frontiers in Education

METHODS
published: 07 June 2019
doi: 10.3389/feduc.2019.00043

Examining Cultural Responsiveness in Large-Scale Assessment: The Matrix of Evidence for Validity Argumentation

*Guillermo Solano-Flores**

Graduate School of Education, Stanford University, Stanford, CA, United States

As cultural products, assessment systems, and instruments are a reflection of the cultures in which they originate and reproduce the characteristics of those cultures. Accordingly, cultural responsiveness in large-scale assessment concerns the entire process of assessment, not simply its tests or the analysis of test scores. This paper focuses on validity as argumentation in the testing of culturally diverse populations in both national and international large-scale assessment programs. It examines the intersection of *m* procedural assumptions (conceptual and methodological criteria that need to be met in order to validly, fairly test culturally diverse populations), and *n* components of the process of assessment. The paper explains the development and use of the *matrix of evidence for validity argumentation.* In this matrix, a cell *[i, j]* contains confirming and disconfirming evidence that Procedural Assumption *i* is met by the practices enacted and the artifacts used or generated in Assessment Process Component *j*. An argument of validity is constructed by integrating the information contained in the cells of the matrix. A series of examples illustrate how each procedural assumption intersects with each assessment process component in the matrix. These examples show the ubiquity of cultural issues in assessment. Attaining cultural responsiveness in large-scale national and international assessment programs entails a serious systemic, societal, global effort.

Keywords: assessment, validity, cultural responsiveness, large scale testing, matrix of evidence

OPEN ACCESS

Edited by:
Mustafa Asil,
University of Otago, New Zealand

Reviewed by:
Giray Berberoglu,
Başkent University, Turkey
Jeffrey K. Smith,
University of Otago, New Zealand

***Correspondence:**
Guillermo Solano-Flores
gsolanof@stanford.edu

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *09 March 2019*
Accepted: *07 May 2019*
Published: *07 June 2019*

Citation:
Solano-Flores G (2019) Examining Cultural Responsiveness in Large-Scale Assessment: The Matrix of Evidence for Validity Argumentation. Front. Educ. 4:43. doi: 10.3389/feduc.2019.00043

INTRODUCTION

Every human invention (e.g., a hamburger, a cell phone, a law, or a science curriculum) is a cultural product that originated and evolved as a result of certain needs in a society (e.g., eating, communicating, controlling, teaching) and which reflects the history, thinking, experience, values, and forms of doing things of those who created it.

Underlying the notion of standardization in tests (i.e., the establishment of uniform sets of observations and procedures; see Geisinger, 2010) is the implicit assumption that all individuals being tested with the same instrument share the same set of cultural experiences and are equally familiar with the features of items (Solano-Flores, 2011). Among other, these features include wording, conventions used to represent information, and contextual information such as stories, situations, and fictional characters used to situate problems with the intent to make items meaningful to the examinee (Ruiz-Primo and Li, 2015).

As with any invention, testing, tests, and assessment systems are cultural products (see Cole, 1999). They are or contain practices and artifacts that reflect culturally-determined world views, sets of knowledge and skills valued, and ways of representing information, building arguments, and asking and responding to questions. Thus, when students from different cultural backgrounds are assessed with the same instrument, fairness the validity of the interpretations of test scores becomes an issue because they may not share the same set of cultural experiences and, therefore, may not have equal access to the content of items (see Camilli, 2006). In this case, test score differences are attributable, at least to some extent, to cultural differences rather than differences on the target knowledge or skills (American Educational Research Association et al., 2014).

For decades, efforts to minimize the effects of cultural differences in large-scale assessment have focused on activities such as item writing, item translation/adaptation, test review, the use of testing accommodations, and the analysis of item bias. At the core of these efforts is the need for ensuring the equivalence of constructs in ways that the scores produced have similar meanings across cultural groups (van de Vijver and Tanzer, 2004; International Test Commission, 2018). Unfortunately, while such efforts are necessary, they may be insufficient to ensure valid, fair testing of culturally diverse populations due to their focus on isolated aspects of assessment. This limited focus may be a reflection of the fact that the conceptual connection between cultural responsiveness and the process of assessment needs to be more explicit across all its practices and artifacts. Unfortunately, while important conceptual developments of cultural responsiveness have focused on teaching (see Ladson-Billings, 1995; Gay, 2000; Aronson and Laughter, 2016), little attention has been paid to the systemic aspects of assessment.

This paper addresses the need for conceptualizing and operationalizing cultural responsiveness in large-scale assessment from the perspective of validity. Building on the notion of validity as argumentation (Kane, 1992, 2006; Mislevy, 1994; Haertel, 1999), I propose the *matrix of evidence for validity argumentation*. This matrix is intended to allow examination of validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick, 1989, p. 13). More specifically, this matrix is a tool for organizing information on whether and how the practices and artifacts involved in the process of assessment meet a series of procedural assumptions (criteria of conceptual and methodological rigor) relevant to validly assessing culturally diverse populations. Using examples from multiple large-scale assessment programs and different culturally diverse populations, I discuss and illustrate the intersection of procedural assumptions and assessment process components. This intersection is key to operationalizing (and, ultimately elucidating what it takes to attain) cultural responsiveness in large-scale assessment.

For the purposes of this discussion, the term *cultural group* is used very broadly, for example based on race/ethnicity, language background, country, or socio-economic status. While none of them is culture by itself, these factors and their combinations shape culture. For example, races have different access to different sets of experiences and opportunities, language is a cultural product, and countries have different histories, identities, and legal and educational systems. Depending on the large-scale assessment program, cultural groups may be countries or linguistic or ethnic groups. Accordingly, the terms, *cultural diversity* and *culturally diverse population* are used in cases in which student populations contain several cultural groups.

METHODS

From the perspective of validity as argumentation, examining cultural responsiveness involves: (1) identifying a set of relevant components in the process of assessment, (2) establishing a set of procedural assumptions critical to valid, fair testing of a culturally diverse population, and (3) examining the intersection of those components and those assumptions by constructing a matrix of evidence for validity argumentation.

Identifying Assessment Process Components

While large-scale assessments may have very different characteristics, it is possible to identify a common set of components in the process of assessment. These assessment process components comprise both activities inherent to testing (e.g., developing and administering tests; analyzing and reporting test results) and artifacts (e.g., normative documents, legislation) that influence assessment activities.

Table 1 identifies an initial set of six assessment components, which reflect current practice and research in large-scale assessment according to literature reviews, normative documents, and handbooks in the field of testing and test development (e.g., American Educational Research Association et al., 1999, 2014; Linn, 1999; Brennan, 2006; Downing and Haladyna, 2006; Lane et al., 2016; International Test Commission, 2018). Consistent with the notion that tests and assessment systems are cultural products, all practices and artifacts involved should be regarded as relevant to examining cultural responsiveness, in addition to those explicitly intended to address cultural diversity in large-scale assessment programs (e.g., item bias analysis, translation, testing accommodations).

The table should not be regarded as attempting to portray a universal process, as the components may vary across social contexts or large-scale programs. Nor should the six components be considered as fixed, as they may be collapsed or broken down into other components, depending on each assessment context. Indeed, the list may be expanded as experience examining cultural responsiveness in large-scale assessment accumulates. The table is intended to show the wide variety of practices that may be enacted and the wide variety of artifacts that may be used or generated in different large-scale assessment programs. The practices listed within and across assessment process components do not necessarily take place in a linear or in the same sequence, and many of the artifacts may be generated iteratively (e.g., through a process of review and revision) or synergistically.

TABLE 1 | Assessment process components with examples of assessment practices and artifacts.

1. Relevant policy and legislation. Legislation and policy establish student populations (e.g., eight grade students) to be tested and target content areas (e.g., science); it also dictates ways in which the results of tests are to be used to (e.g., to make student promotion or retention decisions or to sanction schools or impose conditions for funding). Policy and legislation may also identify student populations of special interest, as is the case of indigenous populations or students who are not proficient in the language of testing. *(Examples of practices and artifacts: legislation on the testing of linguistic minorities; legal definitions of students from certain cultural or linguistic groups; assessment-based accountability for schools with students who are not proficient in the language in which tests are administered).*
2. Normative documents. Normative documents specify a domain of knowledge and skills for a given content and provide guidance for systematic development of tests. These documents reflect a view of the content valued and a form of organizing that content. They influence the ways in which content is sampled and the characteristics of the items that are to be generated. In some cases, these documents may discuss actions to be taken with the intent to address cultural diversity. *(Examples of practices and artifacts: content standards; sampling frameworks; assessment frameworks; item specifications; test development project solicitations).*
3. Test development and adaptation. Tests are developed according to normative documents. Typically, tests are designed and created with the mainstream population of students in mind; then they are adapted for certain cultural groups. Adaptations may include translations or modifications of the wording of items and may be informed by external reviews. Assessment products generated may include testing accommodations (modifications in the formats of tests or the ways in which they are administered) for students with limited proficiency in the language in which they are tested, with the intent to support them to gain access to the content of items. *(Examples of practices and artifacts: test design; test development; trying out tests with pilot students; test translation / adaptation; test review; test translation review; test revision).*
4. Tests and test administration. Students are given tests or translated /adapted tests. Depending on the form of administration (e.g., paper and pencil, computer-based), test administration can include different forms of accommodations or different forms of accessibility resources intended to support students who are not proficient in the language of testing to gain access to the content of items. *(Examples of practices and artifacts: tests; test administration; testing accommodations; translated / adapted tests; online or computer-administered testing platforms; accessibility resources).*
5. Data analysis. Students' responses to tests are scored and analyzed using different methods depending on the characteristics of the assessment program. Most analytical approaches used in large-scale assessment are based on the use of item response theory, as the main goal is to measure students' skills according to a scale. Analyses can be performed to detect culturally biased items. (*Examples of practices and artifacts: scoring; analysis; reporting)*.
6. Uses and consequences. The results of tests are reported and used to inform multiple decisions on students, their teachers, and their schools, or in the form of national or state educational policies. (*Examples of practices and artifacts: grading; score reporting; student placement; student exit and retention decisions; student classification and reclassification according to level of proficiency in the language of schooling; generalizations about groups; accountability*).

The table helps to appreciate the complexity of the assessment process. Notice that rarely are practices or artifacts other than tests, testing, and data analyses considered in discussions of validity. For example, with some exceptions (e.g., Wolf et al., 2010), seldom is the validity of score interpretations across multiple cultural groups referred to the characteristics of the assessment systems that generate tests or to the normative documents that establish how tests are to be created.

Establishing Procedural Assumptions

Procedural assumptions can be conceived as criteria of conceptual and methodological rigor that need to be met throughout the entire assessment process in order to validly, fairly test culturally diverse populations. Altogether, these procedural assumptions formalize the kinds of actions that need to be taken to ensure the defensibility of the interpretation and use of test scores across cultural groups.

Culturally responsive assessment has been conceived as:

> "(being) mindful of the student populations (...), using language that is appropriate for all students (...), acknowledging students' differences in the planning phases of an assessment effort, developing, and/or using assessment tools that are appropriate for different students, and being intentional in using assessment results to improve learning for all students. Culturally responsive assessment (...) calls for student involvement throughout the entire assessment process including the development of learning outcome statements, assessment tool selection/development process, data collection and interpretation, and use of results." (Montenegro and Jankowski, 2017, p. 10).

Based on this definition and on literature reviews, normative documents, and documents on fairness in testing (e.g., American Educational Research Association et al., 1999, 2014; Duran, 1999; Hambleton, 2005; Camilli, 2006; International Test Commission, 2018), 10 initial procedural assumptions can be identified as critical to cultural responsiveness (**Table 2**). Each assumption is worded generically, as a statement of the conditions that need to be met or the actions that need to be taken to effectively contribute to valid, fair testing of culturally diverse populations. Needless to say, how each assumption is met depends on the characteristics of each specific large-scale assessment program.

The procedural assumptions shown are written in generic style to facilitate examination of multiple assessment practices and artifacts of any assessment process component in any large-scale assessment program. They should not be interpreted as a finished set; more assumptions can be added as experience from examining cultural responsiveness of multiple assessment programs accumulates. For example, in international comparisons, procedural assumptions may need to be formalized to address the fact that different societies value different skills (e.g., memorization, problem solving, critical thinking) in different ways and emphasize those skills in different ways in their curricula (Gebril, 2016; Kennedy, 2016). Taking into account those differences are critical to making valid interpretations of countries' score differences. Clearly, the set of procedural assumptions needs to be partitioned, or expanded depending on the complexity and characteristics of each assessment endeavor.

Constructing a Matrix of Evidence for Validity Argumentation

Matrices of evidence have been characterized as heuristic, analytical tools for promoting critical thinking (Averill, 2002; Atkinson et al., 2007). They have been used recently in

TABLE 2 | Procedural assumptions.

1. Population specification. Cultural groups are defined according to criteria such as ethnicity, language, nationality, social class, and history, to the extent to which they reflect, emerge from, or are associated with cultural differences.
2. Conceptual and empirical defensibility. Cultural diversity is addressed in ways that are consistent with current knowledge from disciplines relevant to culture and language (e.g., cultural anthropology, sociolinguistics, semiotics, language acquisition).
3. Inclusion, representation, and sampling. Representative samples of individuals from different cultural groups and their social contexts are used.
4. Data disaggregation. Information on both the performance of students on tests and the technical properties of those instruments are examined separately for each relevant cultural group.
5. Probabilistic reasoning. Error due to uncertainty in the knowledge of the characteristics of cultural groups and to the fallibility of classifications of individuals into cultural groups is recognized and estimated.
6. Heterogeneity. Heterogeneity of individuals within the same given cultural group and heterogeneity between cultural groups is recognized and addressed.
7. Implementation. Resources and efforts are allocated to ensure that methods and procedures are applied with fidelity and consistently across individuals and socio-cultural contexts.
8. Time and timeliness. Activities intended to address cultural diversity are scheduled at points in the process of assessment in which they can influence its outcomes and are given enough time to be completed successfully.
9. Correction mechanisms. Procedures and resources are in place that allow improvements of the assessment process when new information relevant to properly addressing cultural diversity is available.
10. Intersectionality. Information is obtained about cultural groups other than test data allows examination of specific combinations of categories of factors that are particularly relevant to addressing cultural diversity issues.

criminology (e.g., Lum et al., 2011; Veigas and Lum, 2013) and clinical diagnosis (Seidel et al., 2016) as tools for examining and integrating pieces of information (often conflicting or fragmented) from multiple sources.

In the field of educational assessment, one publication reports on the use of evidence matrices to evaluate fidelity in the implementation of translation and cultural adaptation procedures in an international assessment project (Chia, 2012). Regarding validity, only three recent publications report on efforts to unpack the relationships of assumptions and evidence of interpretive arguments through the use of trees (Ruiz-Primo et al., 2012) and evidence matrices (Ruiz-Primo and Li, 2014, 2018). This small number of publications is surprising, given the fact that the importance of relating facts to evaluation criteria as critical to examining validity in testing was discussed more than six decades ago (see Cronbach and Meehl, 1955).

The matrix of evidence for validity argumentation can be conceived as a particular type of evidence matrix. It is a conceptual tool for organizing information on the extent to which the practices and artifacts involved in a large-scale assessment program address cultural diversity according to a set of procedural assumptions. Formally, it consists of a rectangular arrangement of *m* procedural assumptions (rows) and *n* assessment process components (columns), as shown in **Figure 1**. Each cell *[i, j]* contains confirming and/or disconfirming evidence that the characteristics of relevant assessment artifacts and practices from Assessment Process Component *j* meet Procedural Assumption *i*. This confirming and/or disconfirming evidence is obtained from examining relevant sources of information (e.g., documents on the process of development of the assessment, technical reports of the instrument, or reports of reviews of sample items) that are relevant to the corresponding large-scale assessment program.

Figure 2 shows a matrix of evidence for validity argumentation assembled with the set of assessment process components and the set of procedural assumptions presented in **Tables 1**, **2**.

To illustrate how the cells of this matrix are to be filled out, a series of narratives are provided below. A narrative is provided for each of six cells ([3, 1], [3, 2],..., [3, 6]) across the same assumption (Assumption 3: Inclusion, Representation, and Sampling). Altogether, these six narratives show how the same procedural assumption applies to practices and artifacts from different assessment process components.

In a specific large-scale assessment program, all the cell narratives contain information specific to that program. Yet for illustration purposes, these six narratives are from different assessment contexts. This serves the function of showing that the same given assumption is applicable to different large-scale assessment programs with different forms of cultural diversity.

[3, 1]. The term, *inclusion* is often used in the assessment literature to refer to the actions intended to ensure that students from cultural minority groups or traditionally underserved groups participate in large-scale assessment programs. Such is the case of the decree that created the National Institute for Educational Evaluation in Mexico (Secretaría de Educación Pública, 2002), which charged the institute with assessing Indian populations, along with other populations, as part of the strategies intended to evaluate the state of education in the country. While the document states that Indian populations should be included in national tests, no documents are available that provide guidance on how their characteristics (e.g., their tremendous linguistic diversity) need to be addressed if they are to be fairly included in national assessment programs (see Solano-Flores et al., 2014).

[3, 2]. Along with views of disciplinary knowledge, standards documents (e.g., New Generation of Science Standards, 2013) used to inform the process of test development in large-scale assessment pose different sets of language demands and opportunities for learning (see Lee et al., 2013). Yet culture and language issues are discussed only tangentially in many normative documents (see Raiker,

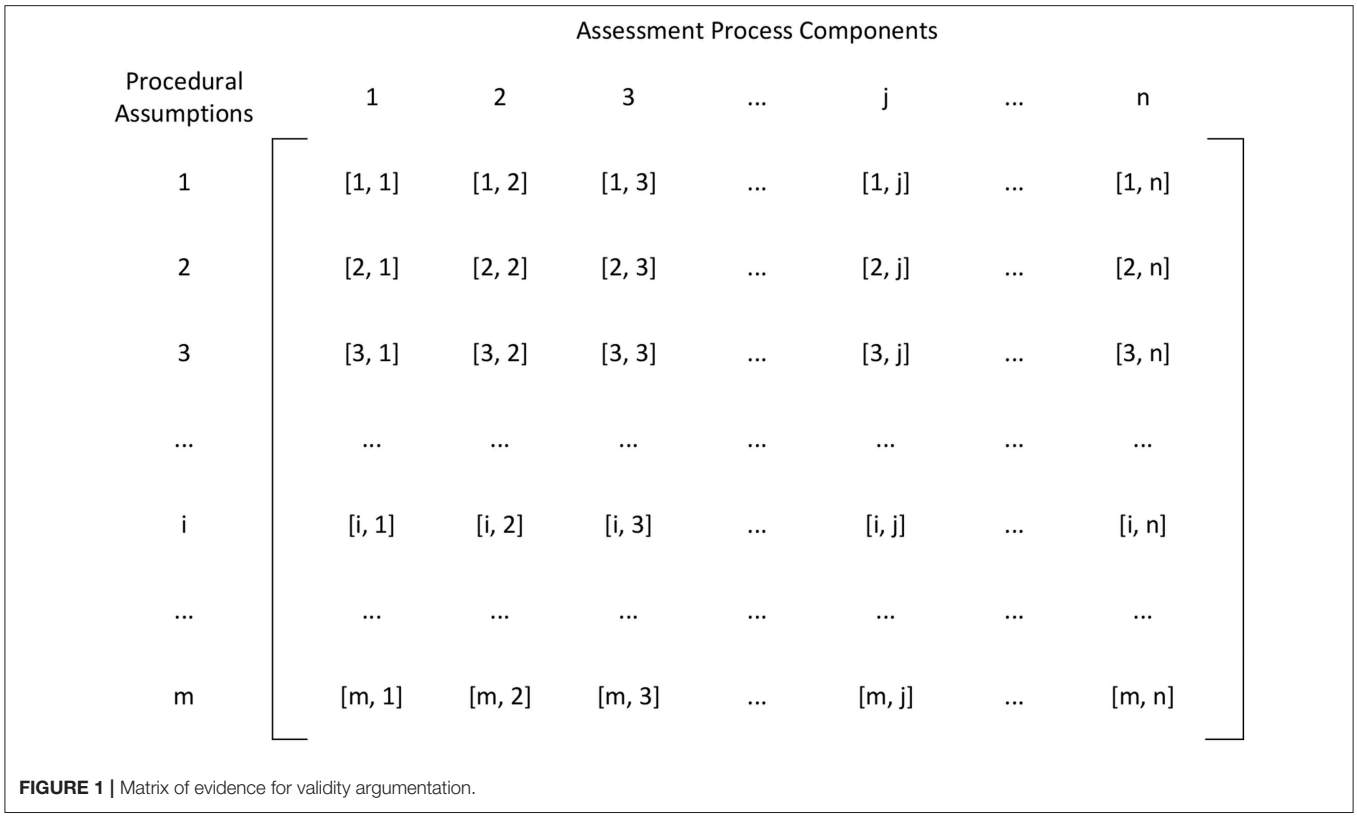

FIGURE 1 | Matrix of evidence for validity argumentation.

2002). At best, they are discussed in the last chapters of those documents, or as appendices, not as issues that cut across topics. As a consequence, scant attention is paid to the role that culture and language play in knowledge construction, the epistemological aspects of learning content, the ways in which language encodes disciplinary knowledge, or the ways in which culturally-determined world views shape content learning. As a result of such disconnect, disciplinary knowledge may be wrongly assumed to be culture-free.

[3, 3]. In many large-scale assessment programs, test developers make serious efforts to include representatives of certain cultural groups in their teams of developers, consultants, or reviewers. While commendable, these efforts may not be sufficient to properly address cultural diversity if representative samples of students from those cultural groups are not included in the process of test development (Solano-Flores, 2009). A great deal of the process of test development has to do with refining the language of items to ensure that students understand them as test developers intend. Therefore, by not including those students in the process, test developers miss the opportunity to obtain valuable information on the characteristics of the items (see National Academies of Sciences, 2018).

[3, 4]. Criticisms to international test comparisons such as PISA point at the fact that their items represent situations, contexts, epistemologies, and values from middle- and upper-class segments of societies in industrialized, Western countries (e.g., Sjøberg, 2016). This under-representation of other cultures can potentially have an adverse impact on the performance of students from cultures that are not those portrayed in the test items. Some cultural adaptations are made on the characteristics of items when they are translated into the languages used in the participating countries (see Arffman, 2013; OECD, 2017). However, these adaptations tend to be few and superficial and do not include the contextual information of items. Of course, it cannot be assumed that limited familiarity with the contexts used in tests necessarily prevents students from understanding test items. However, there is evidence that students may make sense of the contextual information of test items by relating them to socio-cultural experiences that do not always take place in the classroom (Solano-Flores and Nelson-Barber, 2001; Luyks et al., 2007; Solano-Flores and Li, 2009).

[3, 5]. The inclusion of students from different cultural groups in large-scale assessment programs is often predicated upon the availability of analytical techniques for detecting cultural bias (see van de Vijver, 2016), including those based on item response theory (Camilli, 2013). However, because item bias detection is costly and requires the use of data obtained after administering tests to large samples of students, it is difficult or unlikely for this kind of scrutiny to take place in large-scale assessment programs,

Procedural Assumptions	Assessment Process Components					
	1 Relevant Policy and Legislation	2 Normative Documents	3 Test Development and Adaptation	4 Tests and Test Administration	5 Data Analysis	6 Uses and Consequences
1. Population specification	[1, 1]	[1, 2]	[1, 3]	[1, 4]	[1, 5]	[1, 6]
2. Conceptual and empirical defensibility	[2, 1]	[2, 2]	[2, 3]	[2, 4]	[2, 5]	[2, 6]
3. Inclusion, representation, and sampling	[3, 1]	[3, 2]	[3, 3]	[3, 4]	[3, 5]	[3, 6]
4. Data disaggregation	[4, 1]	[4, 2]	[4, 3]	[4, 4]	[4, 5]	[4, 6]
5. Probabilistic reasoning	[5, 1]	[5, 2]	[5, 3]	[5, 4]	[5, 5]	[5, 6]
6. Heterogeneity	[6, 1]	[6, 2]	[6, 3]	[6, 4]	[6, 5]	[6, 6]
7. Implementation	[7, 1]	[7, 2]	[7, 3]	[7, 4]	[7, 5]	[7, 6]
8. Time and timeliness	[8, 1]	[8, 2]	[8, 3]	[8, 4]	[8, 5]	[8, 6]
9. Correction mechanisms	[9, 1]	[9, 2]	[9, 3]	[9, 4]	[9, 5]	[9, 6]
10. Intersectionality	[10, 1]	[10, 2]	[10, 3]	[10, 4]	[10, 5]	[10, 6]

FIGURE 2 | Matrix of evidence for validity argumentation with a specific set of procedural assumptions and a specific set of assessment process components.

even for samples of the items generated (Solano-Flores and Milbourn, 2016).

[3, 6]. Unfortunately, accountability, policy, or education reform decisions based on test scores rarely take into consideration the under-representation of situations, contexts, epistemologies, and values of certain cultural groups in test items. For example, in making sense of results from international test comparisons, policy makers may focus on their countries' relative rankings (see Carnoy, 2015) without taking into consideration the cultural mismatch between the characteristics of the items and the characteristics of the national culture (Solano-Flores, 2019).

It is important to mention that, in examining cultural responsiveness for a specific large-scale assessment program, the cell narratives in the final version of the matrix of evidence for validity argumentation are likely to be longer and more elaborated than the narratives shown. Also, each cell may contain several pieces of evidence. Moreover, the same given cell may contain conflicting (i.e., confirming and disconfirming) pieces of evidence. Furthermore, for some large-scale assessment programs, some cells may be in blank because relevant information is unavailable. While, strictly speaking, absence of evidence is not evidence of absence, many blank cells in the matrix of evidence for validity argumentation may be an indication of poor implementation and poor documentation of the process of assessment.

THE OUTCOME: BUILDING A VALIDITY ARGUMENT

Consistent with the notion of validity as argumentation (Cronbach, 1988; Kane, 2006), the matrix of evidence for validity argumentation allows building a logical evaluative argument on the interpretations and use of test scores for a culturally diverse population of students. More specifically, a validity argument for a given set of interpretations and uses of test scores is built by integrating the information contained in the cells of the matrix. This validity argument can be organized according to the procedural assumptions (by row, across columns), or according to the assessment process components (by column, across rows).

The narrative below illustrates how a validity argument can be built using the matrix of evidence for validity argumentation. It shows how the information from different cells on the same column can be integrated into a coherent narrative for validity argumentation. It discusses one particular set of practices and artifacts (testing accommodations) from the same assessment

process component (tests and testing) in a specific assessment program (the National Assessment of Educational Progress in the U.S.) and for a specific cultural group (English language learners or ELs—students in the U.S. who are not proficient in English but who are tested in English in many large-scale assessment programs in the U.S.) across procedural assumptions. Callouts are inserted in the text at the end of sentences or paragraphs to indicate the specific cells in the matrix (**Figure 2**) that are relevant to the issues being discussed.

In the U.S., many large-scale assessment programs authorize the use of accommodations in the testing of EL students. These accommodations are defined as modifications in the characteristics of the tests or the ways in which they are administered with the intent to support these students to gain access to the content of the items without altering the constructs targeted and without giving these students any unfair advantage over students who are not provided with accommodations (Abedi and Ewers, 2013). For example, some of accommodations authorized by the National Assessment of Educational Progress include: extending the time for students to take the test, allowing them to take the test in small groups, reading for them the test items aloud in English, and allowing them to take breaks during test (Institute of Education Sciences et al., 2018b) [1, 4].

Testing accommodations deviate from the traditional view of standardization as a basis for fair assessments and their use is predicated upon the notion that "surface conditions that differ in principled ways for different learners can provide equivalent evidence" (Mislevy et al., 2013, p. 122). While they are intended to reduce measurement error due to limited proficiency in the language of the test, many of these accommodations have not been sufficiently investigated (see Sireci et al., 2008), do not have sufficient theoretical or empirical support (Wolf et al., 2012), or have been borrowed from the field of special education (see Rivera and Collum, 2006) [2, 4].

Few accommodations (e.g., simplifying the wording of items or providing extra time for completing the test) that can potentially be more effective in supporting ELs have been investigated. However, available empirical evidence on their effectiveness is mixed (e.g., Kieffer et al., 2009) [2, 4]. One possible reason for this inconsistency is that, because of the considerable heterogeneity of EL populations (see IInstitute of Education Sciences et al., 2018a), the samples of EL students used in these investigations may have varied on important factors such as first language, schooling history, SES background, and even level of English proficiency [6, 4]. Another possible reason is that each form of accommodation may be interpreted and implemented in different ways by educators in different contexts (Solano-Flores, 2016) [7, 4].

While there is more consistent evidence on the effectiveness of some accommodations (Pennock-Roman and Rivera, 2011), this effectiveness is somewhat limited by the fact that they are provided as blanket supports (all students classified as ELs receive the same kind of accommodation) and they are not sensitive to the fact that each EL student has a unique set of needs and strengths in English [6, 4].

Difficulties in establishing a principled, effective practice concerning the use of accommodations stem from error and the level of granularity of data in the classifications of EL students. While NAEP reports test results by English proficiency status, it uses only two categories of English proficiency: LEP (Limited English) and Not LEP (e.g., Institute of Education Sciences et al., 2018c). The use of this limited number of categories probably stems from the fact that states are inconsistent in their definitions and classifications of students as *English learners* at different levels of English proficiency (Linquanti et al., 2016) [5, 4]. In addition, NAEP does not report results on English proficiency by student first language or form of testing accommodation provided (e.g., Institute of Education Sciences et al., 2018c) [10, 4]. Owing to this limitation, it is difficult to judge how the effectiveness of accommodations is shaped by students' characteristics.

Notice that the narrative above discusses only one of the six assessment process components considered and focuses only on a specific set of practices and artifacts (testing accommodations) within that component. This speaks to both the analytical capabilities of the matrix of evidence for validity argumentation and the richness of the information that can be used to examine cultural responsiveness in large-scale assessment.

DISCUSSION

In a global economy, and as societies become more diverse, cultural responsiveness in assessment plays an increasingly critical role in ensuring proper use of tests and proper test score interpretation. Yet in many large-scale assessment programs, practical limitations may limit the level of attention given to culture as critical to valid, fair testing of culturally diverse populations. For example, when it is conducted, item bias analysis may take place late in the process of assessment, when it is difficult or impractical to eliminate biased items (see Allalouf, 2003). In many cases, culturally biased items can be detected only after tests have been administered and decisions have been made based on the scores they produce (e.g., Yildirim and Berberoğlu, 2009).

In this paper, I have addressed the challenges of attaining cultural responsiveness in large-scale assessment from a systemic perspective that involves the practices enacted and artifacts used or generated throughout the process of assessment. Consistent with the perspective of validity as argumentation, a matrix of evidence for validity argumentation can be used as a tool for systematically gathering, contrasting, and integrating pieces of evidence that confirm or disconfirm the validity of interpretations and use of test scores for a culturally diverse population in a large-scale assessment program.

Narratives were provided to illustrate the kind of information contained in the cells in the matrix and to illustrate how the information provided in several cells can be integrated to build a validity argument.

The intersection of assessment process components and procedural assumptions speaks to the ubiquity of cultural issues

in the entire process of assessment. Effectively attaining cultural responsiveness in state, national, or international large-scale assessment entails a societal endeavor. It requires that all actors involved (funding agencies, decision makers, test developers, contractors, researchers, governments, and test users) recognize and act upon the fact that, simply because the practices enacted and the artifacts used or generated in the process of assessment are cultural products, their characteristics are relevant to culture. Attaining cultural responsiveness in large-scale assessment is a serious but not impossible endeavor, especially because all the concepts and methods needed to properly address the procedural assumptions in the matrix (e.g., data disaggregation, statistical representation, sampling, population specification, heterogeneity, etc.) and all the practical/logistical aspects of test development (e.g., timeliness, correction mechanisms) are well-known in the field of educational measurement. The matrix of evidence for validity argumentation contributes to ensuring that these concepts and methods are used consistently and systematically.

DATA AVAILABILITY

This paper is based on data published in other studies. All relevant data for this study are included in the paper.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

ACKNOWLEDGMENTS

I am grateful to Dr. Maria A. Ruiz-Primo for her comments on an earlier version of this paper.

REFERENCES

Abedi, J., and Ewers, N. (2013). *Accommodations for English Language Learners and Students With Disabilities: A Research-Based Decision Algorithm*. Smarter Balanced Assessment Consortium.

Allalouf, A. (2003). Revising translated differential functioning items as a tool for improving cross-lingual assessment. *Appl. Meas. Educ.* 16, 55–73. doi: 10.1207/S15324818AME1601_3

American Educational Research Association, American Psychological Association, and National Council for Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association American Psychological Association and National Council for Measurement in Education.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Arffman, I. (2013). Problems and issues in translating international educational achievement tests. *Educ. Meas. Issues Pract.* 32, 2–14. doi: 10.1111/emip.12007

Aronson, B., and Laughter, J. (2016). The theory and practice of culturally relevant education: a synthesis of research across content areas. *Rev. Educ. Res.* 86, 163–206. doi: 10.3102/0034654315582066

Atkinson, M., Wills, J. B., and McClure, A. I. (2007). The evidence matrix: a simple heuristic for analyzing and integrating evidence. *Teach. Sociol.* 35, 262–271. doi: 10.1177/0092055X0803600306

Averill, J. B. (2002). Matrix analysis as a complementary analytic strategy in qualitative inquiry. *Qual. Health Res.* 12, 855–866. doi: 10.1177/10432302012006011

Brennan, R. L. (ed.). (2006). *Educational Measurement, 4th Edn*. Westport, CT: American Council on Education and Praeger Publishers.

Camilli, G. (2006). "Test fairness," in *Educational Measurement, 4th Edn,* ed R. L. Brennan (Westport, CT: American Council on Education and Praeger Publishers), 221–256.

Camilli, G. (2013). Ongoing issues in test fairness. *Educ. Res. Eval.* 19, 104–120. doi: 10.1080/13803611.2013.767602

Carnoy, M. (2015). *International Test Score Comparisons and Educational Policy: A Review of the Critiques.* National Education Policy Center.

Chia, M. (2012). *Fidelity of Implementing an Assessment Translation and Adaptation Framework in a Study of an Emerging International Assessment* (Unpublished doctoral dissertation). School of Education, University of Colorado at Boulder, Boulder, CO.

Cole, M. (1999). "Culture-free versus culture-based measures of cognition," in *The Nature of Cognition,* ed R. J. Sternberg (Cambridge, MA: The MIT Press), 645–664.

Cronbach, L. J. (1988). "Five perspectives on validity argument," in *Test Validity,* eds H. Wainer and H. I. Braun (Hillsdale, NJ: Erlbaum), 3–17.

Cronbach, L. J., and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychol. Bull.* 52, 281–302.

Downing, S. M., and Haladyna, T. M. (eds.). (2006). *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbaum.

Duran, R. P. (1999). "Testing of linguistic minorities," in *Educational Measurement, 3rd Edn.* ed R. L. Linn (New York, NY: American Council on Education and Macmillan Publishing Company), 573–587.

Gay, G. (2000). *Culturally Responsive Teaching: Theory, Practice, and Research*. New York, NY: Teachers College Press.

Gebril, A. (2016). "Educational assessment in Muslim countries: values, policies, and practices," in *Handbook of Human and Social Conditions of Assessment,* eds G. T. L. Brown and L. Harris (New York, NY: Routledge), 420–435.

Geisinger, K. F. (2010). "Test standardization," in *Corsini Encyclopedia of Psychology, 4th Edn*. Vol. IV, eds I. Weiner and W. E. Craighead (New York, NY: Wiley), 1769–1770.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: in search of the evidence. *Educ. Meas. Issues Pract.* 18, 5–9.

Hambleton, R. K. (2005). "Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures," in *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*, eds R. K. Hambleton, P. F. Merenda, and C. D. Spielberger (Mahwah, NJ: Lawrence Erlbaum Associates, Publishers), 3–38.

Institute of Education Sciences, National Center of Education Statistics, National Assessment of Educational Progress. (2018a). *English Language Learners in Public Schools*.Available online at: https://nces.ed.gov/programs/coe/indicator_cgf.asp (accessed April 20, 2019).

Institute of Education Sciences, National Center of Education Statistics, National Assessment of Educational Progress. (2018b). *Inclusion of Students With Disabilities and English Language Learners*. Available online at: https://nces.ed.gov/nationsreportcard/about/inclusion.aspx (accessed April 20, 2019).

Institute of Education Sciences, National Center of Education Statistics, National Assessment of Educational Progress. (2018c). *NAEP Mathematics Report Card*. Available online at: https://www.nationsreportcard.gov/math_2017?grade=4 (accessed April 20, 2019).

International Test Commission (2018). *ITC Guidelines for the Large-Scale Assessment of Linguistically and Culturally Diverse Populations*. Available online at: www.InTestCom.org (accessed April 20, 2019).

Kane, M. T. (1992). An argument-based approach to validity. *Psychol. Bull.* 112, 527–535.

Kane, M. T. (2006). "Validation," in *Educational Measurement, 4th Edn*, ed R. L. Brennan (Westport, CT: American Council on Education and Praeger Publishers), 17–64.

Kennedy, K. J. (2016). "Exploring the influence of culture on assessment: the case of teachers' conceptions of assessment in Confucian heritage cultures," in *Handbook of Human and Social Conditions of Assessment*, eds G. T. L. Brown and L. Harris (New York, NY: Routledge), 404–419.

Kieffer, M. J., Lesaux, N. K., Rivera, M., and Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: a meta-analysis on effectiveness and validity. *Rev. Educ. Res.* 79, 1168–1201. doi: 10.3102/0034654309332490

Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *Am. Educ. Res. J.* 32, 465–491.

Lane, S., Raymond, M. R., and Haladyna, T. M. (eds.) (2016). *Handbook of Test Development, 2nd Edn*. New York, NY: Routledge.

Lee, O., Quinn, H., and Valdés, G. (2013). Science and language for English language learners in relation to Next Generation Science Standards and with implications for Common Core State Standards for English Language Arts and Mathematics. *Educ. Res.* 42, 223–233. doi: 10.3102/0013189X13480524

Linn, R. L. (eds.) (1999). *Educational Measurement, 3rd Edn*. New York, NY: American Council on Education and Macmillan Publishing Company.

Linquanti, R., Cook, H. G., Bailey, A. L., and MacDonald, R. (2016). *Moving Toward a More Common Definition of English Learner: Collected Guidance for States and Multi-State Assessment Consortia*. Washington, DC: Council of Chief State School Officers.

Lum, C., Koper, C. S., and Telep, C. W. (2011). The evidence-based policing matrix. *J. Exp. Ciminol.* 7, 3–26. doi: 10.1007/s11292-010-9108-2

Luyks, A., Lee, O., Mahotiere, M., Lester, B., Hart, J., and Deaktor, R. (2007). Cultural and home language influences on children's responses to science assessments. *Teach. Coll. Rec.* 109, 897–926.

Messick, S. (1989). "Validity," in *Educational Measurement, 3rd Edn,* ed R. L. Linn (New York, NY: American Council on Education/Macmillan), 13–103.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika* 59, 439–483.

Mislevy, R. J., Haertel, G, Cheng, B. H., Ructtinger, L., DeBarger, A., and Murray, E. (2013). A "conditional" sense of fairness in assessment. *Educ. Res. Eval.* 19, 121–140. doi: 10.1080/13803611.2013.767614

Montenegro, E., and Jankowski, N. A. (2017). *Equity and Assessment: Moving Towards Culturally Responsive Assessment* (Occasional Paper No. 29). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA).

National Academies of Sciences, Engineering, and Medicine (2018). *English Learners in STEM Subjects: Transforming Classrooms, Schools, and Lives*. Washington, DC: The National Academies Press.

New Generation of Science Standards (2013). *Appendix D: All Standards, All Students: Making the Next Generation Science Standards Accessible to All Students*. New Generation of Science Standards.

OECD (2017). *PISA 2015 Technical Report*. OECD.

Pennock-Roman, M., and Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: a meta-analysis of experimental studies. *Educ. Measur. Issues Pract.* 30, 10–28. doi: 10.1111/j.1745-3992.2011.00207.x

Raiker, A. (2002). Spoken language and mathematics. *Camb. J. Educ.* 32, 45–60. doi: 10.1080/03057640220116427

Rivera, C., and Collum, E. (eds.) (2006). *State Assessment Policy and Practice for English Language Learners: A National Perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ruiz-Primo, M.A., and Li, M. (2015). The relationship between item context characteristics and student performance: the case of the 2006 and 2009 PISA science items. *Teach. Coll. Rec.* 117:010306.

Ruiz-Primo, M. A., and Li, M. (2014). *Building a Methodology for Developing and Evaluating Instructionally Sensitive Assessments*. NSF Award ID: DRL-0816123. Final Report to the National Science Foundation.

Ruiz-Primo, M. A., and Li, M. (2018). "Evaluating the validity claims of instructionally sensitive assessments," in *Paper Presented in the Symposium: Validity of Educational Assessments: Capturing (Instructional Effects on) Students' Learning Growth AERA Annual Conference* (New York, NY).

Ruiz-Primo, M. A., Li, M., Wills, K., Giamellaro, M., Lan, M-C., Mason, H., et al. (2012). Developing and evaluating instructionally sensitive assessments in science. *J. Res. Sci. Teach.* 49, 691–712. doi: 10.1002/tea.21030

Secretaría de Educación Pública (2002). *Decreto por el Que se Crea el Instituto Nacional para la Evaluación de la Educación*. Diario Oficial, jueves *8 de agosto de* 2002.

Seidel, D., Frank, R. D., and Schmidt, S. (2016). The evidence value matrix for diagnostic imaging. *J. Am. Coll. Radiol.* 13, 1253–1259. doi: 10.1016/j.jacr.2016.05.013

Sireci, S. G., Han, K. T., and Wells, C. S. (2008). Methods for evaluating the validity of test scores for English language learners. *Educ. Assess.* 13, 108–131. doi: 10.1080/10627190802394255

Sjøberg, S. (2016). "OECD, PISA, and globalization: the influence of the international assessment regime," in *Education Policy Perils: Tackling the Tough Issues*, eds C. H. Tienken, and C. A. Mullen (New York, NY: Routledge), 102–133.

Solano-Flores, G. (2009). "The testing of English language learners as a stochastic process: population misspecification, measurement error, and overgeneralization," in *Generalizing from Educational Research,* eds K. Ercikan and W. M. Roth (New York, NY: Routledge), 33–48.

Solano-Flores, G. (2011). "Assessing the cultural validity of assessment practices: an introduction,"in *Cultural Validity in Assessment: Addressing Linguistic and Cultural Diversity,* eds M. R. Basterra, E. Trumbull, and G. Solano-Flores, (New York, NY: Routledge), 3–21.

Solano-Flores, G. (2016). *Assessing English Language Learners: Theory and Practice*. New York, NY: Routledge.

Solano-Flores, G. (2019). "The participation of Latin American countries in international assessments: assessment capacity, validity, and fairness," in *Sage Handbook on Comparative Studies in Education: Practices and Experiences in Student Schooling and Learning,* eds L. E. Suter, E. Smith, and B. D. Denman (Thousand Oaks, CA: Sage), 139–161.

Solano-Flores, G., Backhoff, E., Contreras-Niño, L. A., and Vázquez-Muñoz, M. (2014). Language shift and the inclusion of indigenous populations in large-scale assessment programs. *Int. J. Test.* 15, 136–152. doi: 10.1080/15305058.2014.947649

Solano-Flores, G., and Li, M. (2009). Generalizability of cognitive interview-based measures across cultural groups. *Educ. Meas. Issues Pract.* 28, 9–18. doi: 10.1111/j.1745-3992.2009.00143.x

Solano-Flores, G., and Milbourn, T. (2016). Assessment capacity, cultural validity, and consequential validity in PISA. *Relieve* 22:M12. doi: 10.7203/relieve.22.1.8281

Solano-Flores, G., and Nelson-Barber, S. (2001). On the cultural validity of science assessments. *J. Res. Sci. Teach.* 38, 553–573. doi: 10.1002/tea.1018

van de Vijver, F. J. R. (2016). "Assessment in education in multicultural populations," in *Handbook of Human and Social Conditions of Assessment,* eds G. T. L. Brown and L. Harris (New York, NY: Routledge), 436–453.

van de Vijver, F. J. R., and Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: an overview. *Rev. Eur. Psychol. Appl.* 54, 119–135. doi: 10.1016/j.erap.2003.12.004

Veigas, H., and Lum, C. (2013). Assessing the evidence base of a police service patrol portfolio. *Policing* 7, 248–262. doi: 10.1093/police/pat019

Wolf, M. K., Herman, J. L., and Dietel, R. (2010), Spring. *Improving the Validity of English Language Learner Assessment Systems* (CRESST Policy Brief No. 10 - Executive Summary). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Wolf, M. K., Kao, J. C., Rivera, N. M., and Chang, S. M. (2012). Accommodation practices for English language learners in states' mathematics assessments. *Teach. Coll. Rec.* 114, 1–26.

Yildirim, H. H., and Berberoğlu, G. (2009). Judgmental and statistical DIF analyses of the PISA-2003 mathematics literacy items. *Int. J. Test.* 9, 108–121. doi: 10.1080/15305050902880736

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.