)	7	11

Item Alignment

We assessed the extent to which items aligned with the six different approaches to assessment—assessment for learning, assessment as learning, assessment of learning, culturally responsive assessment, culturally sustaining assessment, and supremacist approach to assessment. Item alignment was not as promising as vignette alignment. Only 23 percent of items attained a match rate of 0.70 or better (Appendix C, Tables C1-C18). The assessment for (33 percent), as (28 percent), and of (28 percent) learning categories had the highest percentage of items meeting these criteria. Mean match rates were fairly consistent across categories, but the percentage of items that worked in each category varied. The percent of items that met alignment criteria was lower for items addressing justice-oriented assessment: culturally responsive assessment (22 percent of items), supremacist approaches (17 percent), culturally sustaining approaches (11 percent; Tables 6-7).

Table 6. Content Validation Indices Across Raters and Items by Assessment Approach

Approach	Mean Match	Mean Relevance	Percent Top Certainty	Percent Top Relevance
Assessment of Learning (n=144)	0.45	2.10	39	29
Assessment for Learning (n=144)	0.57	2.93	56	53
Assessment as Learning (n=144)	0.55	2.67	49	47
Supremacist (n=144)	0.48	2.13	40	21
Culturally Responsive (n=144)	0.48	2.64	39	39
Culturally Sustaining (n=144)	0.41	2.73	40	36
Equity-Oriented (n=288)	0.71	2.17	64	58

Note: Content validation indices were calculated across raters (16 total) and items (9 per assessment approach reviewed by any given rater), such that 144 (16*9) ratings were used in calculations. Equity-oriented approaches included ratings for both culturally responsive and culturally sustaining approaches (144 + 144).

As expected, a smaller percentage of items had acceptable CVI levels, with similar CVI results across dimensions except for the supremacist category—where no items were both matched to the intended dimension and viewed as relevant—and the culturally sustaining category where 11 percent of items met these criteria. Raters were also uncertain about whether their ratings matched items to the intended dimensions, with assessment for learning (33 percent of items rated pretty/very sure) and assessment as learning (22 percent) performing best (Table 7).

Table 7. Percentage of Items with Acceptable Content Validation Indices

Approach	Percent of Items Match ≥ 0.7	Percent of Items Pretty/Very Sure of Match ≥ 0.7	Percent of Items CVI ≥ 0.7
Assessment of Learning (n=18)	28	17	22
Assessment for Learning (n=18)	33	33	28
Assessment as Learning (n=18)	28	22	28
Supremacist (n=18)	17	6	0
Culturally Responsive (n=18)	22	6	22
Culturally Sustaining (n=18)	11	11	11
Equity-Oriented (n=36)	61	44	33

Note. The number of equity-oriented item matches can double the number of other matches (36, as opposed to 18) because there were two items—culturally sustaining and culturally responsive—addressing this dimension.

Based both on qualitative feedback and these findings, we suspected that our culturally responsive and culturally sustaining items may be equity-oriented but might not distinguish between culturally responsive and culturally sustaining practices. For this reason, we also calculated content validity indices by grouping culturally responsive and culturally sustaining items into a broader, equity-oriented category, counting matches of these items to either dimension as a match to equity-oriented practices. This improved alignment indices—61 percent of items correctly matched to the equity-oriented assessment dimension. SME were pretty/very sure about the match for 44 percent of these items and rated 33 percent of matched items as relevant to their intended dimension. See Tables 6-7 and Appendix C for details.

Qualitative Results

The disappointing quantitative results only increased the importance of qualitative findings, which provide the main source of information about the extent to which the instrument achieved its justice-oriented aims. We describe the common themes that emerged between SME reviewing the vignettes and SME reviewing the items because the feedback was remarkably similar for both groups. Themes include: (a) inadvertent stereotyping of minoritized groups, (b) comments that supremacist items were "not supremacist enough," (c) appropriateness of naming specific groups in vignettes, (d) overlap of categories across items, and (e) lack of coherence between vignettes and items. We also examine variations in feedback according to SME expertise, including differences in the perspectives of Chilean and international experts. The MV and JAV frameworks guided all analyses.

Inadvertent Stereotyping of Minoritized Groups. Reviewers noted that several vignettes unintentionally perpetuated deficit-based narratives by portraying immigrant and Indigenous students in ways that reinforced negative stereotypes or positioned them as "others" in the classroom. One SME stayed

after their data collection session to share that there was a systematic problem with the instrument spanning multiple vignettes—they presented minoritized students as performing less well than their peers. Other SME identified similar issues with specific vignettes as discussed next.

One vignette described a situation where two students, Paulina from Chile and María, a new student from Venezuela, initially scored similarly on a standardized test. However, Paulina's score increased substantially the next year whereas María's was lower. As SMME 5 explained, "In the current context, it's common to find Venezuelan students in classrooms. The example of a foreign student who performs poorly on the test can (unintentionally) contribute to the stigmatization of these students."

A similar concern arose in a vignette depicting a situation where all but two students performed well on the most recent test. The last item asked teacher candidates if they would teach the class about the diverse ways in which students from different countries demonstrate their knowledge. SSME3 thought that readers might make assumptions that immigrant students might perform poorly even though the vignette did not specify the background of the lower-performing students.

Based on this feedback, we reviewed the entire instrument and identified seven vignettes that portrayed minoritized students as low performing. We removed five from the next iteration of the instrument and revised one vignette that originally said that Haitian students struggled with mathematical explanations to indicate that the teacher knew that they were mathematically advanced. The other retained vignette illustrates bias against minoritized students in a reading comprehension test focused on a topic that may be unfamiliar—celebration of Chile's Independence Day—and includes items that provide an opportunity to identify this issue.

Not being Supremacist Enough. SME also noted that we often failed to include supremacist items that truly reflected a supremacist stance. As ISME15 explained 'I really struggled to decide if I had enough information to call this "Supremacist."' For example, one option that valued students who "speak like a scientist" was designed to capture how certain ways of speaking are privileged. Another item asked TC if they would include performance on the reading comprehension test focused on the Chilean Independence Day in immigrant student report card grades. In response to this feedback, we reviewed all supremacist items and revised every one.

Appropriateness of Naming Groups in Scenarios. In line with the previous theme, some experts suggested that referring explicitly to communities viewed as facing challenges in Chile reinforces negative stereotypes. They recommended omitting any references to minoritized groups, a recommendation that we appreciated but ultimately rejected as it could potentially reinforce narratives that instruments can be colorblind.

The last item associated with a vignette involving traditional legends asked whether TC would ask Mapuche (indigenous) students to share a family legend and explain the phenomenon it seeks to illustrate. SSME4 noted, "It's not clear to me how much of a contribution is made by including the fact that they are Mapuche. The options are geared toward generic aspects, except for the last one which could apply to any family."

Another vignette pertained to a classroom where newly arrived Haitian students struggled to write their mathematical reasoning in Spanish. The items asked how likely teachers were to: (a) grade in a way that captured the quality of mathematical thinking without penalizing for grammar, (b) collect additional evidence to better support students in expressing their reasoning in Spanish, or (c) allow all students to complement their explanations using other languages and/or visual representations. SSME4 commented, "Since the problem is language, perhaps it could be generic. You have foreign students who speak other languages (French, English), so as not to stigmatize Haitians as having problems."

Overlap of Categories Within Items. SME also commented that some items could reasonably fall into more than one category. ISME1 commented that one item could fit either *assessment for learning* or *assessment as learning*. They stated, "I would also categorize the [item] as assessment for learning, understanding that if self-assessment and peer assessment are promoted, they are part of this broader concept."

ISME15 identified an item that could simultaneously reflect an *assessment for learning approach* and a *supremacist approach*. The item asked teacher candidates about the likelihood of providing feedback to help students learn to communicate like scientists. They stated, "I struggled with *communicate like scientists orally*. I selected *assessment for learning* at first, as the teacher is providing feedback, but then I started thinking about what it means to *communicate like a scientist orally* and decided that this type of communication might show a teacher's biases of who scientists are."

ISME15 also reflected on a vignette portraying a classroom where a Middle Eastern student expressed fractions in a way that is common in their home country (e.g., saying "thirds-two" instead of "two-thirds") and an item asking whether respondents would distinguish between the student's mathematical understanding and the way they expressed themselves. They noted, "Although I labeled the item as assessment of learning, I also feel that it could be labeled culturally responsive." ISME14 suggested, "[For all of the items on all pages], each item could be characterized by its purpose approach (of, for, as learning) and its cultural stance (supremacist, relevant, responsive). In each case [on all pages], I have selected the choice I think is most salient in each item as written, but that leaves most of the items not well described."

SME most frequently identified an overlap between culturally responsive and culturally sustaining assessment items. For a vignette addressing planning the year's writing assessments, one item (Option 3) asked respondents how likely they would be to organize oral activities where students shared experiences from their community and then wrote about one of them. Another (Option 5) addressed the likelihood of assigning a writing task that allowed students to write in any language or combination of languages they chose. ISME1 commented, "Options 3 and 5 made me uncertain about whether they represent an example of culturally responsive or culturally sustaining assessment. I find it difficult to see the difference between the two."

Lack of Coherence Between Vignettes and Items. SME also noted that items sometimes failed to address the vignette's stated purpose. For the vignette presenting a situation in which a teacher was trying to understand why two students—Paulina (Chilean) and María (recently arrived from Venezuela)—showed very different levels of progress on standardized tests, item 1 asked how likely teacher candidates would be to congratulate Paulina because she prepared better for the test than María. Item 2 asked whether they would group María with other low-performing students and give them easier tests in the future. ISME1 commented, "[Item 1] has nothing to do with the question posed in the vignette." ISME4 added, "Not all actions align with the purpose of the vignette."

They also commented on an item involving a math lesson in which students were learning to order fractions, and a significant group answered a question incorrectly. The vignette states that the teacher's goal is to understand the source of this confusion. The item asked how likely respondents would be to "continue moving forward with the content because not all students perform well in every activity." They commented, "Option 1 doesn't address the purpose of the vignette, which is to understand the source of their confusion."

Another vignette described a lesson on the responsible use of natural resources. Students had watched and discussed a documentary and then written a personal narrative about their experiences of caring for nature. For the next activity, they would work in groups to create and present a poster on environmental protection. The teacher wanted to provide an opportunity for students to reflect on their learning. Items asked about the likelihood of (a) providing individual and whole-class feedback during the activities; (b) asking students to evaluate the quality of the evidence supporting their own and their peers' opinions; or (c)

emphasizing that all modes of expression—oral, visual, or written—contribute valuable academic knowledge. Half of the SME noted that the options did not fully align with the intended focus on student reflection. As SSME1 remarked, "The items appeal to the entire context of the scenario, not just reflection, which is the question asked in the vignette."

Feedback Variations Between Chilean and non-Chileans experts. We examined the characteristics of the feedback provided by Chilean experts (n=13) and non-Chilean experts (n=11). As shown in Table 8, Chilean and non-Chilean SME focused on different content. Chilean SME tended to focus heavily on 'Other' topics, such as clarity and coherence (56 percent of comments) and on disciplinary factors (31 percent of comments) while non-Chilean SME mostly focused on disciplinary factors (84 percent). The relationship between nationality and focus on clarity and coherence is as expected: 12 of 13 Spanish language reviewers were Chilean.

Contextual comments, the least common type of comment but mentioned three times as often by Chilean SME, highlighted specific local realities, including cultural and linguistic, political, or demographic issues. For example, SSME1 noted misalignments with educational policy: "This is contradictory given the new evaluation decree." SSME2 commented on demographic shifts, "In the current context, it's common to find Venezuelan students in classrooms." SSME5 focused on linguistic characteristics "Perhaps try another misspelled word instead of *schaleco* because it's uncommon." In contrast, although fewer non-Chilean SME participated, they provided more disciplinary comments (84 percent of comments) compared to the Chilean group (31 percent of comments).

Finally, for both groups, most disciplinary comments were related to classroom assessment rather than justice-oriented assessment. This aligns with the overall composition of the panels. Across both groups (N=24), there were only seven experts in justice-oriented assessment.

Table 8. Summary of types of experts' comments by nationality

		Type of comment			Focus of disciplinary comment	
Experts' nationality	Number of comments	Percent Other	Percent Context	Percent Disciplinary	Percent Classroom Assessment	Percent Justice-oriented Assessment
Chilean (n=13)	48	56	13	31	73	53
Non-Chilean (n=11)	37	8	5	84	84	32

Note. The sum of various types of comments exceeds 100 percent because comments often address multiple topics. Type of comment percentages were calculated by dividing by the total number of comments. Focus of disciplinary comment percentages were calculated by dividing by the number of disciplinary comments (15 Chilean and 31 Non-Chilean). Types of comments are defined as follows:

Other: Comments related to clarity, wording or other aspects not connected to the context where the survey will be administered, nor to disciplinary aspects of classroom assessment.

Context: Comments referring to the context where the survey will be administered (e.g., cultural aspects, political aspects, demographic aspects)

Disciplinary: Comments related to the field of classroom assessment or justice-oriented assessment.

Changes Made Based on Holistic Analysis. Using the results presented above, and keeping the MV and JAV frameworks in mind, we made several changes to the instrument. The simplest revisions involved deleting or revising items and vignettes that were unclear or poorly aligned to their intended Standard or assessment approach, including merging the culturally responsive and culturally sustaining items into an equity-oriented assessment category. These kinds of revisions are consistent with standard content validation efforts (Bandalos, 2018; McCoach et al., 2013).

We also examined results through the lens of the MV and JAV frameworks, paying special attention to the extent to which the measure: (a) presents negative stereotypes of minoritized populations, (b) disrupts negative stereotypes, (c) integrates antiracist content, and (d) privileges linguistic or cultural ways of thinking, eliminating scenarios that reinforced deficit perspectives. For instance, one expert commented that several scenarios referred to immigrant students who had performed poorly. Applying Randall et al.'s (2022) JAV reflective prompt—"Do the test items reflect or reify negative stereotypes of minoritized populations?" (p. 175)—we recognized that the wording indeed risked stereotyping immigrant students. Consequently, we omitted one scenario that mentioned that a Venezuelan student had performed lower than a Chilean student. In the nine scenarios we retained, we also modified language that reinforced negative stereotypes.

Discussion

Findings from our content validation study led to meaningful refinements that improved the survey's conceptual clarity and helped us achieve the measure's justice-oriented goals. Although results supported use of vignettes that covered five of the six intended standard focus areas, item-level results suggested the need for substantial revision to adequately measure each dimension.

However, we wanted to do more than capture the intended construct. We also wanted to address issues of justice, not only in the content of the instrument but also in our approach to content validation. Drawing on Chang and Cochran-Smith's (2022) theoretical, methodological, and relational dimensions, we attended to the theories of classroom assessment used to develop the measure, the composition of the research team (two Chilean and two U.S. researchers), and the composition of the SME reviewers—in terms of language, nationality, scholarly expertise, and work setting—to ensure that the perspectives and interests of diverse stakeholders were considered throughout the instrument development and validation process. We approached this work with curiosity about our own cultural expectations and those of the other researchers and participants, questioning the veracity of assumptions that inequity in Chile is solely a socioeconomic challenge and that Chilean schools are similar to those in the United States.

Importantly, we listened to our diverse group of SMEs who were generous in providing rich qualitative feedback about the instrument, allowing us to reflect on a wide range of "standpoints, epistemologies, methods, and perspectives" (Chang & Cochran-Smith, 2022, p. 8). SME helped us think through how items might be interpreted in the Chilean context, how they might reinforce negative stereotypes, and helped us identify places where our dimensions were not as distinct as we had hoped. This led to questions about the extent to which we can uniquely identify culturally responsive or culturally sustaining assessment practices. It also helped us notice that many equitable assessment practices also captured assessment as, of, or for learning.

In both vignette and item content alignment procedures, SME reviewed items and vignettes together. We believe that our procedure holds value because there is currently no established content validation process specifically designed for vignettes (St. Marie et al., 2020) and because this approach helped identify instances in which vignettes and items worked at cross-purposes. The inclusion of qualitative information, as recommended by Spoto et al. (2025), was a significant strength of our methodology because it provided a rich source of information about issues of justice and equity that we would not have been able to obtain

through quantitative data alone. It also played a crucial role in our selection of scenarios and in the modifications we introduced. There were also instances in which we could confirm our interpretations and decisions. For example, alignment indices for culturally responsive and culturally sustaining items were low, and several experts noted an overlap between these categories. We therefore combined the culturally responsive and culturally sustaining assessment approach and proposed an equity-oriented assessment dimension and could corroborate that ratings improved substantially. Consequently, we refined our framework, reducing the three equity approaches to two (supremacist vs. equity-oriented).

The JAV framework developed by Randall et al. (2022, 2024a, 2024b) was particularly influential in shaping our approach. Interrogating the qualitative feedback while attending to negative stereotypes, antiracist content, and instances in which dominant group linguistic or cultural ways of thinking were privileged improved the measure considerably. A key tension that emerged from this work involved consideration of the extent to which it is necessary to name specific minoritized groups to integrate antiracist content and to examine the extent to which educators give privilege to linguistic or cultural ways of thinking. We opted to include details about student backgrounds as it is impossible to gauge the extent to which a TC is equity-oriented or supremacist in their assessment approach in the absence of information about student race, language, culture, or socioeconomic background.

Some SME identified ways in which the measure inadvertently reinforced negative stereotypes by identifying immigrant students as low performing. This created an opportunity for revision that instead disrupted negative stereotypes. They also pointed out other ways in which the instrument was not justice-oriented enough and encouraged us to make items more explicitly supremacist to better capture the full array of ways in which TC think about assessment. Finally, some SME recommended avoiding potential biases by omitting identity-based information. We respectfully disagree. Identifying student backgrounds provides important context to help gauge the extent to which educators embrace equity-oriented approaches.

The JAV framework helps guide analyses and identify and revise several items that unintentionally reflected deficit-based assumptions. This strengthened our ability to explore TC beliefs about the roles of justice, equity, language, and culture in enacting classroom assessment. However, we wish we had done more.

Although SME did provide equity- and justice-oriented feedback after learning about the focus of the instrument, we did not explicitly ask questions that directly tied to the MV or JAV frameworks. The use of structured prompts to guide SME reflection on justice-oriented issues could have promoted deeper critical reflection and more systematically identify content that either reproduced or disrupted inequities. This might have also strengthened our ability to reveal deficit-oriented perspectives or implicit cultural assumptions embedded in item wording or vignette design. This omission highlights the need to intentionally embed justice-oriented protocols throughout the validation process, not only in selecting SME, but also while eliciting their feedback.

The current study is particularly significant for several reasons. First, Chile's educational context—characterized by high socioeconomic segregation and diverse student population (Valenzuela et al., 2014)—makes it an ideal setting for demonstrating how justice-oriented validation can inform culturally responsive assessment practices. Second, the comprehensive model presented here can inform future instrument development efforts across diverse educational contexts. The model integrates principles of justice into instrument development by carefully selecting SME, working multilingually, and by fully integrating qualitative and quantitative results to review items with a "culturally conscious" lens (Badrinarayan et al., 2025, p. 115). By grounding the instrument in both local standards and international assessment theory, we illustrate how justice-oriented validation can bridge contextual specificity and broader theoretical perspectives.

Received: 8/16/2025. **Accepted:** 12/19/2025. **Published:** 1/12/2026.

Citation: Zunino-Edelsberg, V., Welsh, M.E., Santelices, V., & Albano, T. (2025). Application of justice-oriented content validation to a classroom assessment literacy measure. *Practical Assessment, Research, & Evaluation*, 30(2)(5). Available online: https://doi.org/10.7275/pare.3657

Corresponding Author: Valeria Zunino-Edelsberg, University of California, Davis.
Email: vczunino@ucdavis.edu

References

Almanasreh, E., Moles, R. J., & Chen, T. F. (2022). A practical approach to the assessment and quantification of content validity. In S. P. Desselle, V. García-Cárdenas, C. Anderson, P. Aslani, A. M. H. Chen, & T. F. Chen (Eds.), *Contemporary research methods in pharmacy and health services* (pp. 583–599). Academic Press. https://doi.org/10.1016/B978-0-323-91888-6.00013-2

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Badrinarayan, A., Lyons, S., Miranda, A., Klyachkina, A., & Leonard, P. (2025). Leveraging students' cultural and linguistic assets for assessment: A Framework for culturally conscious assessment in the Chicago Public Schools. *Educational Assessment*, *30*(2), 115–140. https://doi.org/10.1080/10627197.2025.2497775

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. The Guilford Press.

Banitalebi, Z., Estaji, M., & Brown, G. T. (2025). Measuring teacher assessment literacy in digital environments: Development and validation of a scenario-based instrument. *Educational Technology & Society*, 28(2), 169-215. https://doi.org/10.30191/ETS.202504_28(2).RP10

Bennett, R. E. (2023). Toward a theory of socioculturally responsive assessment. *Educational Assessment*, *28*(2), 83-104. https://doi.org/10.1080/10627197.2023.2202312

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Bonner, S. M., & Chen, P. P. (2021). Development and validation of the survey of unorthodox grading beliefs for teachers and teacher candidates. *Journal of Psychoeducational Assessment*, 39(6), 746-760. https://doi.org/10.1177/07342829211015462

Centro de Perfeccionamiento, Experimentación e Investigaciones Pedagógicas (CPEIP). (2022). *Estándares pedagógicos y disciplinarios para carreras de pedagogía en educación general básica [Standards]*. Ministerio de Educación. https://estandaresdocentes.mineduc.cl/Categoria-p/pedagogias/

Chang, W. C., & Cochran-Smith, M. (2022). Learning to teach for equity, social justice, and/or diversity: Do the measures measure up? *Journal of Teacher Education*. https://doi.org/10.1177/00224871221075284

Chang, W. C. C., Ludlow, L. H., Grudnoff, L., Ell, F., Haigh, M., Hill, M., & Cochran-Smith, M. (2019). Measuring the complexity of teaching practice for equity: Development of a scenario-format scale. *Teaching and Teacher Education*, *82*, 69-85. https://doi.org/10.1016/j.tate.2019.03.004

D'Agostino, J. V., Welsh, M. E., Cimetta, A. D., Falco, L. D., Smith, S., VanWinkle, W. H., & Powers, S. J. (2008). The rating and matching item-objective alignment methods. *Applied Measurement in Education*, *21*(1), 1-21. https://doi.org/10.1080/08957340701580728

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016). Approaches to classroom assessment inventory: A new instrument to support teacher assessment literacy. *Educational Assessment*, *21*(4), 248–266. https://doi.org/10.1080/10627197.2016.1236677

DeLuca, C., Coombs, A., MacGregor, S., & Rasooli, A. (2019). Toward a differential and situated view of assessment literacy: Studying teachers' responses to classroom assessment scenarios. In Frontiers in Education, 4, 94. https://doi.org/10.3389/feduc.2019.00094

Earl, L. M. (2013). *Assessment as learning: Using classroom assessment to maximize student learning* (2nd ed.). Corwin Press.

Gerst, M. D., Dillard, M., & Loerzel, J. (2025). Methodological recommendations for content validation of community resilience indicators. *Natural Hazards Review*, 26(2), 04025010. https://doi.org/10.1061/NHREFO.NHENG-2179

Heritage, M., & Wylie, C. (2018). Reaping the benefits of assessment for learning: Achievement, identity, and equity. *ZDM*, 50(4), 729–741. https://doi.org/10.1007/s11858-018-0943-3

McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). *Instrument development in the affective domain*. Springer.

Nortvedt, G. A., Wiese, E., Brown, M., Burns, D., McNamara, G., O'Hara, J., & Taneri, P. O. (2020). Aiding culturally responsive assessment in schools in a globalising world. *Educational Assessment, Evaluation and Accountability*, 32(1), 5-27. https://doi.org/10.1007/s11092-020-09316-w

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, *29*(5), 489–497. https://doi.org/10.1002/nur.20147

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

Randall, J., Poe, M., Oliveri, M. E., & Slomp, D. (2024a). Justice-oriented, antiracist validation: Continuing to disrupt white supremacy in assessment practices. *Educational Assessment*, *29*(1), 1-20. https://doi.org/10.1080/10627197.2022.2042682

Randall, J., Poe, M., Slomp, D., & Oliveri, M. E. (2024b). Our validity looks like justice. Does yours? *Language Testing, 41*(1), 203–219. https://doi.org/10.1177/02655322231202947

Sireci, S., & Benítez, I. (2023). Evidence for test validation: A guide for practitioners. *Psicothema*, *35*(3), 217–226. https://dx.doi.org/10.7334/psicothema2022.477

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment*, *5*(4), 299–321. https://doi.org/10.1207/s15326977ea0504_2

Solano-Flores, G., & Nelson-Barber, N. (1999, March 28 – 31). Developing culturally responsive science assessments. Workshop paper presented at the 1999 Meeting of the National Association for the Research of Science Teaching. Boston, Massachusetts.

Spoto, A., Nucci, M., Prunetti, E., & Vicovaro, M. (2025). Improving content validity evaluation of assessment instruments through formal content validity analysis. *Psychological Methods,* 30(2), 203–222. https://doi.org/10.1037/met0000545

St. Marie, B., Jimmerson, A., Perkhounkova, Y., & Herr, K. (2020). Developing and establishing content validity of vignettes for health care education and research. Western Journal of nursing research, *43*(7), 677-685. https://doi.org/10.1177/0193945920969693

Valenzuela, J. P., Bellei, C., & Ríos, D. D. L. (2014). Socioeconomic school segregation in a market-oriented educational system. The case of Chile. Journal of education Policy, 29(2), 217-241. https://doi.org/10.1080/02680939.2013.806995

Wolf, M. K., Sova, L., Janssen, G., López, A. A., Gooch, R. M., Pooler, E., & Lee, J. (2025). Equity for multilingual learners: Leveraging formative assessment and socioculturally responsive assessment principles. *Bilingual Research Journal*, 1–21. https://doi.org/10.1080/15235882.2025.2492675

Appendix A

Standards for the Teaching Profession for Elementary Education Careers

STANDARD 4: ASSESSMENT PLANNING

Plan the assessment, incorporating various modalities that allow the production of evidence aligned with the learning objectives, monitor the level of achievement of these and provide feedback to their students.

Description

The graduated teacher plans the evaluation process, understanding that this is part of the teaching preparation process, in order to have quality, timely and pertinent evidence, according to the disciplinary and transversal learning objectives defined in the planning. In order to fulfill its double purpose –formative and summative–, it selects and designs various evaluation instruments and strategies that allow the analysis, monitoring, feedback and qualification of the level of learning achievement of its students. It proposes activities to evaluate and for the self and co-evaluation of the students, that respond to their specific needs and that provide information for feedback. Analyzes and critically reflects on the results of the evaluations from an ethical dimension, to identify possible interpretative biases that create barriers to the learning of their students. Likewise, it analyzes and reflects on the evaluation process and the quality of its evidence, to plan and readjust their assessment practices.

Descriptors

Focus: Construction and collection of learning evidence

4.1 Plans the assessment considering the appropriate moments and various techniques and instruments for it, including self- and peer-evaluation by their students, so that everyone can demonstrate what they have learned, and their results provide timely and relevant information regarding progress and achievement of the learning objectives.

4.2 Design assessments that allow diversifying and broadening the evidence, formative, to monitor and track learning, and summative, to collect information on the level of achievement of learning objectives.

4.3 Adapt, in collaboration with their peers, the evaluation strategies and procedures to diagnose, give timely feedback and qualify the learning of students who require specific support.

4.4 Builds, selects and adapts evaluation criteria consistent with the learning objectives, to guide their observation.

Focus: Analysis of learning evidence and feedback

4.5 Analyzes the data and evidence provided by the assessments, to identify gaps between the expected learning and the actually achieved, as well as changes with respect to previous assessments, and to improve the evaluation procedures and techniques used.

4.6 Plan different instances – oral and written, individual and group – to provide students with timely feedback that helps them reflect on and regulate their own learning.

4.7 Addresses the ethical dimension of evaluation and the use of evidence, to interpret the results and detect biases that may reflect inequities in learning opportunities.

4.8 Determines procedures to qualify the performance of its students, using evaluation criteria in a fair, rigorous and transparent way, and that accurately communicate through a number, symbol or concept, the level of learning achievement.

STANDARD 9: ASSESSMENT AND FEEDBACK FOR LEARNING

Use assessment and feedback to monitor and enhance learning, based on evaluation criteria and relevant evidence, adjusting supports in a timely and specific manner, and promoting self-assessment in students.

Description

The graduated teacher uses a variety of assessment and feedback strategies during learning activities, which allow them to obtain evidence of the achievement of the objectives, make decisions and reduce learning gaps. The teacher continuously provides timely and descriptive feedback to his students, suggesting options for them to continue learning, and encouraging perseverance in challenging tasks, in order to maintain their involvement until they are completed. In addition, they communicate specific and precise indicators of success, so that students can monitor their progress, adjust their learning process and use various self- and peer-assessment strategies, so that they acquire autonomy and responsibility. As they progress in their development, they encourage their students to determine evaluation indicators and criteria in order to achieve a better understanding of the expectations and to promote greater transparency in the assessment process.

Descriptors

Focus: Criteria for assessing and monitoring of learning

9.1 Explain the assessment criteria to their students, aligned with the learning objective, giving them examples of the expected performance so that they gradually participate in the definition of these criteria.

9.2 Checks during class, through questions or relevant activities, the level of understanding of their students and identifies difficulties and errors to redirect teaching.

9.3 Use the monitoring results to carry out additional and differentiated activities or to reorganize learning experiences, providing support according to the rhythms, characteristics and needs of their students.

Focus: Feedback

9.4 Offers students descriptive feedback in a timely manner, based on assessment criteria and indicators, so that they have differentiated information on the levels of achievement of the knowledge, skills and attitudes defined in the evaluated learning objectives; and to establish strategies that allow them to overcome the gaps.

9.5 Communicate to students the grades obtained, making sure they understand the number, symbol or concept that represents the level of learning achievement, so that they define their own improvement goals and commit to the following learning processes.

9.6 Develop timely strategies to address the potential effects of assessment and grades on students' emotions and motivation, in order to protect their academic self-esteem and promote perseverance in learning the discipline taught.

Focus: Self-assessment of learning

9.7 Teaches and guides students to use criteria, indicators and attributes for self- and peer-assessment processes, with the purpose of observing their learning and that of others, and determining the learning achieved and those that require improvement.

Appendix B

Table B1. Vignette-level Alignment Results (n=8)

Scenario #	Descriptor	Match to Descriptor	Focus Area	Match to Focus Area
2	4.1	0.69	AP1	1.00
6	4.2	0.56	AP1	0.75
18	4.3	0.25	AP1	0.25
13	4.4	0.00	AP1	0.56
12	4.5	0.63	AP2	0.94
3	4.6	0.44	AP2	0.44
7	4.7	0.75	AP2	0.88
9	4.7	0.75	AP2	0.94
14	4.8	0.19	AP2	0.31
8	9.1	0.81	FFL1	0.81
1	9.2	0.88	FFL1	0.88
5	9.3	0.44	FFL1	0.44
15	9.3	0.50	FFL1	0.57
17	9.3	0.75	FFL1	0.75
10	9.4	0.44	FFL2	0.63
16	9.5	0.81	FFL2	0.81
11	9.6	0.63	FFL2	0.81
4	9.7	0.81	FFL3	0.81

Note. Values represent match rates between scenarios, descriptors, and focus areas.

Appendix C

Table C1. Item-level Alignment Results: Scenario 1

Scenario	Item	Approach	Match Rate	Mean Relevance	Proportion Pretty/Very Sure of Match	Content Validity Index
You are a 3rd grade teacher. You are planning the writing assessments you will administer throughout the year for your language class.	Ask students to create a plan for writing a text that includes a timeline for meeting specific goals.	Assessment as learning	0.63	3.00	0.50	0.63
	Ask your students to write a text only requiring skilled students to write in formal Spanish.	Supremacist	0.38	2.00	0.38	0.25
	Organize oral activities where students share experiences they have had with their community and then write about one of them.	Culturally sustaining assessment	0.38	3.00	0.38	0.38
	Plan assessments in which students draft, receive your feedback, and submit a final version.	Assessment for learning	0.50	2.75	0.50	0.50
	Assign a writing task allowing your students to write in whatever language or languages they choose as long as they meet the requirements of the assignment.	Culturally responsive assessment	0.88	2.71	0.63	0.75
	Plan a graded assessment at the end of each unit.	Assessment of learning	1.00	2.88	1.00	1.00

Table C2. Item-level Alignment Results: Scenario 2

Scenario	Item	Approach	Match Rate	Mean Relevance	Proportion Pretty/Very Sure of Match	Content Validity Index
You have developed three quizzes with your grade level team to assess your students' skill in adding fractions. Your team noticed that a group of students who seemed to understand the contents during lessons performed below the learning expectations on the three quizzes. You want to get better information about these students' understanding of fractions.	Use the results because it is important that students know how to show their learning on tests.	Supremacist	0.50	2.00	0.50	0.25
	Analyze the assessment to identify what errors were most often made by the students-so that you can reinforce concepts.	Assessment for learning	0.88	2.86	0.88	0.88
	Ask students to explain how they arrived to their answer to better understand what is unclear.	Assessment as learning	0.25	3.00	0.25	0.25
	Let students retake the test using an assessment that covers the same content.	Assessment of learning	0.75	2.83	0.75	0.75
	Review the test to see if it is suitable for all students in the same way.	Culturally responsive assessment	0.75	3.00	0.75	0.75
	Reassess the students by having them write and solve problems with fractions that addresses an issue they experience in their family.	Culturally sustaining assessment	0.38	3.00	0.38	0.38

Table C3. Item-level Alignment Results: Scenario 3

Scenario	Item	Approach	Match Rate	Mean Relevance	Proportion Pretty/Very Sure of Match	Content Validity Index
Paulina, a student from Chile, and Maria, a new student from Venezuela, got the same score on a standardized test in 7th grade. In 8th grade, Paulina got a much higher score than Maria. You want to better understand this result.	Analyze the vocabulary of the test to see if it may have affected Maria's performance.	Culturally responsive assessment	0.50	3.00	0.50	0.50
	Congratulate Paulina because she prepared better for the test than Maria.	Assessment of learning	0.00	0.00	0.00	0.00
	Review the test together with Maria to identify how far she is from meeting the expectations and in which aspects she needs additional support.	Assessment for learning	0.50	3.00	0.50	0.50
	Group Maria with other students who performed poorly and give them easier tests in the future.	Supremacist	0.38	2.33	0.38	0.25
	Have the students work on projects that use the tested math skills to address a social issue of their choosing and see if Maria performs better using this approach.	Culturally sustaining assessment	0.75	3.00	0.75	0.75
	Ask Paulina and Maria to compare their answers and discuss what they have understood differently	Assessment as learning	0.75	2.67	0.75	0.75

Table C4. Item-level Alignment Results: Scenario 4

Scenario	Item	Approach	Match Rate	Mean Relevance	Proportion Pretty/Very Sure of Match	Content Validity Index
You are a 6th-grade teacher, and some of your immigrant students performed worse than other students on a reading comprehension assessment in which they read a text about Chile's Independence Day celebration. The students have consistently demonstrated strong reading skills on classroom assessments.	Don't use the results for grading and continue with your next session.	Assessment of learning	0.00	0.00	0.00	0.00
	Ask students to share what they found confusing to help them reflect on their learning.	Assessment as learning	0.63	2.60	0.63	0.63
	Have students retell the text, including key celebration activities, from the perspective of an indigenous Chilean to explore how others may feel about the celebration.	Culturally sustaining assessment	0.38	3.00	0.38	0.38
	Gather additional evidence to identify student strengths and areas where they may need more support, including if some students were disadvantaged by the topic of the reading.	Assessment for learning	0.25	3.00	0.25	0.13
	Use the test results to grade all students because everyone should be familiar with Chilean culture.	Supremacist	0.63	2.00	0.63	0.25
	Have students take a new reading comprehension test using texts that address independence days in a culture you haven't taught.	Culturally responsive assessment	0.00	0.00	0.00	0.00

Table C5. Item-level Alignment Results: Scenario 5

Scenario	Item	Approach	Match Rate	Mean Relevance	Proportion Pretty/Very Sure of Match	Content Validity Index
You have developed a rubric to assess the writing skills of your 4th-grade students. You want to make sure that the students understand the rubric.	Ask students to suggest rubric criteria that embrace the writing styles used in a variety of communities.	Culturally sustaining assessment	0.13	3.00	0.13	0.13
	Analyze if some of the rubric criteria are focused on following directions instead of writing skills (e.g., writing a certain number of sentences instead of including a well-constructed topic sentence) to ensure its fair to all students.	Culturally responsive assessment	0.25	2.50	0.13	0.25
	Explain to students that higher levels on the rubric represent end of the year expectations.	Assessment of learning	0.50	2.67	0.38	0.38
	Create the rubric with your students so they can analyze their work knowing what the expectations are.	Assessment as learning	0.63	3.00	0.63	0.63
	Would assume that fourth-grade students should already know how to use rubrics and need no explanation.	Supremacist	0.75	1.60	0.50	0.25
	Discuss students' work with them by comparing their work to the rubric and ask them to rewrite it.	Assessment for learning	0.38	3.00	0.38	0.38

Table C6. Item-level Alignment Results: Scenario 6

Scenario	Item	Approach	Match Rate	Mean Relevance	Proportion Pretty/Very Sure of Match	Content Validity Index
You give an assessment to determine which 1st-grade students need support in developing their decoding skills. Students performed very differently, with some students doing very well and others struggling with differentiating the sound /r/ and /rr.	Use the results to create a learning group to practice these skills.	Assessment for learning	0.63	3.00	0.63	0.63
	Recommend that the student is enrolled in a pull-out reading program specialized for students with difficulties.	Supremacist	0.50	2.67	0.25	0.38
	Have students provide feedback on a friend's efforts to sort pictures of things that are pronounced /r/ or /rr/.	Assessment as learning	0.63	3.00	0.38	0.63
	Discuss with parents that their child has difficulty reading words.	Assessment of learning	0.13	3.00	0.13	0.13
	Determine if students can correctly match words with /r/ and /rr/ to pictures even if they can't roll their rs well.	Culturally responsive assessment	0.63	3.00	0.25	0.50
	Encourage students to share how to pronounce words that they use at home with /r/ and /rr/ to explore how these letters can be read.	Culturally sustaining assessment	0.13	3.00	0.13	0.13

Table C7. Item-level Alignment Results: Scenario 7

Scenario	Item	Approach	Match Rate	Mean Relevance	Proportion Pretty/Very Sure of Match	Content Validity Index
You just finished grading the first test of the year. Many students are immigrant students from countries with different grading systems. You want to help students understand their grades and reflect on performance.	Explain to your students that the numeric grade relates to a descriptor (e.g., 7 is excellent; 6 is very good) and that a higher number show that they are meeting learning objectives.	Assessment of learning	0.25	2.50	0.25	0.25
	Put written feedback on assessments to tell students what they need to do better to receive a higher numeric grade on the next assessment.	Assessment for learning	0.38	2.67	0.25	0.38
	Have students correct their tests and think through how they lost points to see how errors affected how their grade was calculated.	Assessment as learning	0.71	2.60	0.57	0.71
	Explain what kinds of answers are considered high-quality in Chile.	Supremacist	0.50	1.75	0.57	0.25
	Help immigrant students translate Chilean grades into those used in their home countries to help them understand how well they are doing.	Culturally responsive assessment	0.71	2.60	0.43	0.71
	Ask your immigrant students to explain to the class how the grading system works in their country, what information it provides, and how they use it.	Culturally sustaining assessment	0.43	3.00	0.43	0.43

Table C8. Item-level Alignment Results: Scenario 8

Scenario	Item	Approach	Match Rate	Mean Relevance	Proportion Pretty/Very Sure of Match	Content Validity Index
You are planning a peer assessment for a writing activity where each student will write a description of a painting. You want to guide students so they can provide each other with useful feedback.	Teach students how to apply the rubric criteria to different responses to show that the criteria can be met in different ways.	Culturally responsive assessment	0.38	3.00	0.38	0.25
	Discuss the ways in which people's different backgrounds and ways of speaking make written descriptions more interesting. Jointly create with students' rubric criteria that reward the use of slang, dialect, multilingual terms, etc.	Culturally sustaining assessment	0.50	2.67	0.50	0.38
	Teach your students how to interpret rubric scores.	Assessment of learning	0.25	1.50	0.25	0.13
	Model how to provide useful feedback using a rubric, identifying ways in which the writing is and is not consistent with the rubric.	Assessment as learning	0.50	3.00	0.38	0.38
	Explain to students that the rubric defines what is expected of a good writer.	Supremacist	0.38	3.00	0.25	0.25
	Collect the feedback students provide to each other and use it to design lessons on giving helpful feedback.	Assessment for learning	0.75	3.00	0.75	0.63

Table C9. Item-level Alignment Results: Scenario 9

Scenario	Item	Approach	Match Rate	Mean Relevance	Proportion Pretty/Very Sure of Match	Content Validity Index
When reading student's work, you notice that many students write "schaleco" instead of "chaleco." You think that this misspelling may be related to how they pronounce the word.	Use the instance to discuss how differences in how people say, and spell words can lead to injustice.	Culturally sustaining assessment	0.86	3.75	0.86	0.71
	Because words can be pronounced in different ways, have students make a list of words that are spelled differently from how people may say them.	Culturally responsive assessment	0.43	2.50	0.29	0.29
	Emphasize that they pronounced the word incorrectly and this can cause people to think they don't know how to speak properly and disadvantage them.	Supremacist	0.71	2.33	0.57	0.29
	Ask students to work together to identify other words that are pronounced in different ways and are therefore spelled incorrectly in their own writing.	Assessment as learning	0.43	4.00	0.29	0.43
	Since this isn't a spelling quiz and you think they spelled the word the way they pronounce it, don´t deduct points.	Assessment for learning	0.14	3.00	0.14	0.14
	Model the correct way to pronounce the word to show them why it is spelled with "ch".	Assessment of learning	0.14	0.00	0.00	0.00

Table 10. Item-level Alignment Results: Scenario 10

Scenario	Item	Approach	Match Rate	Mean Relevance	Proportion Pretty/Very Sure of Match	Content Validity Index
Your 4th grade class is diverse. It includes immigrants and students from the Mapuche community. You are in the middle of the unit about traditional legends and want to identify what students have learned so far.	Ask students from the Mapuche community to share a legend that their family taught them and to share what phenomena it helps to explain.	Culturally sustaining assessment	0.50	2.75	0.38	0.50
	Highlight that legends may vary across different regions and ask them to identify the importance of a specific legend to the region it comes from.	Culturally responsive assessment	0.38	3.00	0.25	0.38
	Give all of your students some traditional Chilean legends so they can read them at home. Ask them to name the area or person in history that they focus on.	Supremacist	0.13	3.00	0.13	0.13
	Ask your students to review a legend they created and to analyze if it has all the characteristics studied.	Assessment as learning	0.50	3.00	0.50	0.50
	Ask the students to tell the class a legend to help you identify what characteristics of a legend they are missing so you can reteach them.	Assessment for learning	0.75	3.00	0.75	0.75
	Apply a summative test to identify how much they know.	Assessment of learning	0.88	2.43	0.88	0.75

Table 11. Item-level Alignment Results: Scenario 11

Scenario	Item	Approach	Match Rate	Mean Relevance	Proportion Pretty/Very Sure of Match	Content Validity Index
Your students are working on math problems that require them to explain their thinking. Your Haitian students, who are new to Chile, are struggling to write their explanations in Spanish. You want to support their mathematical thinking.	Allow all students complement their Spanish explanation with other languages and/or visual representations.	Culturally responsive assessment	0.50	3.00	0.38	0.50
	Ask students if they know a way of solving the problem in a different country or community and to share these other approaches with the class.	Culturally sustaining assessment	0.50	2.75	0.50	0.50
	Approach scoring in a way that captures quality of mathematical thinking without docking points for grammar or spelling.	Assessment of learning	0.38	2.67	0.38	0.38
	Ask students to share explanations to help them identify the characteristics of strong mathematical explanations.	Assessment as learning	0.63	2.80	0.63	0.63
	Apply the same test to all the students and require them to answer in the same way so that results are comparable.	Supremacist	0.63	1.80	0.63	0.25
	Gather more evidence by applying similar exercises to determine how to better support students in writing mathematical explanations.	Assessment for learning	0.63	3.00	0.63	0.63

Table 12. Item-level Alignment Results: Scenario 12

Scenario	Item	Approach	Match Rate	Mean Relevance	Proportion Pretty/Very Sure of Match	Content Validity Index
You are teaching a unit on the importance of the appropriate use of natural resources. Your students watched and discussed a documentary and wrote a journal entry about their experiences taking care of nature. In the next activity, they will work in groups to present a poster about how to take care of the environment. You want to include an activity in which they reflect on their grading.	Focus your grading on the final product of the poster using criteria that you explain to the students.	Assessment of learning	0.75	2.83	0.63	0.75
	Provide feedback both to individual students and to the whole group as they work on each activity.	Assessment for learning	1.00	0.00	1.00	1.00
	Ask students to evaluate the evidence supporting both their own opinion and others' opinions.	Assessment as learning	0.88	2.71	0.88	0.88
	Provide feedback to help students learn to communicate like scientists orally, in writing, or on a poster.	Supremacist	0.25	1.50	0.13	0.13
	Set the expectation that all modalities of expression (e.g. orally-recorded or live-, drawings, writing) contribute valuable academic knowledge.	Culturally responsive assessment	0.50	2.75	0.50	0.50
	Ask questions during all the activities to highlight the variety of ways that different cultures protect natural resources.	Culturally sustaining assessment	0.50	3.00	0.38	0.50

Table 13. Item-level Alignment Results: Scenario 13

Scenario	Item	Approach	Match Rate	Mean Relevance	Proportion Pretty/Very Sure of Match	Content Validity Index
You and your colleagues are developing grading policies for the coming school year and have noticed that some students do not do well when they are tested at the beginning of the marking period, requiring you to reteach some concepts. You want to make some suggestions to the team.	Recommend that the team uses assessments to inform instructional choices throughout the year, but not to use beginning of marking period assessment results to determine report card grades.	Assessment for learning	0.75	0.00	0.75	0.75
	Recommend that teachers put students who don't do well at the beginning of the year in a special group to reteach last year's concepts.	Supremacist	0.38	3.00	0.38	0.29
	Recommend that the team develops procedures to help students use assessments to reflect on what they know and what they need to work.	Assessment as learning	0.88	2.86	0.88	0.88
	Recommend that the team progressively weights later assessments more in calculating report card grades.	Assessment of learning	0.75	2.67	0.63	0.63
	Suggest that the team identifies the core characteristics of grade level performance and generates work samples to share with students and parents that demonstrate that proficiency doesn't have to be presented in one specific way.	Culturally responsive assessment	0.50	3.00	0.38	0.50
	Recommend that students are provided with multiple opportunities throughout the year to design their own end-of-unit assessments that apply course content to issues important to them.	Culturally sustaining assessment	0.00	0.00	0.00	0.00

Table 14. Item-level Alignment Results: Scenario 14

Scenario	Item	Approach	Match Rate	Mean Relevance	Proportion Pretty/Very Sure of Match	Content Validity Index
While teaching a unit about fractions, you ask your students to order the fractions with the same numerators but different denominators from largest to smallest. About 15% of your students put the fraction with the largest denominator in the space for the biggest number. You wish you could understand the source of their confusion.	Have students explain the way that their families talk about fractions (e.g. two-thirds or thirds-two) and explain how the same fraction can be written multiple ways.	Culturally sustaining assessment	0.38	2.67	0.13	0.38
	Reinforce the concept of numerator and denominator using a donut to represent the parts.	Supremacist	0.13	3.00	0.13	0.13
	Keep moving forward with the content because not everyone performs well on every activity.	Assessment of learning	0.13	2.00	0.00	0.13
	Have students write down questions that they have about fractions.	Assessment as learning	0.88	2.57	0.71	0.88
	Ask the students questions to help you reflect on the ways in which you could do a better job of explaining fractions.	Assessment for learning	0.63	2.80	0.63	0.63
	Reinforce the concept of numerator and denominator using different representations that can resonate to all students.	Culturally responsive assessment	0.75	2.83	0.63	0.75

Table 15. Item-level Alignment Results: Scenario 15

Scenario	Item	Approach	Match Rate	Mean Relevance	Proportion Pretty/Very Sure of Match	Content Validity Index
Your colleague asks for advice on how to provide feedback to her fourth-grade students who are finishing a unit on fables. She is worried because a Haitian student wrote a draft of a fable that included all of the important parts, except it did not teach a lesson about the world. His tale included some words written in Creole.	Suggest that she tell the student that he mostly met the criteria for writing a fable but remind him to follow the instructions more carefully next time to receive a higher score.	Assessment of learning	0.63	2.40	0.50	0.63
	Suggest that students switch papers and compare them with a check list to identify needed improvements.	Assessment as learning	0.38	3.00	0.38	0.38
	Suggest that she tell the student that being multilingual is impressive and ask him to teach the class about the Creole terms he used.	Culturally sustaining assessment	0.50	2.75	0.50	0.50
	Recommend that she reviews all parts of a fable with the student, using examples to show how lessons are presented in fables.	Assessment for learning	0.50	3.00	0.50	0.50
	Recommend that she tell the student to write it again replacing the Creole words with Spanish words.	Supremacist	0.50	2.00	0.38	0.25
	Suggest that she allow him to use Creole words to help express more nuanced meanings and add a glossary with the definition of the terms.	Culturally responsive assessment	0.50	2.50	0.50	0.50

Table 16. Item-level Alignment Results: Scenario 16

Scenario	Item	Approach	Match Rate	Mean Relevance	Proportion Pretty/Very Sure of Match	Content Validity Index
All but two of your 4th-grade students did well on the most recent test. You want to support the students to help them persevere.	Tell them that this is only one piece of evidence to examine their learning and that they will have opportunities to show what they know in many different ways.	Culturally responsive assessment	0.38	3.00	0.25	0.38
	Encourage them to study more for their next assessment so they can get a good grade.	Assessment of learning	0.38	2.33	0.38	0.38
	Explain that assessments can help to identify areas where students need more support and that you will use the results to help them learn.	Assessment for learning	0.88	2.86	0.88	0.86
	Tell them the grade is a warning that they aren't studying enough.	Supremacist	0.50	1.75	0.25	0.25
	Present examples of how students in different countries express their knowledge and are assessed in different ways and incorporate some of these other assessment approaches in future assessments.	Culturally sustaining assessment	0.63	2.60	0.63	0.57
	Talk to students to help them reflect on their learning and to help them prepare for a test retake.	Assessment as learning	0.50	3.00	0.50	0.43

Table 17. Item-level Alignment Results: Scenario 17

Scenario	Item	Approach	Match Rate	Mean Relevance	Proportion Pretty/Very Sure of Match	Content Validity Index
You have a diverse classroom where one of your students is from a middle eastern country. You applied a test and realized that they answered "thirds-two" instead of "two-thirds", a common way of expressing fractions in their home country.	Explain how fractions are written in Chile and require the student to express their knowledge in the Chilean way.	Supremacist	1.00	2.13	0.88	0.63
	Try to integrate the students' way of expressing fractions into the next lesson.	Assessment for learning	0.13	3.00	0.13	0.13
	Have students develop directions to translate Chilean fractions into Arabic fractions and vice-versa.	Assessment as learning	0.25	2.50	0.13	0.25
	Give him full credit because he understands the mathematical concept.	Assessment of learning	0.50	2.75	0.50	0.50
	Work to separate what the students know about fractions from the way they express themselves when you grade them because mathematical skill is different than Spanish expression.	Culturally responsive assessment	0.38	2.67	0.38	0.38
	Ask the student to teach his classmates the Arabic way of expressing fractions and have students apply it.	Culturally sustaining assessment	0.38	3.00	0.38	0.38

Table 18. Item-level Alignment Results: Scenario 18

Scenario	Item	Approach	Match Rate	Mean Relevance	Proportion Pretty/Very Sure of Match	Content Validity Index
You are finishing the grammar unit on the use of articles. You give students a test in which they must write the correct article before the noun. Several of your students wrote the article "la" before the word "calor". You want to determine whether the student has not mastered with feminine/masculine nouns or whether this reflects a difference in how "calor" is used in different communities.	Have students read a book whose characters use informal language and identify the ways in which atypical uses of "la" and "el" make the story interesting.	Culturally responsive assessment	0.25	2.50	0.13	0.14
	Ask them to review which nouns use feminine/masculine articles at home and allow them to retake the test.	Assessment of learning	0.63	2.40	0.38	0.57
	Reassess the students using different nouns. Use the results to inform how you might differentiate instruction to support the students who answered incorrectly.	Assessment for learning	0.63	2.80	0.63	0.57
	Have students write a story in which they apply "la and el" in the ways that they use them at home to make the story more authentic.	Culturally sustaining assessment	0.13	3.00	0.13	0.00
	Explain "calor" is an abstract masculine noun that corresponds to the article "el."	Supremacist	0.50	2.00	0.38	0.29
	Ask them to practice sorting words that do not end in "a" or "o" into masculine or feminine and to check the accuracy of their sorting.	Assessment as learning	0.00	0.00	0.00	0.00