frontiers in Education

REVIEW
published: 14 April 2021
doi: 10.3389/feduc.2021.637993

The Semiotics of Test Design: Conceptual Framework on Optimal Item Features in Educational Assessment Across Cultural Groups, Countries, and Languages

*Guillermo Solano-Flores**

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

This paper offers a conceptual framework on test design from the perspective of social semiotics. Items are defined as arrangements of features intended to represent information, convey meaning, and capture information on the examinees' knowledge or skills on a given content. The conceptual framework offers a typology of semiotic resources used to create items and discusses item representational complexity—the multiple ways in which the semiotic resources of an item are related to each other—and item semiotic alignment—the extent to which examinees share cultural experience encoded by items. Since the ability to make sense of items is shaped by the examinees' level of familiarity with the social conventions underlying the ways in which information is represented, unnecessary representational complexity and limited semiotic alignment may increase extraneous item cognitive load and adversely impact the performance of examinees from certain populations. Semiotic test design allows specification of optimal pools of semiotic resources to be used in creating items with the intent to minimize representational complexity and maximize semiotic alignment for the maximum number of individuals in diverse populations of examinees. These pools of semiotic resources need to be specific to the content assessed, the characteristics of the populations of examinees, the languages involved, etc., and determined based on information produced by cross-cultural frequency analyses, cognitive interviews, focus groups, and expert panels.

Keywords: test design, semiotics, item features, semiotic resources, cultural groups

OPEN ACCESS

Edited by:
Olga Zlatkin-Troitschanskaia, Johannes Gutenberg University Mainz, Germany

Reviewed by:
Sara Magdalena Lenninger, Kristianstad University, Sweden
Alin Olteanu, RWTH Aachen University, Germany

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

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Education

Received: *04 December 2020*
Accepted: *10 March 2021*
Published: *14 April 2021*

Citation:
Solano-Flores G (2021) The Semiotics of Test Design: Conceptual Framework on Optimal Item Features in Educational Assessment Across Cultural Groups, Countries, and Languages. Front. Educ. 6:637993. doi: 10.3389/feduc.2021.637993

INTRODUCTION

Current views of assessment as evidentiary reasoning emphasize the importance of systematic approaches for determining the numbers, formats, and features of items or tasks that are to be used in assessing a given domain of knowledge (Martinez, 1999; Pellegrino et al., 2001; National Research Council, 2006; Mislevy and Haertel, 2007). In large-scale assessment, these views support the process of test development (National Research Council, 2014) and the development of item

specifications documents that prescribe the general characteristics of items to be included in a given test (e.g., Council of Chief State School Officers, 2015).

Unfortunately, given their scope and level of analysis, such documents cannot pay detailed attention to the multiple textual and non-textual features of items. At present, no methodology is available that allows systematic selection, development, and use of the hundreds of features used in items, such as graphs, lines, arrows, labels, font styles, speech balloons, abbreviations, graph axes, ways of asking questions, ways of arranging options in multiple-choice items, buttons to click, cascade menus, and boxes to type or write answers to questions. While these features may or may not be directly related to the target knowledge domain, all of them contribute to representing information and may influence examinees' understanding of items. Many of these features may be used inconsistently across items within the same assessment program and, to a large extent, their use may be shaped more by idiosyncratic factors or tradition than by principled practice.

Concerns about this lack of a principled practice are even more serious for assessment programs that test culturally and linguistically diverse populations. For example, efforts oriented to minimizing cultural bias and ensuring the comparability of measures of tests across cultural and linguistic groups focus almost exclusively on the text of tests (e.g., Hambleton, 2005; Downing and Haladyna, 2006; International Test Commission, 2017). Little is known about whether and how the non-textual features of items should be adapted for students from different countries or cultural backgrounds. Yet we know that individuals from different cultural backgrounds may differ on the level of attention they pay to focal objects or contextual and background information (Nisbett, 2003; Chua et al., 2005); that the relative frequency of some features of item illustrations vary substantially across different assessment programs (Wang, 2012); and that the extent to which item illustrations influence student performance on science items in international comparisons varies across high- and low-ranking countries (Solano-Flores and Wang, 2015). Among many other, these findings speak to the need for a perspective of test design that allows systematic, detailed selection, and examination of the features of items.

This paper offers a conceptual framework on semiotic design focused on the testing of diverse populations across cultural groups, countries, and languages. It contributes to closing an important gap in the intersection of testing and semiotics: while education has captured the attention of semioticians for decades (e.g., Lemke, 1990; Stables, 2016; Pesce, 2018), the focus has been mainly on learning, text, and the classroom; little attention has been paid to tests and testing. The goal is not to offer a semiotic theory of testing, but rather a reasoning on the ways in which key concepts from the field of semiotics can be used to systematically analyze and design the features of test items in ways intended to minimize error variance and promote fair test development practices.

The first section provides some basic concepts from the field of social semiotics—the study of the ways in which information is represented and meaning is made according to implicit and explicit social conventions (van Leeuwen, 2004). A perspective on semiotic resources as socially made tools for conveying meaning (Kress, 2010) provides the conceptual foundation for reasoning about meaning making as cultural practice and the ways in which the features of items can be selected or created systematically. The second section offers a classification of semiotic resources used in tests and discusses their use in the testing of culturally and linguistically diverse populations. The third section offers some ideas for semiotic test design based on the notion of representational complexity— the multiple ways in which the semiotic resources of an item are related to each other—and semiotic alignment—the intersection of the cultural experience encoded by semiotic resources and the examinees' cultural experience.

SEMIOTICS, TESTS, AND DIVERSE POPULATIONS

Features, Semiotic Resources, and Multimodality

At the core of this conceptual framework is the concept of semiotic resource. van Leeuwen (2004) defines semiotic resources as

> "the actions, materials and artifacts we use for communicative purposes, whether produced physiologically—for example, with our vocal apparatus, the muscles we use to make facial expressions and gestures—or technologically—for example, with pen and ink, or computer hardware and software—together with the ways in which these resources can be organized."
>
> "Semiotic resources have a meaning potential, based on their past uses, and a set of affordances based on their possible uses, and these will be actualized in concrete social contexts where their use is subject to some form of semiotic regime" (van Leeuwen, 2004, p. 285)."

This definition allows appreciation of the vastness of actions, materials, and artifacts that have the potential to communicate meaning. For example, in certain cultural contexts, the letter *A* can be a letter used in combination with other letters to create words, an option in a multiple-choice item, a grammatical article, a marker of the beginning of a sentence, a referent of hierarchy or priority, a letter denoting a variable, etc.

The definition also allows appreciation of the critical role that history plays in encoding meaning. Semiotic resources have been characterized as means for meaning making. But because they encode cultural experience, their affordances are not constant across social and cultural contexts (Kress, 2010). The ability of individuals to make meaning of semiotic resources depends on the extent to which they share that encoded cultural experience.

According to this reasoning, a test item can be viewed as an arrangement of multiple semiotic resources used in combination with the intent to represent information, convey meaning, and capture information on the examinee's knowledge or skills on a given knowledge domain. Proper interpretation of items

greatly depends on the individual's familiarity with the social conventions underlying the features of items and, therefore, their ability to make meaning of them. Those social conventions may be explicit or implicit, formally taught at school or acquired through informal experience, relevant or irrelevant to the content assessed, or specific or external to tests and testing.

Given their interrelatedness, no semiotic resource can be assumed to be intrinsically trivial. For example, a decimal point and a decimal comma are not intended to play a critical role in assessing computation skills respectively in the items *3.1416* $\times r^2 =$ _____ and *3,1416* $\times r^2 =$ _____, which are intended to assess exactly the same kind of skill. Yet, since the use of decimal separators varies across countries (Baecker, 2010), in an international test comparison, not using the proper decimal separator in each country could constitute a source of measurement error.

The terms *item semiotic resource* and *item feature* are used as interchangeable in this paper. However, the former is used to emphasize purposeful design (e.g., *a team of test developers identifies the set of* ***semiotic resources*** *to be used in an international test*). In contrast, the latter is used more generically to refer to the characteristics of items, regardless of whether they are a result of a systematic process of design (e.g., *a researcher develops a system for coding the* ***features*** *identified in existing items from different countries*).

For the purposes of this conceptual framework, the term, *semiotic modes* is used to refer to broad categories of ways of representing information integrally (e.g., ***textual*** *and* ***visual*** *modes*) and the term, *multimodality* is used to refer to the use of semiotic resources belonging to different modalities (Kress and van Leeuwen, 2006). It is important to bear in mind that semiotic modalities should not be understood as clearly, fixed, and stable categories, but rather as interacting categories with fuzzy boundaries. For example, text contains visual features such as margins, font sizes, bold letters, etc., which contribute to conveying meaning. Also, a map has limited value as a visual device in the absence of labels and legends.

Culture, Cultural Groups, and Cultural Experience

Broadly, *culture*, as a phenomenon, is understood here as the set of practices, views, values, attitudes, communication and socialization styles, ways of knowing, and ways of doing things among the members of a community, and which are the result of shared experience and history and learned through either formal and informal experiences or acquired through multiple forms of social participation and interaction with other individuals. The definition of culture as "the *non-hereditary memory of the community*, a memory expressing itself in a system of constraints and prescriptions," (Lotman et al., 1978, p. 213, italics in the original) provides a perspective that is sensitive to the process of testing as a communication process (Solano-Flores, 2008). This definition is also consistent with the view, that, since it is the medium in which humans live and develop, culture "should be defined in terms of the artifacts that mediate human activity" (Packer and Cole, 2020, p. 11).

The term *cultural experience* or *cultural background* is used to refer to the set of experiences that an individual has from their contact with a given cultural context or with several cultural contexts. This set of experiences is assumed to be unique to each examinee, although multiple individuals can be regarded as a cultural group when they share many cultural experiences.

Items as Samples of Encoded Cultural Experience

Current thinking in the field of educational measurement views the items of a test as samples of observations from a knowledge domain (Kane, 1982). According to this view, writing an item is equivalent to drawing a sample from that knowledge domain. Items are drawn (generated) systematically according to dimensions such as topic, type of knowledge, and disciplinary practice, etc. (Lane et al., 2016).

Unfortunately, item features do not receive the same level of attention in test development as these dimensions do. For example, while item specifications documents of assessment programs may provide detailed prescriptions regarding the alignment of the items to a set of standards, scant consideration is given to features beyond item format (e.g., multiple-choice or constructed-response) or text length. Such neglect dismisses the multimodal nature of disciplinary knowledge—the fact that disciplines develop elaborate ways of representing information in multiple textual and non-textual forms used in combination (see Lemke, 1998).

A wealth of evidence speaks to the influence of different item features on the examinees' performance on tests. For example, we know that the performance of students is instable across item formats (Ruiz-Primo et al., 1993); that construct equivalence may vary depending on the ways in which items are designed (Rodriguez, 2003); and that even small changes in wording may cause translated items to function differentially (Ercikan et al., 2014).

Semiotic resources effectively convey meaning to the extent that they encode cultural experience shared by the examinees. Items indeed can be viewed not only samples of a knowledge domain, but also as samples of encoded cultural experience. These samples may be biased if they predominantly reflect the cultural experience of specific segments of a society or the specific population of students for which tests are originally developed.

The amount of effort needed to minimize such bias should not be taken lightly, as the following example illustrates:

> An assessment program intends to create a list of names of fictitious characters to be used in the contexts of its mathematics word problem items (e.g., ***Joe*** *and* ***Clara*** *need to cut a pizza into seven slices of the same size. What measure should they use to make sure that the slices have the same size?*). The intent is to have a restricted list of names that are recognizable by students with different cultural backgrounds. Using only the names included in that list should contribute to minimizing reading demands and creating equally meaningful contexts for students with different cultural backgrounds. While assembling a list of names is a simple project in principle, to serve its

intended purpose, the list should meet multiple criteria. For example: (1) female and male names should be equally represented; (2) all names should be easy to spell and read; (3) no name should have an unintended meaning in a different language; (4) all names should be familiar to many cultural groups; (5) no name should be associated to cultural stereotypes; (6) no name should be longer than ten characters; etc. Given this level of specificity, serious systematic work needs to be done to assemble a list of names that fit these rules. This work should include, among other things performing searches and asking individuals from the target populations of examinees about the suitability of the names.

Thus, even seemingly simple item features may need to be carefully designed if cultural bias is to be effectively minimized. Unfortunately, the impact on student performance of item features is yet to be investigated with this level of detail and, with some exceptions (e.g., Solano-Flores et al., 2014a), assessment programs have not paid attention to their systematic design.

TYPES OF ITEM SEMIOTIC RESOURCES

This conceptual framework classifies item semiotic resources into six types, summarized in **Table 1**. The classification is not necessarily exhaustive. Also, the six categories and types of semiotic resources discussed should not be regarded as mutually exclusive. For the sake of simplicity, the examples provided can be viewed as basic semiotic resources—those that, in the context of design, act as building blocks of more complex semiotic resources.

Consistent with the notion that disciplinary knowledge is represented, communicated, and interpreted using multiple semiotic modes (Lemke, 1998), the categories discussed should be considered as being interconnected.

Language Resources

For the purposes of this paper, language is understood as a *system* of socially established conventions for conveying meaning orally, in signed language, or in written/printed form (Halliday, 1978) and language resources are defined as specific aspects of language used as semiotic resources in items. The category of language resources is vast, as it comprises resources as small and simple as a punctuation sign or a letter and as vast and complex as the language or the multiple language modes (oral, aural, textual) in which a test administered.

Because language is the vehicle through which testing takes place, examinees' limited proficiency in the language in which tests are administered or limited familiarity with the ways in which language is used constitutes a major threat to the validity of interpretations of test scores (American Educational Research Association [AERA] et al., 2014; Sireci and Faulkner-Bond, 2015). Even minimal aspects of language use may constitute important influences that shape examinees' interpretations of test items. For example, there is evidence that subtle variations on the ways in which items are worded can make a difference in the ways in which students interpret items (Ercikan, 2002). Also, the misalignment between the textual features of items in an international test and the textual features of items in national examinations (Anagnostopoulou et al., 2013) has been documented. Potentially, such misalignment could unfairly increase the difficulty of items in international tests.

Examples at three levels of complexity illustrate the wide range of language resources and their design implications. At a very basic level, text size illustrates how features of printed language may appear deceptively trivial. Because languages differ on word length and grammatical complexity (Coupé et al., 2019), the text size of items may vary considerably across different language versions of the same test. If text size ratios are not considered at a planning stage in the development of a test, the display of the items may look crowded for some of its language versions.

At another level of complexity, the ways in which vocabulary is addressed in testing illustrates the gap between what is known about language and how that knowledge is incorporated in testing practices. While there are sources that document the frequency of words in English (e.g., Nagy and Anderson, 1984; Davies and Gardner, 2010; Nation, 2014), that information is not used routinely to decide the wording and minimize the lexical complexity in items not intended to assess vocabulary knowledge.

At a higher level of complexity, issues in test translation illustrate the challenges of testing diverse populations in different languages, mainly because translation may alter the nature of the constructs assessed by items (Hambleton, 2005; Winter et al., 2006; Arffman, 2013). A great deal of the effort and time invested in the process of assessment development concerns refining the wording of items to ensure that examinees understand them as their developers intend (Abedi, 2006, 2016). Yet, compared to the time allocated for test development, assessment programs allocate considerably less time for test translation and adaptation (Solano-Flores, 2012). Tight timelines seriously limit the opportunities for examining students' interpretations of translated items (e.g., through verbal protocols and cognitive interviews) and conducting differential item functioning analyses with the purpose of detecting cultural bias. These practical constraints underscore the need for improved judgmental

TABLE 1 | Types of semiotic resources used in test items.

Type	Main property	Examples
(1) Language resources	Systemic	Vocabulary, grammar, syntactical structures, discourse, idiomatic expressions, quotation marks, formal language, sign language
(2) Images	Mimetic	Photographs, illustrations, drawings
(3) Metaphorical devices	Diegetic	Light bulb representing an idea, speech balloons, arrows, lines connecting labels and elements in an illustration
(4) Abstract representational devices	Analytic	Graphs, tables, symbols, formulas, schemata, flowcharts, color codes
(5) Contexts	Episodic	Characters, places, situations, stories
(6) User Interface Elements	Interactive	Text boxes, cascade menus, cursors, buttons

translation review procedures (Allalouf, 2003; Zhao and Solano-Flores, 2021).

An emerging realization concerning language resources is that language issues in testing cannot be effectively addressed without taking into consideration non-textual ways of representing information (Kopriva and Wright, 2017). Moreover, a broader view of translation as both a meaning making and meaning taking enterprise, reveals the need to recognize multiple forms of translation as intrinsic to the act of representing information (Marais, 2019, p. 122). This broader view appears to be consistent with the ultimate goal of ensuring construct equivalence across cultures and languages. A wealth of possibilities emerge. For example, in addition to replacing text in one language with text in another language, should translation concern semiotic modalities other than text (e.g., replacing illustrations used in tests)? Also, are there cases in which translation should be transmodal (e.g., replacing text with illustrations or illustrations with text)? Of course, substantial conceptual developments need to take place before these thoughts can be incorporated into testing practices.

Images

Images are semiotic resources intended to convey meaning through mainly graphic, non-textual components. Photographs, illustrations, and drawings are examples of images. Images can be characterized as *mimetic* artifacts—they serve descriptive (rather than interpretive) purposes; they are intended to show entities, rather than to tell about their characteristics.

While images vary on their level of realism (the extent to which the representation of an object resembles the object represented as it would be seen in its presence), tangibility (the extent to which the characteristics of the object represented are concrete), and completeness (the extent to which the representation includes all the elements of the object), there is always a minimum of topological correspondence between the characteristics of the object and its representation. This topological correspondence is preserved, at least to some extent, even in cartoons—which deliberately distort, magnify, minimize, or omit components of the objects they represent. The assumption that meaning in images is self-evident neglects the role of the viewer in the communicative role of images, as there is evidence that individuals with different cultural backgrounds focus on different aspects of images (e.g., Boduroglu et al., 2009).

Research on the use of images in education has been uneven and unsystematic. Through history, images have attracted the attention of researchers at scattered points in time and the aspects investigated have not followed a coherent thematic line (e.g., Fleming, 1966; Miller, 1938; Levie and Lentz, 1982). Research on the use of images in educational assessment has been, in addition, scant (e.g., Washington and Godfrey, 1974). While assessment frameworks and other documents recognize the importance of images in assessment (e.g., NGSS Lead States, 2013), they do not provide clear conceptualizations for systematic image development. As a result, items may contain images whose intended functions (e.g., as supports of the text of items, as stimulus materials, or as decorative components) are unclear or vague, and whose characteristics (e.g., complexity, style) are not consistent across items.

An important notion in the field of social semiotics is that text and image are interconnected, in the sense that the user makes meaning based on using the textual and non-textual information in combination (Kress, 2010). Consistent with this notion, there is evidence that, in making sense of items accompanied by illustrations, examinees not only use the images to make sense of the text but also use the text of items to make sense of the images (Solano-Flores et al., 2014b). Also, evidence from international test comparisons suggests that, in making sense of items, examinees from high-ranking countries have a stronger tendency than examinees from low-ranking countries to cognitively integrate text and image (Solano-Flores et al., 2016). This evidence speaks to the importance of addressing the multiple ways in which disciplinary knowledge is represented throughout the entire process of test development. Since the inception of items, images (as well as other semiotic resources) should be developed along with the text of the items.

The use of images as potential visual supports for students to understand the text of items has originated a wide variety of types of images, such as those intended to illustrate the text of an item as a whole (Kopriva, 2008; Solano-Flores, 2011; Turkan et al., 2019) and those intended to illustrate the options of multiple choice items (Noble et al., 2020). Also, thanks to the ability of computers to interact with their users, it is possible to provide pop-up images that illustrate specific words or terms and which appear on the screen when the examinee clicks on them (Guzman-Orth and Wolf, 2017; Solano-Flores et al., 2019). Due to the recency of these innovations, empirical evidence on effective design and use is just beginning to appear.

Metaphorical Devices

Metaphorical devices are representations of tangible or visible objects, events, actions, or conditions intended to represent invisible or intangible events, actions, or conditions figuratively. While the term, *metaphor* has a long use history in semiotics (see Eco and Paci, 1983), in this conceptual framework the word metaphorical is reserved to this type of semiotic resource.

Metaphorical devices originate from the need to overcome the limitations imposed by the medium in which information is represented. For example, the need to use lines to represent movement or the direction of actions originates from the limitations of representing certain actions in a given medium (e.g., Krull and Sharp, 2006; Lowe and Pramono, 2006). Arrows departing from labels and pointing at different parts of a flower are effective as a semiotic resource because they are associated to the idea of direction and precision. The cross section of a volcano showing its chimney and lava concretizes a hypothetical situation (*If we would cut a volcano by the half and see what is inside.*). A bubble representing the thoughts of a person is a proxy to intangibility and ephemerality; the text inside the balloon makes those thoughts accessible to the viewer.

Typically used in combination with images, metaphorical devices may have textual or non-textual components or both textual and non-textual components. Metaphorical devices are *diegetic*—they serve a narrative function, rather than a descriptive function. They inform the viewer about something being shown; they explain, clarify, or emphasize. An implicit assumption in the

use of metaphorical devices is that the viewer understands that they are not part of the objects represented. The arrows pointing at different parts of a flower in a science item are not intended to be interpreted by the viewer as being in the same place as the flower; the volcano is not supposed to be interpreted as actually being cut; the thought bubble is not part of the story told—the thoughts represented with words (although not the words) are.

While they are common in instructional materials, textbooks, tests, and other materials, it is possible that individuals do not learn to use and interpret most of the metaphorical devices through formal learning experiences. Indeed, it is possible that many metaphorical devices used in instructional materials and tests have been borrowed from popular culture. At least in the case of the representation of motion in static materials, the use of different semiotic resources tends to originate from the work of illustrators and graphic designers rather than from systematic work on visual literacy (de Souza and Dyson, 2007).

As with images, many metaphorical devices may be used in tests intuitively, under the assumption that they are universal and, therefore, their meaning is self-evident. However, while some semiotic resources can be readily used by individuals to represent and interpret abstract ideas such as sequence and causation (Heiser and Tversky, 2006), this may not apply to other metaphorical devices.

Abstract Representational Devices

Abstract representational devices convey meaning through the interplay of multiple representational textual and non-textual components (e.g., words, symbols, and lines). Tables, graphs, and formulas are examples of abstract representational devices. Abstract representational devices are *analytic*; they present information on different aspects or parts of an object or phenomenon in ways intended to make relationships (e.g., proportion, causation, equivalence, sequence, hierarchy, magnitude, etc.) between entities explicit (e.g., through contrast or comparison).

Abstract representational devices have no topological correspondence with the objects or phenomena they represent—most of them are based on abstractions and generalizations about the objects represented. Instead, the precision in the way in which information is presented and the relevance of the information included play a critical role in their construction. For instance, the expressions $7x + (4/3)y$ and $(7x + 4)/3y$ have different meanings due to a difference in the location of the parentheses.

Following language resources, abstract representational devices are probably the second type of semiotic resource most commonly taught in formal instruction (Macdonald-Ross, 1977). However, this does not mean that they can be used without worrying about challenges for interpretation. For example, there is evidence that it takes a great deal of time and effort for individuals to develop the habit of communicating ideas with diagrams (Uesaka and Manalo, 2012).

The belief that, because they are part of disciplinary knowledge, formal information representation devices are universal and, therefore, everybody within the same discipline interprets and use them in the same way has been long discredited (see Pimm, 1987). For example, mathematical notation varies considerably across countries (Libbrecht, 2010). As with images, the complexities of properly developing and using abstract representational devices in tests may have been underestimated by assessment programs and their characteristics are not discussed in detail in assessment frameworks and item specifications documents. For example, tables summarizing information provided by items as stimulus materials do not have a consistent style across items within the same assessment program (Solano-Flores et al., 2009).

While standards, assessment frameworks, and other normative documents address the use of graphs, charts, schemata, and other abstract representational devices, the prescriptions they provide focus on the interpretation of content-related data (National Research Council, 2012). Yet it is not uncommon for assessment programs such as PISA (e.g., OECD, 2019) to include, in items not intended to assess data interpretation or representation, tables as resources to provide contextual information and for examinees to provide their answers. Also, rarely do normative documents address the complexity of these devices as a factor to control for in the design of tests. There is evidence on the effectiveness of abstract representational devices in supporting examinees with different cultural backgrounds to understand the content of items (Martiniello, 2009). However, this evidence is difficult to generalize because available literature is not sufficiently explicit about the complexity of those representational devices. In addition, different authors classify abstract representational devices in different ways, for example, by referring to different representational devices with the same name or to the same representational device with different names (see Wang, 2012).

Contexts

Contexts are plots, scenarios, or stories used with the intent to make tasks or problems meaningful to examinees. Contexts are very common in current large-scale assessment programs. For example, a study on the use of contexts in PISA 2006 and PISA 2009 items found that about one third of the sample of items examined contained contexts in the form of a narrative (Ruiz-Primo and Li, 2016).

Contexts are *episodic*—they involve a fictitious or non-fictitious event, a set of circumstances. This event and these circumstances give rise to a problem that needs to be solved. The events or objects involved are assumed to be familiar to all examinees. Contexts may vary on their degree of concretion (the extent to which the problem resembles the kinds of problems the examinee would encounter in real life) and authenticity (the extent to which problem resembles the problems and situations that are characteristic of a given discipline or professional activity).

Although the use of contexts is not necessarily a guarantee that items tap into higher order thinking skills, their popularity may have been fueled by constructivist thinking in the field of instruction, which emphasizes situated learning (Schoenfeld, 2004). Since the 1990s, tasks situated in meaningful contexts have been regarded as potential instruments for both promoting and assessing higher order thinking skills (e.g., Shavelson et al., 1990). Yet little is known about what makes contexts effective and how

exactly they contribute to make items better (see Ruiz-Primo and Li, 2015; Ruiz-Primo et al., 2019).

At the college level, efforts to assess critical thinking have led to the development of constructed-response tasks situated in realistic, complex scenarios (Zlatkin-Troitchanskaia and Shavelson, 2019). Accomplishing context authenticity across countries takes careful work. For example, in the International Performance Assessment of Learning initiative, a great deal of the work on test development focuses on ensuring that the same context is presented in different versions according to the characteristics of each country. Also, a great effort is put into ensuring that stimulus materials such as e-mails, newspaper clips, letters, and reports that examinees are asked to read have the same appearance and style of real documents they would encounter in their countries (Shavelson et al., 2019).

While rarely is engagement mentioned in assessment normative documents, contexts are semiotic resources that potentially can capture examinees' interest during test taking (Fensham, 2009). At the same time, contexts may be distracting. There is evidence that some examinees may not be skilled enough to tell apart the problem posed by an item and the contextual information used to introduce the problem (Solano-Flores, 2011). Also, contexts may account for more for differences in student performance than the skills items are intended to assess. An investigation on inferential reading comprehension in which the narrative structure and the linguistic complexity of the texts used as stimulus materials were kept constant found that the topic of the story, more than any other factor, was the main source of score variation among second language learners in the U.S. tested in both their first language and the second language (González-Otero, 2021).

Altogether, this evidence shows that item contexts are tremendously complex, delicate semiotic resources that need to be developed carefully. If the characters, events, or objects depicted (and their appearance) are not equally familiar to all examinees, contexts may end up adding information that is irrelevant to the target construct and unnecessarily increase item difficulty.

User Interface Elements

User interface elements are textual, visual, and auditory components embedded in a computer-administered environment and intended to facilitate the interaction of the examinee with the computer for the examinee to obtain and enter information with ease. User interface elements are *interactive*—they react to the examinees' actions. They include cursors, pointers, cascade menus, buttons, boxes, hyperlinks, icons, and navigation arrows, among many other features. They are operated or activated by actions that include hovering, clicking, dragging objects, etc.

Due to globalization, the ubiquity of some platforms, and the widespread presence of certain websites in many countries, certain user interface elements may be in the process of global standardization, may be familiar to multiple populations of examinees, and may be mimicked by many other platforms—including testing platforms. However, the influence of local and regional cultural factors in this process should not be underestimated. There is evidence that the design of websites reflects the preferences, worldviews, and communication styles of the cultural contexts in which they originate. Important differences have been documented on attributes such as layout, color, links, navigation, etc. (Alexander et al., 2016). In addition, many user interface elements should not be assumed to be static. For example, icons tend to change with every version of the same software or platform (Familant and Deteweiler, 1993)—which potentially may be a challenge for interpretation.

The field of information technology has developed a wide variety of methods for website localization—the adaptation of the websites to the characteristics of a specific target country, cultural group, language, or region. These methods are intended to address subtle cultural differences (Aykin, 2005). Regrettably, while those methods are frequently used in marketing and business, they are yet to be adopted as part of the translation and adaptation practices in international test comparisons.

The assumption that a given user interface element is interpreted in the same way by everybody may not hold equally for different populations. Differences in the popularity and cost of certain devices and differences in access to computers and the internet (OECD, 2020) may create important differences in the examinees' level familiarity with different user interface elements. In online or computer-based testing, the characteristics of interface user elements may be determined by factors such as the technical properties of the software, processing speed, or hardware requirements, which may constrain or support the design possibilities of computer-administered tests in different ways (International Test Commission, 2005). Also, due to the specific characteristics of online tests (i.e., types of tasks, content area, skills targeted, school grade), certain user interface elements may need to be designed for specific tests (Bennett, 2015).

Since the early days of the internet, web designers have incorporated in their design practices the notion that different cultural groups ascribe different meanings to different colors and other object features. Those features may be used purposefully with the intent to communicate danger, joy, importance, etc. Indeed, it is well known that certain icons, colors, font styles, and other design elements can be used so frequently and consistently in the design of websites in a given country that they become cultural markers (e.g., Barber and Badre, 1998; Cyr et al., 2010). However, whether or how the interpretation of a specific user interface element varies across certain populations of students may be difficult to anticipate. An issue that adds to the challenges to fair, valid testing is the underrepresentation of certain cultural groups in the data that feeds the algorithms used by websites and search engines (Henrich et al., 2010; Noble, 2018).

Current cognitive-based approaches to test design pay special attention to the interplay between the characteristics of items and the characteristics of the knowledge and skills being assessed. Consistent with the notion that response processes do not take place separately for the target constructs and the means through which tests are administered (Ercikan and Pellegrino, 2017), a sound methodology for computer-administered and online test development should enable test developers to treat the constructs assessed and the characteristics of the interface in an integrated manner.

Reasoning from the field of cognitive psychology allows examination of the usability of user interface components—the ease with which they can be used or learned (see Preece et al., 1994; Norman, 2013). Because items, by definition, present examinees with novel situations, the creation of online items involves the design of microinteractions—contained product moments that involve a single use case (Saffer, 2014). To a large extent, the design of an online item is the design of a microinteraction whose complexity is shaped by the content assessed and the characteristics of the user interface.

BASIC IDEAS FOR SEMIOTIC TEST DESIGN

Defining Semiotic Test Design

The term, *semiotic test design* should not be confused with the term, *test design*, which is typically used in relation to the technical properties of tests (e.g., Wendler and Walker, 2006; van der Linden, 2016) and the ways in which content is covered through item and population sampling (Gonzalez and Rutkowski, 2010). While assessment frameworks and item specifications documents address the format, structure, complexity, and number of items to be included in tests (e.g., National Assessment Governing Board, 2017), they are not intended to provide detailed information on the multiple features of items.

In contrast, *semiotic test design* is concerned with the selection of optimal sets of item semiotic resources intended to meet the examinees' cultural backgrounds. Ideally, since the inception of a test, and based on the characteristics of the population of examinees, decisions should be made about the characteristics of semiotic resources to use consistently across items in ways intended to minimize challenges for interpretation due to cultural differences.

Nor the term, *semiotic test design* should be confused with *universal design* and *universal test design,* which refer to the set of basic principles and practices intended to ensure that the needs of diverse students are taken into account during the entire process of test development and to maximize accessibility to all examinees (see Lidwell et al., 2003; American Educational Research Association [AERA] et al., 2014; Thurlow and Kopriva, 2015; Sireci and O'Riordan, 2020).

While *semiotic test design* shares those goals and basic principles, it has a more explicit theoretical foundation from the field of social semiotics and its relation to cognitive science, sociolinguistics, and socio-cultural theory. More specifically, semiotic test design aims at minimizing unnecessary cognitive load in items by optimizing item representational complexity and item semiotic alignment.

Cognitive Load

Cognitive load theory comes handy in reasoning about the design of items and its impact on the working memory that an individual needs to use in responding to an item. Cognitive load theory distinguishes three types of cognitive load—intrinsic, germane, and extraneous. While intrinsic and germane cognitive load involve respectively mental processing of information that is needed to complete the task and mental processing of information into knowledge structures and their storage in long-term memory, extraneous cognitive load involves mental processing resulting from the manner in which information is presented (Sweller, 1988; Sweller et al., 1998).

A recurrent issue in testing is the increase in an item's extraneous cognitive load that takes place when, in addition to thinking about the problems posed, examinees need to figure out how they need to give their responses to items (Clariana and Wallace, 2002; Carpenter and Alloway, 2018). This concern arises, for example, in online testing endeavors that involve populations with varying levels of familiarity with computers. The generalizability of findings from research that compares the cognitive load imposed by paper-and-pencil and computer-administered tests (e.g., Priscari and Danielson, 2017) appears to be shaped by factors such as the content assessed, the socioeconomic characteristics of the population of examinees, and the examinees' familiarity with computers or the specific testing platform.

Item Representational Complexity

Item representational complexity is defined here as the multiple ways in which the semiotic resources of an item are related to each other. It is the combination, not only the sum of semiotic resources, what influences the ways in which examinees make sense of items. A key tenet in testing is that unnecessary complexity (e.g., too much wording, a crowded item layout) is a source of construct-irrelevant variance because it contributes to increasing extraneous cognitive load. Also, information provided in different sensory modalities without proper organization of different components and pieces of information may hamper, rather than facilitate, information processing because individuals need to split their attention between information provided in disparate modalities and then mentally integrate that information (see Chandler and Sweller, 1992; Mayer et al., 2001).

One of the goals of semiotic test design is to minimize the cognitive load of items by minimizing semiotic item complexity. In online testing, the inclusion of too many features in the user interface (Norman, 2013) may lead to an unnecessary increase of extraneous cognitive load. For example, an item whose response requires from the examinee building a graph by dragging and dropping bar lines into a box, labeling the axes of the graph, and typing a number in a panel, may be too complex compared to the complexity of the specific knowledge the item is intended to assess.

Experience from item writing provides good examples of the intricacies of examining representational complexity. The work on linguistic simplification as a form of testing accommodation for second language learners has focused on minimizing the lexical and syntactical complexity of items with the intent to reduce their reading demands for students who are second language learners. While some lexical variables have been found to be good predictors of item difficulty (Shaftel et al., 2006; Martiniello, 2009), linguistic simplification has been, at best, moderately effective in minimizing limited language proficiency in the language of testing as a source of error variance

(Abedi et al., 2006; Sato et al., 2010; Haag et al., 2015; Noble et al., 2020). These moderate effects suggest that linguistic simplification does not necessarily reduce the reading demands imposed by items. For example, expressing the same idea in fewer and shorter sentences may require a higher level of encoding and the use of more precise words with lower frequencies. While a shorter sentence has fewer words to read, the level of mental processing needed to decode the sentence may be higher.

Research on images provides another set of good examples of the intricacies of examining representational complexity. Consistent with approaches to measuring visual complexity based on the number of components (Forsythe et al., 2003), the analysis of complexity of illustrations used in items has been based on counting the number of different types of features they contain (e.g., color, black and white, or grayscale tonalities; zooming; symbols), as shown in **Table 2**. Based on examining items from different assessment programs, Wang (2012) identified over a hundred features of illustrations used in science items and coded the presence and absence of illustration features as dichotomous (1–0) variables, classified into several categories of illustration features. Unlike other approaches to characterizing images (which are based on broad categories such as "chart," "table," or "graph"), this coding approach has allowed systematic examination of illustrations used in different assessment programs (Solano-Flores et al., 2013, 2016; Solano-Flores and Wang, 2015; Shade, 2017).

Quantifying representational complexity also makes it possible to ensure consistency in the complexity of images across items in a test or assessment program. For example, using a set of design criteria that specified the characteristics of illustrations to be added to the text of middle school science items, Wang et al. (2012) were able to create images that had, on average, about 16 features. This number contrasts with the average (rounded) number of 22, 21, and 21 different features observed in Grades 4–12 science items respectively from China, the U.S., and TIMSS (the Trends in Mathematics and Science international assessment program).

Using number of different features as a measure of item representational complexity also makes it possible to compare in detail the characteristics of items from different countries. For example, Wang (2012) compared items from Chinese science assessment programs and items from American assessment programs. She found that, while the average number of different types of features are similar across countries, the most frequent types of features were not necessarily the same across countries. For example, photographs in illustrations were 3.52 times more frequent in items from China than in items from the U.S., whereas analogic line drawings were 3.36 times more frequent in items from the U.S. than in items from China.

TABLE 2 | Segment of the list used to code non-textual components in different assessment programs.

OBJECTS AND BACKGROUND
Image concreteness: photo; scanned document; text clip; realistic line drawing; schematic; map; silhouette; cartoon; logo; icon; emblem; metonymy; symbol; reference; entity; geometric shape
Background: with background; without background
Zooming: no zooming; zoom in; zoom out
View: external; internal; from above object; from below object; from side of object
Dimension: three dimensional; two dimensional
Relative scale of objects: proportionate; disproportionate
Color: black and white; multicolor; gray scale
Composition: single image; compound image; image in an object
TEXT IN ILLUSTRATION
Text unit: non-math/scientific sign; math/scientific sign, and notation; abbreviation; Roman numeral; Arabic numeral; letter; word; phrase; sentence; paragraph; acronym
Text function: provide label; provide a code (legend); title/caption/heading; elaborate/explain/describe; comment/note; provide instructions; provide data; text in an object
Text emphasis: capitalization; bolding; italicizing; underlying; circling
Text direction: between left and right; between top and bottom; oblique direction
CONTEXT
Socio-cultural focus: an undefined person; peers/teachers; media celebrities (characters); family/home; school/class; community/neighborhood; state/province; home country; world/global

Adapted from Wang (2012).

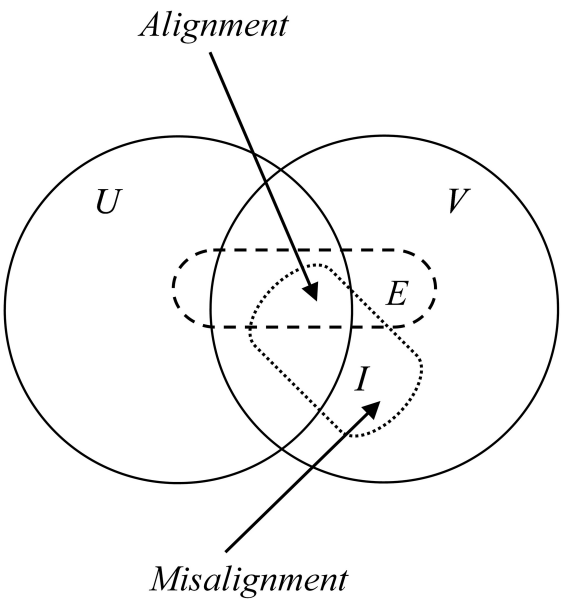

FIGURE 1 | Item semiotic alignment and misalignment.

Item Semiotic Alignment

Item semiotic alignment is defined here as shared cultural experience, the intersection of the cultural experience encoded in the semiotic resources used in an item and the examinee's cultural experience. Conversely, semiotic item misalignment can be defined as the cultural experience encoded in the semiotic resources used in an item but not shared by the examinee.

Figure 1 represents that intersection in a Venn diagram. *U* and *V* are different cultural contexts, *I* is an item originated in *U*, and *E* is an examinee's cultural experience.

Because cultural groups are not isolated, *U* and *V* are shown as intersecting. The item is assumed to encode cultural experience predominantly from *U* but also from *V*—which is represented respectively as the intersection of *I* and *U* and the intersection of *I* and *V*. Similarly, because people do not live their lives in isolation within one single cultural context, a given individual's cultural experience is assumed to develop within both *U* and *V*—a notion that is represented in the diagram respectively as the intersection of *E* and *U* and the intersection of *E* and *V*. The figure shows misalignment as partial, as it is very unlikely for an examinee not to share any cultural experience encoded by the semiotic resources of an item.

Of course, semiotic alignment is difficult to evaluate, given the thousands of possible features of items and the uniqueness of every individual's cultural experience. Yet the notion is helpful in reasoning about the ways in which the examinees' assumed cultural experience (or the lack of knowledge on the examinees' cultural experience) needs to be taken into account when developing or examining tests. For example, experience from research examining students' interpretations of contexts indicates that semiotic misalignment increases item extraneous cognitive load. The notion that an individual's socio-cultural activity takes place at different levels of social participation (apprenticeship, guided participation, and participatory appropriation; Rogoff, 1995) is key to interpreting the findings. There is evidence that, in attempting to make sense of items, examinees make connections between the contexts of items and their own personal experiences (Solano-Flores and Li, 2009, 2013). Item contexts are more meaningful to examinees when they portray situations in which they are actors, rather than observers or apprentices (Solano-Flores and Nelson-Barber, 2001; Le Hebel et al., 2013). An implication of this evidence is that, if the situations and lifestyles portrayed by items are predominantly those of a given cultural group, then contexts may fail to provide the same level of support to all students, even if those items are seemingly familiar to all.

The emotional impact of an excessive representation of a privileged segment of the society in tests may also adversely affect the performance on tests of students from certain cultural groups. There is evidence that the sole impression of being excluded or treated differently in a testing situation may affect the performance of examinees in a test (Steele and Aronson, 1998). Also, there is evidence that individuals with different cultural backgrounds may interact in different ways with tests (Cizek and Burg, 2006; Madaus and Russell, 2010). Given this evidence, it does not seem unreasonable to expect that examinees from certain cultural groups who do not have difficulty interpreting certain contexts may still feel alienated when the contexts used in items do not reflect their everyday lives and that feeling of alienation may adversely affect their performance on tests (Solano-Flores et al., 2014a).

Note that this reasoning on item semiotic alignment applies to all types of semiotic resources equally. While literature on testing and diversity has paid attention almost exclusively to language resources and contexts, other types of semiotic resources need to be considered in examining item semiotic alignment. For example, speech balloons and thought bubbles illustrate how and how frequently semiotic resources used in different cultural contexts may shape its effectiveness as means for item meaning making. Probably it is not an overstatement to say that these semiotic resources are used in many societies (Cohn, 2013). However, this does not

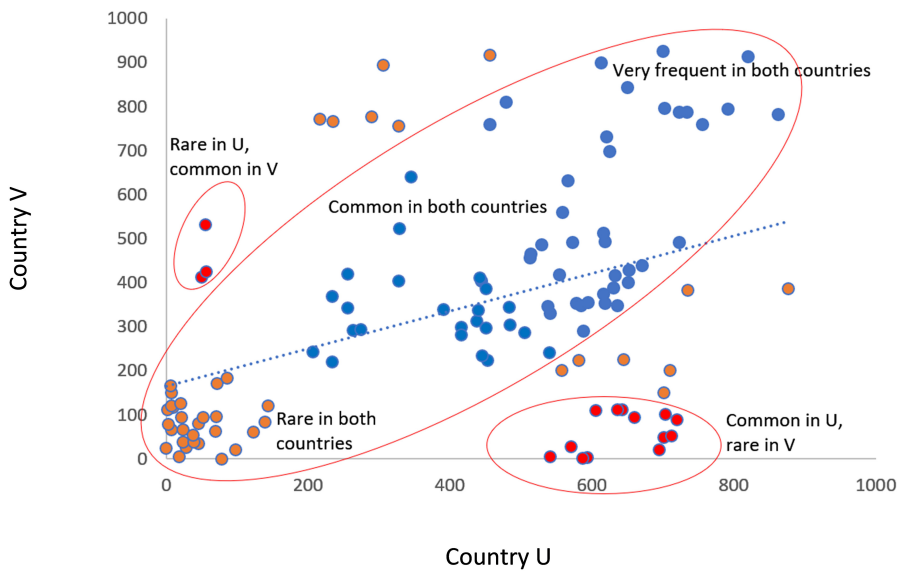

FIGURE 2 | Scattergram of the frequency of 120 item features in two hypothetical samples of 1,000 Grade 5 science items from two countries.

necessarily mean that their communicative value in tests is the same for any population. Due to their association with visual mass media (see Lefèvre, 2006), in some societies these metaphorical devices may be rarely used in textbooks and instructional materials; they may even be regarded as inappropriate for educational contexts. Even if they are common in textbooks and instructional materials, their use may not be customary in tests.

Identification of Item Features and Selection of Item Semiotic Resources in International Tests

When a test involves multiple countries, two issues need to be addressed: (1) To what extent individuals from different countries are likely to interpret the features of items as intended? and (2) How similar is the frequency with which the features of items occur in different countries?

Regarding the first question, cognitive interviews, expert panels, and focus groups can produce data on response processes (e.g., Leighton, 2017; Zhao, 2018) and, more specifically, information on the ways in which the features of items influence examinees' interpretations of items. These methods have been discussed extensively (e.g., Ericsson and Simon, 1993; Megone et al., 1994) and are not discussed here. However, it is important to mention that, because these methods are costly and time consuming, their use may need to be restricted to small numbers of item semiotic resources.

Comparative frequency analyses can produce data relevant to the second question. Frequency is used as a proxy of familiarity: if a given feature occurs with similar frequencies in different countries, it is assumed that individuals from these countries are equally familiar with it and are likely to interpret it in the same way.

Lessons from investigations like Wang's (2012), discussed above, can guide actions oriented to identifying the types of semiotic resources that are or are not likely to successfully convey the intended meaning in testing culturally and linguistically diverse populations of examinees. **Figure 2** shows a hypothetical scatterplot of the frequency of 120 features in two samples of items from two countries, *U* and *V*. In this hypothetical example, each sample contained 1,000 items and the two samples were equivalent—they comprised items of the same grade and the same content area.

The trend (dotted) line shows that, in general, the features tend to appear more frequently in items from Country *U* than in items from Country *V*. Three main types of features can be identified according to the frequencies with which they appear in the two countries: Those that are more common in *U* than in *V,* those that tend to be more common in *V* than in *U*, and those that are equally common in *U* and *V*.

If a test intended to assess populations from Countries *U* and *V* were to be created, the features with substantially different frequencies (red color) would be the first candidates for exclusion from the pool of potential item semiotic resources to be used in creating the items. In contrast, features with similar frequencies (blue color) would be the first candidates for inclusion. After this initial selection stage, a more manageable number of features would remain yet to be examined in detail. Among these semiotic resources would be, first, those with important different frequencies—outliers in the pattern of distribution of the scatterplot—and second, those with similar but low frequencies in both countries. The viability of these two types of features as semiotic resources to be used in the test could be determined through cognitive interviews, focus groups, and expert panels.

It is important to mention that the use of this approach in international test comparisons contributes to minimizing test bias across countries, not within countries. International test comparison programs are typically silent about the tremendous cultural and socio-economic differences and countries are treated as homogeneous. Yet there is evidence of tremendous test score differences attributable to socio-economic inequalities (e.g., Carnoy and Rothstein, 2013).

TABLE 3 | Use specifications for the design parameter, *Division Notation* in a hypothetical international mathematics test.

Division notation	Use specifications
$\frac{x}{y}$	In all countries, in fill-in the blank problems. Do not use in item stems.
x/y	In all countries, except Country H and Country M, in item stems.
x÷y	Only in Country H, in item stems.
x:y	Only in Country M, in item stems.
$y\sqrt{x}$	Do not use.

TABLE 4 | Design parameters of illustrations used to create illustrations accompanying the text of items for students who were not proficient in the language in which they were tested.

Design Parameter and Categories	Value or Category Selected
Framing: Yes/No	Framing
Position relative to text: Left/Right Above/Below Text	At the right of the text of the item
Drawings: Yes/No	Drawings
Color: Full Color/Gray Tone/Black and White	Only Black and white
Realistic/Fantastic representations	Only Realistic
Cartoon: Yes/No	No cartoons
Concrete objects/Abstract ideas	Only concrete objects
View level: Horizontal/From Above/From Below	Only Horizontal view
Relative Scale of Components Preserved: Yes/No	No changes in the relative scale
Perspective: Yes/No	Perspective
Labels: Yes/No	No labels
Sequences-stages: Yes/No	No stages
Backgrounds: Yes/No	No background
Metaphorical devices: Yes/No	No metaphorical devices

Adapted from Solano-Flores et al. (2014b).

Item Design Parameters

Item design parameters are variables that specify the set of semiotic resources that are to be used in the items of a

test or assessment program and the conditions under which their values or categories are to be used (Solano-Flores et al., 2014b). The specification of design parameters is intended to ensure consistency in the characteristics of items and minimize interpretation challenges for individuals from all cultural backgrounds. Current testing practices do not reach that level of standardization because item specification documents or test translation and adaptation guidelines generated by large-scale assessment programs are not sufficiently explicit about the parameters to be used in developing items.

Table 3 shows a design parameter and the use specifications for each of its categories for a hypothetical mathematics test involving multiple countries. **Table 4** provides an example of a set of design parameters used in an investigation that evaluated the effectiveness of vignette illustrations (illustrations added to the text of items with the intent to support students who were not proficient in the language in which the tests were administered to gain access to the content of items). The figure shows only one subset of parameters from a much larger possible set of design parameters that could be identified as relevant to creating vignette illustrations (Solano-Flores et al., 2014b).

Note that the specification of design parameters is not specific to semiotic resources clearly related to the content assessed. Also, which design parameters are relevant and which of their values or categories need to be selected need to be determined according to the characteristics of each assessment endeavor, such as the target populations of examinees, the content, and the cultural groups involved.

To date, design parameters have been used only in a few studies and programs (Kachchaf, 2018; Solano-Flores et al., 2019; Smarter Balanced Assessment Consortium, 2020) to produce pop-up illustrations glossaries (visual representations of words that appear on the screen when examinees click on words they do not understand) and other accessibility resources intended to provide support to students with special needs. These efforts show that it is possible to ensure standardization and efficiency in the selection and use of item semiotic resources.

SUMMARY AND CONCLUDING REMARKS

Approaches to examining cultural bias in items tend to focus on the ways in which, due to cultural differences, the characteristics of items may prevent students from properly understanding the content of items. In the absence of a conceptual framework on semiotic test design, it is difficult to link specific characteristics of items to the performance on tests of different cultural or linguistic groups or to translate the lessons learned from those experiences into improved testing practices. More specifically, in the absence of a conceptual framework on semiotic test design, it is difficult to establish the set of item features that are likely to minimize cultural bias. Item specifications documents provide coarse-grain information useful for systematically generating items according to the content and type of knowledge assessed, but they cannot provide design parameters to be used across all items within the same assessment program.

This paper has presented a conceptual framework for test design from the perspective of social semiotics. It has offered a typology for characterizing the wide variety of semiotic resources used in items and discussed challenges and possibilities in their use in the testing of culturally and linguistically diverse populations. The conceptual framework also discusses basic ideas on semiotic test design, which is intended to support the systematic selection and use of sets of semiotic resources in tests. According to the framework, differences in the frequency of semiotic resources in different societies may produce different degrees of semiotic alignment for different cultural groups. Semiotic test design allows identification of an optimal pool of semiotic resources for a test or assessment program intended to minimize extraneous cognitive load in items by minimizing item representational complexity and maximizing item semiotic alignment for the maximum number of examinees.

The conceptual framework offered makes it possible to imagine a stage in the process of test development focused on specifying design parameters that are relevant to the design of items and decide on the categories or values to apply for each design parameter. Naturally, these decisions need to be supported by information from multiple sources, such as comparative studies of tests across countries, cognitive interviews, expert panels, and focus groups with individuals from the target populations of examinees.

Semiotic test design allows development of test items based on identifying and selecting the optimal features of test items, given the cultural and linguistic characteristics of the target populations. In sum, semiotic test design offers the opportunity to address the complex representational nature of disciplinary knowledge in multicultural, multilingual contexts.

AUTHOR CONTRIBUTIONS

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

REFERENCES

Abedi, J. (2006). "Language issues in item development," in *Handbook of Test Development*, eds S. M. Downing and T. M. Haladyna (Mahwah, NJ: Lawrence Erlbaum Associates, Publishers), 377–398.

Abedi, J. (2016). "Language issues in test development," in *Handbook of Test Development*, eds S. M. Downing and T. M. Haladyna (Mahwah, NJ: Lawrence Erlbaum), 355–373.

Abedi, J., Courtney, M., Leon, S., Kao, J., and Azzam, T. (2006). *English Language Learners and Math Achievement: A Study of Opportunity to Learn and Language Accommodation*. Los Angeles, CA: University of California.

Alexander, R., Thompson, N., and Murray, D. (2016). Towards cultural translation of websites: a large-scale study of Australian, Chinese, and Saudi Arabian design preferences. *Behav. Inf. Technol.* 36, 1–13. doi: 10.1080/14781700.2019.1664318

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 [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME] (2014). *Standards for Educational and Psychological Testing*. Washington, DC: Joint Committee on Standards for Educational and Psychological Testing (U.S.).

Anagnostopoulou, K., Hatzinikita, V., Christidou, V., and Dimopoulos, K. (2013). PISA test items and school-based examinations in Greece: exploring the relationship between global and local assessment discourses. *Int. J. Sci. Educ.* 35, 636–662. doi: 10.1080/09500693.2011.604801

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

Aykin, N. (2005). "Overview: where to start and what to consider," in *Usability and Internationalization of Information Technology*, ed. N. Aykin (Mahwah, N.J: Lawrence Elbaum Associates, Publishers). doi: 10.1201/b12471

Baecker, C. R. (2010). "*A cross-cultural study on the effect of decimal separator on price perception*," in *A Work Project Presented as Part of the Requirements for the Award of a Masters Degree in Management*, (Rua da Holanda: Faculdade de Economia da Universidade Nova de Lisboa).

Barber, W., and Badre, A. (1998). "Culturability: the merging of culture and usability," in *Proceedings of the 4th Conference on Human Factors and the Web, June 5, 1998*, Basking Ridge, NJ.

Bennett, R. E. (2015). The changing nature of educational assessment. *Rev. Res. Educ.* 39, 370–407. doi: 10.3102/0091732X14554179

Boduroglu, A., Shah, P., and Nisbett, R. E. (2009). Cultural differences in allocation of attention in visual information processing. *J. Cross Cult. Psychol.* 40, 349–360. doi: 10.1177/0022022108331005

Carnoy, M., and Rothstein, R. (2013). *What do International tests Really Show about U.S. Students Erformance?*. Washington, DC: Economic Policy Institute.

Carpenter, R., and Alloway, T. (2018). Computer versus paper-based testing: are they equivalent when it comes to working memory? *J. Psychoeduc. Assess.* 37, 382–394. doi: 10.1177/0734282918761496

Chandler, P., and Sweller, J. (1992). The split-attention effect as a factor in the design of instruction. *Br. J. Educ. Psychol.* 62, 233–246. doi: 10.1111/j.2044-8279.1992.tb01017.x

Chua, H. F., Boland, J. E., and Nisbett, R. E. (2005). Cultural variation in eye movements during scene perception. *Proc. Natl. Acad. Sci. U.S.A.* 102, 12629–12633. doi: 10.1073/pnas.0506162102

Cizek, G. J., and Burg, S. S. (2006). *Addressing Test Anxiety in a High-Stakes Environment*. Thousand Oaks, CA: Corwin press.

Clariana, R., and Wallace, P. (2002). Paper-based versus computer-based assessment: key factors associated with the test mode effect. *Br. J. Educ. Technol.* 33, 593–602. doi: 10.1111/1467-8535.00294

Cohn, N. (2013). Beyond speech balloons and thought bubbles: the integration of text and image. *Semiotica* 197, 35–63. doi: 10.1515/sem-2013-0079

Council of Chief State School Officers (2015). *Science Assessment Item Collaborative item Specifications Guidelines for the Next Generation Science Standards*. Washington, DC: Council of Chief State School Officers.

Coupé, C., Oh, Y. M., Dediu, D., and Pellegrino, F. (2019). Different languages, similar encoding efficiency: comparable information rates across the human communicative niche. *Sci. Adv.* 5:eaaw2594. doi: 10.1126/sciadv.aaw2594

Cyr, D., Head, M., and Larios, H. (2010). Colour appeal in website design with and across cultures: a multi-method evaluation. *Int. J. Hum. Comput. Stud.* 68, 1–21. doi: 10.1016/j.ijhcs.2009.08.005

Davies, M., and Gardner, D. (2010). *Word Frequency List of American English*. Avaliavble at: https://www.wordfrequency.info/files/entries.pdf. (accessed February 20, 2021).

de Souza, J. M. B., and Dyson, M. C. (2007). "An illustrated review of how motion is represented in static instructional graphics," in *First Global Conference on Visual Literacies*, Seville: University of Seville. doi: 10.1016/b978-0-240-81010-2.50004-3

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

Eco, U., and Paci, C. (1983). The scandal of metaphor: metaphorology and semiotics. *Poetics Today* 4, 217–257. doi: 10.2307/1772287

Ercikan, K. (2002). Disentangling sources of differential item functioning in multi-language assessments. *Int. J. Test.* 2, 199–215. doi: 10.1207/s15327574ijt023264_2

Ercikan, K., and Pellegrino, J. W. (2017). "Validation of score meaning using examinee response processes for the next generation of assessments," in *Validation of Score Meaning in the Next Generation of Assessments*, eds K. Ercikan and J. Pellegrino (New York, NY: Routledge), 1–8. doi: 10.4324/9781315708591-1

Ercikan, K., Roth, W.-M. Simon, M., Sandilands, D., and Lyons-Thomas, J. (2014). Inconsistencies in DIF detection for sub-groups in heterogeneous language groups. *Applied Measurement in Education* 27, 275–285. doi: 10.1080/08957347.2014.944306

Ericsson, K. A., and Simon, H. S. (1993). *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA: The MIT Press. doi: 10.7551/mitpress/5657.001.0001

Familant, M. L., and Deteweiler, M. C. (1993). Iconic reference: evolving perspectives and an organizing framework. *Int. J. Man Mach. Stud.* 39, 705–728. doi: 10.1006/imms.1993.1080

Fensham, P. J. (2009). Real world contexts in PISA science: implications for context-based science education. *J. Res. Sci. Teach.* 46, 884–896. doi: 10.1002/tea.20334

Fleming, M. L. (1966). *Instructional Illustrations: A Survey of Types Occurring in Print Materials for Four Subject Areas*. Washington, D.C: U.S. Department of Health, Education & Welfare. Project No. 1381 MDEA Tittle VIIA-1381, Grant OE-7-24-0210-279.

Forsythe, A., Sheehy, N., and Sawey, M. (2003). Measuring icon complexity: an automated analysis. *Behav. Res. Methods Instrum. Comput.* 35, 334–342. doi: 10.3758/bf03202562

Gonzalez, E., and Rutkowski, L. (2010). Principles of multiple bookle matrix designs and parameter recovery in large-scale assessments. Issues and methodologies in large-scale assessments. *IERI Monogr. Ser.* 3, 125–156.

González-Otero, S. (2021). *Habilidades Lectoras Como Función de la Heterogeneidad Lingüística en Estados Unidos*. Tesis Doctoral, Universidade da Coruña, Spain.

Guzman-Orth, D., and Wolf, M. (2017). "Illustration glossaries for English learners: Findings from cognitive labs," in *Paper Presented at the Annual Conference of the American Educational Research Association*, San Antonio, TX.

Haag, N., Heppt, B., Roppelt, A., and Stanat, P. (2015). Linguistic simplification of mathematics items: effects for language minority students in Germany. *Eur. J. Psychol. Educ.* 30, 145–167. doi: 10.1007/s10212-014-0233-6

Halliday, M. A. K. (1978). *Language as a Social Semiotic*. London: Edward Arnold.

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). doi: 10.4324/9781410611758

Heiser, J., and Tversky, B. (2006). Arrows in comprehending and producing mechanical diagrams. *Cogn. Sci. A Multidiscip. J.* 30, 581–592. doi: 10.1207/s15516709cog0000_70

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behav. Brain Sci.* 33, 61–135. doi: 10.1017/s0140525x0999152x

International Test Commission (2005). *ITC Guidelines on Computer-Based and Internet Delivered Testing*. Avaliable at: www.intestcom.org. (accessed February 20, 2021).

International Test Commission (2017). *The ITC Guidelines for Translating and Adapting Tests*, 2nd Edn. Avaliable at: www.InTestCom.org. (accessed February 20, 2021).

Kachchaf, R. (2018). "Illustration glossaries: update on post-pilot analysis," in *Presentation at the Smarter Balanced Technical Advisory Committee*, Minneapolis, MN.

Kane, M. T. (1982). A sampling model for validity. *Appl. Psychol. Meas.* 6, 125–160. doi: 10.1177/014662168200600201

Kopriva, R. J. (ed.) (2008). *Improving Testing for English Language Learners*. New York, NY: Routledge.

Kopriva, R. J., and Wright, L. (2017). "Score processes in assessing academic content of non-native speakers: literature review and ONPAR summary," in *Validation of Score Meaning in the Next Generation of Assessments: The use of*

Response Processes, eds K. Ercikan and J. Pellegrino (New York, NY: Routledge), 100–112. doi: 10.4324/9781315708591-9

Kress, G. (2010). *Multimodality: A Social Semiotic Approach to Contemporary Communication*. New York, NY: Routledge. doi: 10.4324/9780203970034

Kress, G., and van Leeuwen, T. (2006). *Reading Images: The Grammar of Visual Design*, 2nd Edn. New York, NY: Routledge. doi: 10.4324/9780203619728

Krull, R., and Sharp, M. (2006). Visual verbs: using arrows to depict the direction of actions in procedural illustrations. *Inf. Des. J.* 14, 189–198. doi: 10.1075/idj.14.3.01kru

Lane, S., Raymond, M. R., Haladyna, T. N., and Downing, S. M. (2016). "Language issues in test development," in *Handbook of Test Development*, eds S. M. Downing and T. M. Haladyna (Mahwah, NJ: Lawrence Erlbaum), 3–18.

Le Hebel, F., Tiberghien, A., and Montpied, P. (2013). "Sources of difficulties in PISA science items," in *ESERA Conference 2013, Sept 2013*, eds C. P. Constantinou, N. Papadouris, and A. Hadjigeorgiou Nicosia, 76–84. Strand 11: Evaluation and assessment of students learning and development.

Lefèvre, P. (2006). "The battle over the balloon: the conflictual institutionalization of the speech balloon in various European cultures," in *Image Narrative: Online Magazine of the Visual Narrative*, 14. Available online at: http://www.imageandnarrative.be/inarchive/painting/pascal_levevre.htm (accessed February 21, 2021).

Leighton, J. P. (2017). "Collecting and analyzing verbal process data in the service of validity and interpretive arguments," in *Validation of Score Meaning in the Next Generation of Assessments: The use of Response Processes*, eds K. Ercikan and J. Pellegrino (New York, NY: Routledge), 25–38. doi: 10.4324/9781315708591-3

Lemke, J. L. (1990). *Talking Science: Language, Learning, and Values*. Norwood, NJ: Ablex Publishing.

Lemke, J. L. (1998). "Multiplying meaning: visual and verbal semiotics in scientific text," in *Reading Science: Critical and Functional Perspectives on Discourses of Science*, eds J. R. Martin and R. Veel (New York, NY: Routledge), 87–113.

Levie, W. H., and Lentz, R. (1982). Effects of text illustrations: a review of research. *Educ. Commun. Technol. J.* 30, 195–232.

Libbrecht, P. (2010). "Notations around the world: census and exploitation," in *Proceedings of the 10th ASIC and 9th MKM International Conference, and 17th Calculemus Conference on Intelligent Computer Mathematics*, New York, NY: ACM, 398–410. doi: 10.1007/978-3-642-14128-7_34

Lidwell, W., Holden, K., and Butler, J. (2003). *Universal Principles of Design: 125 Ways to Enhance Usability, Influence Perception, Increase Appeal, Make Better Design Decisions, and Teach Through Design*. Beverly, MA: Rockport Publishers, Inc.

Lotman, Y. M., Uspensky, B. A., and Mihychuk, G. (1978). On the semiotic mechanism of culture. *New Lit. Hist.* 9, 211–232. doi: 10.2307/468571

Lowe, R., and Pramono, H. (2006). Using graphics to support comprehension of dynamic information in texts. *Inf. Des. J.* 14, 22–34. doi: 10.1075/idj.14.1.04low

Macdonald-Ross, M. (1977). Graphics in texts. *Rev. Res. Educ.* 5, 49–85. doi: 10.2307/1167172

Madaus, G., and Russell, M. (2010). Paradoxes of high-stakes testing. *J. Educ.* 190, 21–30. doi: 10.1177/0022057410190001-205

Marais, K. (2019). *A (bio)semiotic Theory of Translation: The Emergence of Social-Cultural Reality*. New York, NY: Routledge. doi: 10.4324/9781315142319

Martinez, M. E. (1999). Cognition and the question of test item format. *Educ. Psychol.* 34, 207–218. doi: 10.1207/s15326985ep3404_2

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educ. Assess.* 14, 160–179. doi: 10.1080/10627190903422906

Mayer, R. E., Heiser, J., and Lonn, S. (2001). Cognitive constraints on multimedia learning: when presenting more material results in less understanding. *J. Educ. Psychol.* 93, 187–198. doi: 10.1037/0022-0663.93.1.187

Megone, M. E., Cai, J., Silver, E. A., and Wang, N. (1994). Validating the cognitive complexity and content quality of a mathematics performance assessment. *Int. J. Educ. Res.* 21, 317–340. doi: 10.1016/s0883-0355(06)80022-4

Miller, W. A. (1938). Reading with and without pictures. *Elem. Sch. J.* 38, 676–682. doi: 10.1086/462248

Mislevy, R. J., and Haertel, G. D. (2007). Implications of evidence centered design for educational assessment. *Educ. Mea. Issues Pract.* 25, 6–20. doi: 10.1111/j.1745-3992.2006.00075.x

Nagy, W. W., and Anderson, R. C. (1984). How many words are there in printed school english?. *Read. Res. Q.* 19, 304–330. doi: 10.2307/747823

Nation, P. (2014). How much input do you need to learn the most frequent 9,000 words. *Read. Foreign Lang.* 26, 1–16.

National Assessment Governing Board (2017). *Mathematics Framework for the 2017 National Assessment of Educational Progress*. Washington, CG: National Assessment Governing Board.

National Research Council (2006). "Systems for state science assessment. committee on test design for K-12 science achievement," in *Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education*, eds M. R. Wilson and M. W. Bertenthal (Washington, DC: The National Academies Press).

National Research Council (2012). *A Framework for Science K-12 Education: Practices, Cross-Cutting Concepts, and Core Ideas*. Washington, DC: The National Academy Press.

National Research Council (2014). *Developing Assessments for the Next Generation Science Standards*. Washington, DC: The National Academies Press, doi: 10.17226/18409

NGSS Lead States (2013). *Next Generation Science Standards: For states, by States*. Washington, DC: The National Academies Press.

Nisbett, R. (2003). *The Geography of Thought*. New York, NY: Free Press.

Noble, S. U. (2018). *Algorithms of Oppression. How Search Engines Reinforce Racism*. New York, NY: New York University Press. doi: 10.2307/j.ctt1pwt9w5

Noble, T., Sireci, S. G., Wells, C. S., Kachchaf, R. R., Rosebery, A. A., and Wang, Y. C. (2020). Targeted linguistic simplification of science test items for English learners. *Am. Educ. Res. J.* 57, 2175–2209. doi: 10.3102/0002831220905562

Norman, D. (2013). *The Design of Everyday Things: Revised and Expanded Edition*. New York, NY: Basic Books. (accessed February 20, 2021).

OECD (2019). *PISA Test*. Avaliable at: https://www.oecd.org/pisa/test/. (accessed February 20, 2021).

OECD (2020). *Access to Computers From Home (indicator)*. Paris: OECD, doi: 10.1787/a70b8a9f-en

Packer, M., and Cole, M. (2020). "The institutional foundations of human evolution, ontogenesis, and learning," in *Handbook of the Cultural Foundations of Learning*, eds N. S. Nasir, C. D. Lee, R. Pea, and M. M. de Royston (New York, NY: Routledge), 3–23. doi: 10.4324/9780203774977-2

Pellegrino, J. W., Chudowsky, N., and Glaser, R. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, DC: National Academy Press.

Pesce, S. (2018). "From comprehensive research to semiotic approaches to education: a subjective genealogy of educational semiotics," in *Semiotic Theory of Learning*, eds A. Stables, W. Nöth, A. Olteanu, S. Pesce, and E. Pikkarainen (New York, NY: Routledge), 145–157. doi: 10.4324/9781315182438-11

Pimm, C. (1987). *Speaking Mathematically: Communication in Mathematics Classrooms*. London: Routledge & Kegan Paul Ltd.

Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S., and Carey, T. (1994). *Human-Computer Interaction*. Workingham: Addison-Wesley.

Priscari, A. A., and Danielson, J. (2017). Computer-based versus paper-based testing: investigating testing mode with cognitive load and scratch paper use. *Comput. Hum. Behav.* 77, 1–10. doi: 10.1016/j.chb.2017.07.044

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: a random effects synthesis of correlations. *J. Educ. Meas.* 40, 163–184. doi: 10.1111/j.1745-3984.2003.tb01102.x

Rogoff, B. (1995). "Observing sociocultural activity on three planes: participatory appropriation, guided participation, and apprenticeship," in *Sociocultural Studies of Mind*, eds J. V. Wertsch, P. del Río, and A. Alvarez (New York, NY: Cambridge University Press). doi: 10.1017/CBO9781139174299.008

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. Record* 117, 1–36. Available online at: https://www.tcrecord.org (accessed March 21, 2021).

Ruiz-Primo, M. A., Baxter, G. P., and Shavelson, R. J. (1993). On the stability of performance assessments. *J. Educ. Meas.* 30, 41–53. doi: 10.1111/j.1745-3984.1993.tb00421.x

Ruiz-Primo, M. A., and Li, M. (2016). PISA science contextualized items: the link between the cognitive demands and context characteristics of the items. *RELIEVE* 22:art.M11. doi: 10.7203/relieve.22.1.8280

Ruiz-Primo, M. A., Li, M., Minstrell, J., Kanopka, J., Hernandez, P., Dong, D, et al. (2019). Contextualized science assessments: addressing the use of information and generalization of inferences of students' performance. *Paper presented at the AERA Annual Meeting*, Toronto, ON: Canada.

Saffer, D. (2014). *Microinteractions: Designing With Details*. Sebastopol, CA: O'Reilly.

Sato, E., Rabinowitz, S., Gallagher, C., and Huang, C.-W. (2010). *Accommodations for English Language Learner Students: The Effect of Linguistic Modification of Math Test Item Sets*. Washington, DC: National Center for Education Evaluation and Regional Assistance. (NCEE Report 2009-4079).

Schoenfeld, A. H. (2004). The math wars. *Educ. Policy* 18, 253–286. doi: 10.1177/0895904803260042

Shade, C. (2017). *Mathematics Assessment in the Race to the Top era: An Exploratory Study of the Semiotic Resources in Large-Scale Assessment and Their use by Emergent and Non-Emergent Bilingual Students*. Doctoral dissertation, University of Colorado Boulder, Boulder, CO.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., and Poggio, G. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educ. Assess.* 11, 105–126. doi: 10.1207/s15326977ea1102_2

Shavelson, R. J., Carey, N. B., and Webb, N. M. (1990). Indicators of science achievement: options for a powerful policy instrument. *Phi Delta Kappan* 71, 692–697.

Shavelson, R. J., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., and Marino, J. P. (2019). Assessment of university students' critical thinking: next generation performance assessment. *In. J. Test.* 19, 337–362. doi: 10.1080/15305058.2018.1543309

Sireci, S. G., and Faulkner-Bond, M. (2015). Promoting validity in the assessment of ELs. *Rev. Res. Educ.* 39, 215–252. doi: 10.3102/0091732X14557003

Sireci, S. G., and O'Riordan, M. (2020). "Comparability when assessing Individuals with disabilities," in *Comparability of Large-Scale Educational Assessments: Issues and Recommendations*, eds A. I. Berman, E. H. Haertel, and J. W. Pellegrino (Washington, DC: National Academy of Education), 177–204.

Smarter Balanced Assessment Consortium (2020). *Usability, Accessibility, and Accommodations Guidelines*. Avaliable at: https://portal.smarterbalanced.org/library/en/usability-accessibility-and-accommodations-guidelines.pdf. (accessed February 20, 2021).

Solano-Flores, G. (2008). Who is given tests in what language by whom, when, and where? The need for probabilistic views of language in the testing of English language learners. *Educ. Res.* 37, 189–199. doi: 10.3102/0013189x08319569

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. D. R. Basterra, E. Trumbull, and G. Solano-Flores (New York, NY: Routledge), 3–21.

Solano-Flores, G. (2012). *Translation Accommodations Framework for Testing English Language Learners in Mathematics*. Avaliable at: https://portal.smarterbalanced.org/library/en/translation-accommodations-framework-for-testing-english-language-learners-in-mathematics.pdf (accessed September 18, 2012).

Solano-Flores, G., Backhoff, E., and Contreras-Niño, L. A. (2009). Theory of test translation error. *Int. J. Test.* 9, 78–91. doi: 10.1080/15305050902880835

Solano-Flores, G., Barnett-Clarke, C., and Kachchaf, R. (2013). Semiotic structure and meaning making: the performance of English language learners on mathematics tests. *Educ. Eval.* 18, 147–161. doi: 10.1080/10627197.2013.814515

Solano-Flores, G., Chía, M. Y., and Kachchaf, R. (2019). Design and use of pop-up illustration glossaries as accessibility resources for second language learners in computer-administered tests in a large-scale assessment system. *Int. Mul. Res. J.* 13, 277–293. doi: 10.1080/19313152.2019.1611338

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

Solano-Flores, G., and Li, M. (2013). Generalizability theory and the fair and valid assessment of linguistic minorities. *Educ. Res. Eval.* 19, 245–263. doi: 10.1080/13803611.2013.767632

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

Solano-Flores, G., Shade, C., and Chrzanowski, A. (2014a). *Item Accessibility and Language Variation Conceptual Framework. Submitted to the Smarter Balanced Assessment Consortium*. Avaliable at: https://portal.smarterbalanced.org/library/en/item-accessibility-and-language-variation-conceptual{\penalty-\@M}-framework.pdf. (accessed February 20, 2021).

Solano-Flores, G., Wang, C., Kachchaf, R., Soltero-Gonzalez, L., and Nguyen-Le, K. (2014b). Developing testing accommodations for English language learners: illustrations as visual supports for item accessibility. *Educ. Assess.* 19, 267–283. doi: 10.1080/10627197.2014.964116

Solano-Flores, G., and Wang, C. (2015). Complexity of illustrations in PISA-2009 science items and its relationship to the performance of students from Shanghai-China, the United States, and Mexico. *Teach. Coll. Record* 117, 1–18.

Solano-Flores, G., Wang, C., and Shade, C. (2016). International semiotics: item difficulty and the complexity of science item illustrations in the PISA-2009 international test comparison. *Int. J. Test.* 16, 205–219. doi: 10.1080/15305058.2015.1099534

Stables, A. (2016). Edusemiotics as process semiotics: towards a new model of semiosis for teaching and learning. *Semiotica* 212, 45–58. doi: 10.1515/sem-2016-0126

Steele, C. M., and Aronson, J. (1998). "Stereotype threat and the test performance of academically successful African Americans," in *The Black–White Test Score Gap*, eds C. Jencks and M. Phillips (Washington, DC: Brookings Institution Press), 401–427.

Sweller, J. (1988). Cognitive load during problem solving: effects on learning. *Cogn. Sci.* 12, 257–285. doi: 10.1207/s15516709cog1202_4

Sweller, J., van Merrienboer, J. J., and Paas, F. G. (1998). Cognitive architecture and instructional design. *Educ. Psychol. Rev.* 10, 251–296. doi: 10.1023/A:1022193728205

Thurlow, M. L., and Kopriva, R. J. (2015). Advancing accessibility and accommodations in content assessments for students with disability and English learners. *Rev. Res. Educ.* 39, 331–369. doi: 10.3102/0091732X14556076

Turkan, S., Lopez, A., Lawless, R., and Tolentino, F. (2019). Using pictorial glossaries as an accommodation for English learners: an exploratory study. *Educ. Assess.* 24, 235–265. doi: 10.1080/10627197.2019.1615371

Uesaka, Y., and Manalo, E. (2012). Task-related factors that influence the spontaneous use of diagarams in math word problems. *Appl. Cogn. Psychol.* 26, 251–260. doi: 10.1002/acp.1816

van der Linden, W. J. (2016). "Optimal test assembly," in *Handbook of Test Development*, 2nd Edn, eds S Lane, M. R. Raymond, and T. M. Haladyna (New York, NY: Routledge), 507–530.

van Leeuwen, T. (2004). *Introducing Social Semiotics*. New York, NY: Routledge. doi: 10.4324/9780203647028

Wang, C. (2012). *The Use of Illustrations in Large-Scale Science Assessment: A Comparative Study*. Doctoral dissertation, University of Colorado Boulder, Boulder, CO.

Wang, C., Chia, M., Kachchaf, R., and Solano-Flores, G. (2012). "Item illustration complexity and the performance of English language learners in a science test," in *Paper Presented at the Annual Conference of the American Educational Research Association*, Vancouver.

Washington, W. N., and Godfrey, R. R. (1974). The effectiveness of illustrated items. *J. Educ. Meas.* 11, 121–124. doi: 10.1111/j.1745-3984.1974.tb00981.x

Wendler, C. L. W., and Walker, M. E. (2006). "Practical issues in designing and maintaining multiple test forms for large-scale programs," in *Handbook of Test Development*, eds S. M. Downing and T. M. Haladyna (New York, NY: Lawrence Erlbaum Associates, Publishers), 445–467.

Winter, P., Kopriva, R. J., Chen, S., and Emick, J. (2006). Exploring individual and item factors that affect assessment validity for diverse learners: results from a large-scale cognitive lab. *Learn. Individ. Differ.* 16, 267–276. doi: 10.1016/j.lindif.2007.01.001

Zhao, X. (2018). *Test Translation Review Procedures in International Large-Scale Assessment: Sensitivity to Culture and Society*. Doctoral dissertation, University of Colorado Boulder, Boulder, CO.

Zhao, X., and Solano-Flores, G. (2021). Testing across languages in international comparisons: cultural adaptation of consensus-based test translation review procedures. *J. Multiling. Multicult. Dev.* doi: 10.1080/01434632.2020.1852242

Zlatkin-Troitchanskaia, O., and Shavelson, R. J. (2019). Editorial: advantages and challenges of performance assessment of student learning in higher education. *Br. J. Educ. Psychol.* 89, 413–415. doi: 10.1111/bjep.12314

Conflict of Interest: 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.