Smarter Balanced Assessment Consortium:
Item Accessibility and Language Variation Conceptual Framework

Submitted to the Smarter Balanced Assessment Consortium by
Guillermo Solano-Flores
Chelsey Shade
Ashley Chrzanowski

October 10, 2014

ITEM ACCESSIBILITY AND LANGUAGE VARIATION CONCEPTUAL FRAMEWORK

Table of Contents

1. EXECUTIVE SUMMARY 1
Goals 1
Theoretical Perspectives 2
Strategy of Inquiry 3
Data Interpretation 3
Outcomes 4
Organization of the Conceptual Framework 5
2. THEORETICAL UNDERPINNINGS 6
Sociocultural Theory 6
Sociolinguistics 7
Social Semiotics 8
Language as a Source of Measurement Error 9
Test Translation Error 10
Key Issues on Language Variation and Item Accessibility 12
Interaction of Linguistic, Cultural, Social, and Design Factors 12
Language, Language Varieties, and Linguistic Groups as Dynamic Categories 12
Test Development and Test Translation as a Continuum 13
3. BASIC CONCEPTS IN LANGUAGE VARIATION AND ITEM ACCESSIBILITY 15
Language Variation 15
English Language Learners and English Language Proficiency 16
Dialect 16
Register, Academic Language, and Natural Language 16
Translation 17
Constituent 18
Glossary 18
Item Accessibility 18
Construct 19
Semiotic Mode 19
Cognitive Load 20
Affordance 20

Visibility 21
Usability 21
Item Accessibility Resource 21
Design 22
User Interface 22
4. PRINCIPLES FOR THE DESIGN OF ITEM ACCESSIBILITY RESOURCES 23
Consistency 23
User Interface 23
Text 25
Symmetry 26
Contrast 27
Highlighting 27
Contrasting in Stacked Translations 28
Meaningfulness 29
Culture and Meaningfulness 29
Equitable Meaningfulness 30
Design Constraints 31
Text Length 32
Grammatical Features 32
Restricted-Use Constituents 32
Fictitious Character Names 33
Customization and Design from Scratch 35
Pragmatic Suitability 36
Standardization 38
5. DEVELOPMENT OF GLOSSARIES 42
Glossability Analysis 42
Linguistic Demands of Constituents in Natural and Academic Language 43
Morphological and Semantic Correspondence of Constituents 45
Glossary Design 47
Types of Glossaries 47
Semantic Space 49
Selection of Optimal Constituents 51
Consistency of Features 52

Glossing Consistency 53
Horizontal Consistency. 53
Vertical Consistency. 54
Functional Consistency. 54
Glossing Density 54
6. MODEL FOR THE INCLUSION OF LANGUAGES IN ASSESSMENT SYSTEMS 55
Basic Concepts 55
Relevance Factors 55
Viability Factors 56
Priority Space 58
Steps for Language Selection 59
1. *Creation of a Language Inclusion Committee* 59
2. *Call for Nomination of Languages* 59
3. *Preliminary Analysis* 61
4. *Priority Analysis* 61
5. *Reporting* 62
Notes 63
References 64

ITEM ACCESSIBILITY AND LANGUAGE VARIATION CONCEPTUAL FRAMEWORK

1. EXECUTIVE SUMMARY

The Smarter Balanced Assessment Consortium has developed a set of usability, accessibility, and accommodation resources intended to support students in gaining access to the content of items in its Mathematics and English Language Arts assessments. These accessibility resources comprise a wide variety of tools. Some of them are non-embedded accessibility resources that are provided locally. Others, which are the main focus of this study, are provided through the platform with which Smarter Balanced assessments are to be administered online. They allow customization of the display of items and the ways in which they are administered to students according to their needs. Ultimately, these accessibility resources are intended to increase the validity of interpretations of Smarter Balanced assessment scores by reducing the amount of measurement error attributable to factors that are irrelevant to the constructs test items are intended to assess.

Language variation is one of the sources of measurement error addressed by the item accessibility resources offered by Smarter Balanced. Regardless of the type of test item (e.g., multiple-choice, constructed response, hands-on, computer simulation, etc.), testing depends largely on the use of language as a vehicle for administering tests and capturing students' responses. If students are not familiar with the language or style in which the directions and the contextual information of items are worded and the ways in which problems and tasks are formulated, then their knowledge of the target domain is confounded with their proficiency in the language of testing. This is especially the case for English language learners (ELLs)—students who are developing English as a second language while they continue developing their first languages—and users of non-standard forms of English.

Smarter Balanced offers two broad types of accessibility resources intended to address language variation as a source of measurement error—stacked translations and pop-up glossaries (see Smarter Balanced Assessment consortium, 2014). Stacked translations (currently available only in Spanish) are full-text translations of items into the native, first language (L1) of ELL students. Pop-up glossaries are lists of synonyms or definitions of strings of words or terms that act together as a whole to encode meaning in a specific way and which are not a part of the academic language students should possess as part of the knowledge assessed and yet may pose challenges to properly understanding the content of items. Academic language constituents (i.e., those that are part of the disciplinary knowledge) are excluded from pop-up glossaries because they are part of the knowledge assessed. Two types of pop-up glossaries are offered, L1 glossaries and English glossaries. The former provide translations of selected constituents into the ELL students' L1; the latter provide synonyms or definitions of selected constituents that are likely to pose linguistic challenges beyond the content being assessed to both native English speaking and ELL students.

While, in principle, translation and glossary supports are simple notions, important conceptual and practical challenges need to be properly addressed if they are to be effective accessibility resources. Examples of the many tasks that should be guided by principled practice are: (1) identifying constituents that are likely to pose linguistic challenges that hinder students' ability to make sense of an item; (2) determining how constituents should be translated in ways that are likely to be understood as intended by users of multiple dialects of the same language; (3) determining when a given constituent is or is not part of the academic language of the knowledge being assessed; and (4) displaying the translation of a constituent in a way that is clear and simple to all test takers.

Goals

In January 2014, Smarter Balanced hired researchers from the University of Colorado Boulder to develop a conceptual framework on language variation and item accessibility. This conceptual framework should allow test developers and test translators to properly address language variation

through the pop-up glossaries and stacked translation accessibility resources offered by the consortium.

Smarter Balanced also commissioned the researchers to develop a decision model for the inclusion of languages in the assessment consortium. This decision model should enable Smarter Balanced to determine, given a set of criteria and information on certain, critical variables, which languages or dialects should be included for translation to properly serve ELLs.

This document contains the conceptual framework on language variation and item accessibility and the decision model for the inclusion of languages and dialects.

Theoretical Perspectives

To develop the conceptual framework, the researchers adopted a probabilistic, bottom-up, context-based, and design-oriented stand. First, a probabilistic view recognizes uncertainty and fuzziness in the boundaries between language varieties and linguistic groups; it alerts about the limitations of treating language varieties and linguistic groups as non-overlapping categories. Second, a bottom-up view of language is descriptive, rather than prescriptive, in the ways in which language use is examined; it focuses on the ways in which educators, their students, and their communities use language, and is sensitive to language variation across regional, ethnic, and cultural groups. Third, a context-based perspective is sensitive to the fact that item accessibility resources need to be item-specific. A given constituent's translation or rephrasing should be based on the characteristics of the item in which it appears, such as the contextual information it provides and the grammatical structure of the entire sentence in which the constituent appears in the original text. Fourth, a design-oriented perspective aims at determining systematically the characteristics of item accessibility resources based on current principles from the field of design and the language sciences.

The researchers identified five theoretical perspectives as critical to the development of the conceptual framework. These fields informed on the specific aspects of language variation and item accessibility as summarized below:

- Sociocultural theory—how culture influence mind, thinking, and individuals' worldviews.
- Sociolinguistics—the nature of languages, linguistic groups, and varieties of language.
- Social semiotics—how meaning is conveyed and constructed through various forms of representation of information, and how this process is mediated by culture.
- View of language as a source of measurement error—the relation between language variation and error variance, and the actions that can contribute to minimizing error variance due to language through the process of test development
- Theory of test translation error—the nature of languages as partially-equivalent meaning encoding systems, and the view of translated test items as entities within a probabilistic space determined by the frequency and severity of translation errors.

Strategy of Inquiry

As a proof of concept for this set of theoretical perspectives, the team of researchers conducted, during the month of July of 2014, two panels on language variation and item accessibility with educators who taught in classrooms with high linguistic diversity. The overall goal was to examine how effectively these theoretical perspectives allow characterization of main issues in the development and use of Smarter Balanced item accessibility resources.

Two two-day panels were conducted in a Smarter Balanced state on the West Coast. Panel 1 focused on language variation among ELL students. Six bilingual (English-Spanish) educators participated in this panel. These educators taught in classrooms or were language resource professionals in schools with high enrollments of ELLs whose first language was Spanish. These professionals examined a sample of 60 Smarter Balanced items and their L1 and English glossaries. For two items, the panelists also examined the stacked, full translations made available as accessibility resources. The sample of items represented the wide variety of formats used by Smarter Balanced in Grades 3-12. With facilitation from project staff, and based on their knowledge of the use of Spanish in their schools and communities, the panelists identified and discussed constituents not glossed, and which could still be challenging to ELL, native Spanish speaking students. Also, they discussed the pertinence of the glossaries offered in the items and proposed, when appropriate, alternative Spanish versions of those glossaries.

Panel 2 focused on English variation among native English speakers. Seven educators participated in this panel. These educators taught in classrooms with high linguistic diversity and high enrollments of African American students. These professionals examined the English glossaries available in the same sample of items examined in Panel 1. Also with facilitation from project staff, and based on their knowledge of the use of English in their schools and communities, the panelists identified and discussed constituents not glossed, and which could still be challenging to their students. Also, they discussed the pertinence of the glossaries offered in the items and proposed, when appropriate, alternative versions of those glossaries.

Data Interpretation

Generalizations of the outcomes from the panels should be made cautiously. The two sets of panel participants and their students are representative of important demographic segments of the U.S. population. Yet, given the small number of panels and their limited duration, the issues identified should not be assumed to exhaust the wide variety of linguistic and sociocultural issues that would be likely identified in multiple, linguistically-diverse classrooms. Therefore, no attempt was made to generalize the findings from these panels to all classroom contexts in the United States.

Keeping those limitations in mind, the information collected was used in the development of the conceptual framework to: (1) determine whether the issues identified in the panels could be characterized conceptually using the theoretical perspectives mentioned above; (2) gain knowledge on issues of language variation that are relevant to the design of Smarter Balanced items; (3) assess the viability of the actions that should be taken to properly address language variation and item accessibility; (4) identify the critical aspects that should be refined in order to streamline the design and analytical approaches described by the framework; and (5) illustrate main issues of test translation, adaptation, and glossing that are relevant to the development of Smarter Balanced accessibility resources.

The information on the themes that emerged during the panel discussions and on the content of the comments, suggestions, and concerns expressed by the participants was organized according to eleven types of issues:

- Selection and display of constituents in the Spanish pop-up glossary and stacked translation
- Selection and display of constituents in the English glossary
- Design, structure, and wording of the items
- Dialect
- Use of language in academic contexts
- Culture
- Equity
- Safety of untargeted test takers (students who do not need the item accessibility resources)
- Sensitivity of the accessibility resources to individual test takers' needs
- Fidelity with which any improvement on the accessibility resources can be implemented
- Usability

Outcomes

The experience gained from the panels allowed the researchers to confirm that the five theoretical perspectives allow proper characterization and interpretation of a wide variety of issues relevant to language variation and the design of item accessibility resources. Using this knowledge and a probabilistic, bottom-up, context-based, design-oriented approach, the team of researchers was able to identify four major sets of challenges that test developers and test translators need to address in order to properly address language variation and item accessibility.

One set of challenges pertains to the scarcity of procedures and criteria available for properly creating accessibility resources and evaluating their effectiveness. This set of challenges also has to do with the tremendous dialect variation that may exist within the first language of certain ELL groups. Questions that need to be answered include: What procedures should be used to ensure that a translation truly supports ELLs who are speakers of a given L1, in spite of the tremendous dialect variation that may exist within that L1? What criteria should be used to determine which terms are likely to pose more or fewer challenges to ELLs? How should dialect variation be addressed to ensure that the glossaries included in a given item as accessibility resources provide equal support for the majority of the target students?

A second set of challenges relates to language variation and the wide variety of non-standard forms of English. When the knowledge and skills a given item intends to assess are unrelated to vocabulary or grammar, then the vocabulary, discursive structures, and styles used in the wording of items can privilege mainstream students, users of a standard form of English, over students who are users of non-standard forms of English, especially ethnic minorities and students of low socio-economic status. English glossaries are an ideal accessibility tool that can potentially contribute to fair assessment by providing students with a range of alternative constituents for certain terms and expressions. Nevertheless, some criteria and procedures are yet to be developed that allow for their systematic development to ensure their effectiveness.

A third set of challenges concerns the fact that the categories, *academic language* and *everyday language* are highly imprecise. For example, while some constituents can be easily identified as belonging to the academic language that is part of the knowledge in a given content area, the

distinction between academic and non-academic language is somewhat arbitrary for many constituents. Such is the case of constituents used in both academic and non-academic contexts, constituents with the same morphology and different meanings in these contexts, or constituents that are not typically associated with disciplinary knowledge, yet are more frequent in academic contexts.

A fourth set of challenges has to do with cultural aspects relevant to usability in the design of items. In general, usability can be defined as the extent to which the characteristics of an object tell the user of that object how it must be used. When usability is not properly attained, test takers, as users of the software with which they are assessed, may experience considerable difficulty figuring out what items are about and how they should respond to them. Often, an implicit assumption in the design of items is that all test takers are equally familiar with the arrangement of their components (e.g., the relative position of their statements, tables, figures, and questions, and the spaces where responses are to be provided). Another implicit assumption is that the icons and tools in computer-administered items convey the same meaning to all students. In actuality, cultural and linguistic backgrounds influence how individuals interpret images, how they make sense of visual information, and how they interact with computers.

Organization of the Conceptual Framework

Taking into consideration these sets of challenges, the team of researchers decided to organize the conceptual framework into six chapters:

1. Executive Summary
2. Theoretical Underpinnings
3. Basic Concepts in Language Variation and Item Accessibility
4. Principles for the Design of Item Accessibility Resources
5. Development of Glossaries
6. Model for the Inclusion of Languages in Assessment Systems

The document contains a specific chapter on the development of glossaries. While stacked translation is the other main type of accessibility resource concerning language variation, glossing poses an especially complex set of challenges for proper design. In addition, relevant issues concerning full translation are discussed in the assessment translation framework created for Smarter Balanced (Solano-Flores, 2012).

2. THEORETICAL UNDERPINNINGS

This conceptual framework on language variation and item accessibility can be described as probabilistic, bottom-up, context-based, and design-oriented. First, it addresses language variation without assuming the existence of few, non-overlapping language varieties; it recognizes uncertainty and fuzziness in the boundaries between language varieties and linguistic groups. Second, it focuses on information provided by the users of language, particularly educators from diverse linguistic and cultural contexts who are familiar with the varieties of languages used in their schools and communities. Third, it addresses the fact that many translations and adaptations need to be specific to the characteristics of each item. How a constituent is translated, rephrased, or defined needs to be in accord with the characteristics of the specific item in which it appears. Fourth, it aims at ensuring that the characteristics of item accessibility resources are in accord with current knowledge in both the field of design and the language sciences.

The first section of this chapter discusses five main theoretical perspectives that inform the conceptual framework, and which are compatible with the probabilistic, bottom-up, context-based, and design-oriented approach. These five theoretical perspectives are: sociocultural theory, sociolinguistics, social semiotics, the view of language as a source of measurement error, and the theory of test translation error. Although a bit overly simple, the discussion illustrates the complex challenges that language variation poses for item accessibility.

No attempt has been made to discuss in full these theoretical perspectives. Rather, the discussion focuses on examining the aspects that are more relevant to the conceptual framework or which provide its theoretical foundation. It is important to mention that although these theoretical perspectives are discussed as separate entities, they are interrelated. For example, the perspective of language as a source of measurement error uses both methods from generalizability theory (a psychometric theory of measurement error) and reasoning and principles from sociolinguistics. Also, as sociolinguistics and social semiotics evolve, they cannot be thought of as independent fields of knowledge.

The second section examines main issues that should be taken into consideration throughout the entire process of test development and test translation in order to properly address the intersection of language variation and item accessibility. The need for interdisciplinary approaches and the coordinated work of test developers and test translators is emphasized.

Sociocultural Theory

Sociocultural theory—the theory that examines thinking as a cultural phenomenon—allows identification of the social and cultural aspects that influence students' interpretations of items. Sociocultural theory is grounded in the premise that individuals do not act directly on the world; the human mind is mediated through tools (symbolic and/or concrete) and activities.

According to Lave and Wenger (1991) and Rogoff (2003), learning processes are situated in the context of specific cultures, situations, and activities that deeply influence them. Individuals learn through a process of active participation in sociocultural activities that are mediated by the most competent members of the community and its prevailing cultural and social values. The nature of the learner's participation gradually evolves from being relatively peripheral to appropriating the cultural activity (Lave & Wenger, 1991; Rogoff, 2003). As this process of appropriation progresses, they also redefine their membership in the community of learners (Lave & Wenger, 1991) and eventually acquire the capability of influencing the community's practice itself (Rogoff, 2003).

According to this perspective, culture plays an important role in learning. Also, learners and their mentors are interdependent partners who have active and dynamically changing roles in their social

group (Rogoff, 2003). An individual's participation in an activity (e.g., learning in school) is shaped by sociocultural processes, cultural tools (e.g., language), and interaction with others (Nasir & Hand, 2006). Students from one culture may learn Mathematics and English Language Arts using language, tools, and activities that differ from those used by students from another culture or even from different classrooms within the same school.

Physical activities and tools offer people different sets of opportunities to change the world, the circumstances under which they live, and the relationships with each other and themselves (Lantolf, 1996). Since activities and tools are influenced by the cultures in which they originate, individuals from different cultural backgrounds may not be able to make meaning from those tools and activities in the same ways.

Sociocultural perspectives provide a lens for examining how culture and social participation levels shape how students make meaning of items. As cultural products, tests assume students have certain levels of social participation in certain cultural settings. Different sets of cultural experiences influence how students gain access to items.

Sociolinguistics

Sociolinguistics—the study of language as a reflection of choice and social variation (Coulmas, 2013)—examines the relation between linguistic variables and social categories (e.g., race, sex, class), and how individuals co-construct their identities through the use of such variables during social interaction (Eckert, 2012).

Early work from sociolinguistics differentiated and defined cultures according to the characteristics of their language patterns. Current thinking in the field no longer treats languages, varieties of language, and language communities as bounded and uniform. Figure 2.1 models this gradual change in the thinking about language and language variation. Dots represent language features such as vocabulary, spelling, pronunciation, and discursive structure, among many others, and the frequencies of those features. In the past (upper portion of the figure), languages and language varieties were thought of as bounded and separate from each other. According to modern views (bottom part of the figure), language and language variation are dynamic and unbounded systems that share, borrow, transform, and reinforce each other's characteristics.

Dialects of a language vary according to the social characteristics of the language community, such as social class, geographic location, and age (Wolfram, Adger, & Christian, 1999). Everybody speaks a dialect. From a sociolinguistics perspective, no one dialect is better than another. Thus, while Ebonics or African American Vernacular English may differ from Standard English on pronunciation, syntax, lexicon, etc., both dialects are equally sophisticated rule-governed systems of conventions. What makes a dialect be perceived as better or more logical than another is social prestige and familiarity.

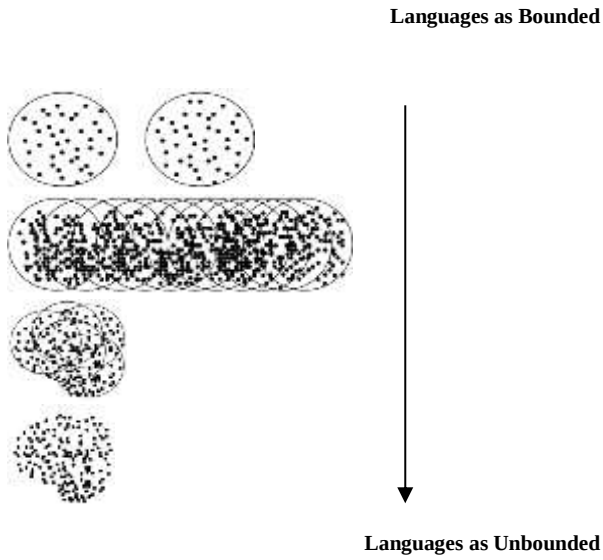

Figure 2.1. Evolution of the view of languages and language variation.

In addition to thinking about language variation due to dialect differences and the differences and commonalities between natural and academic languages, sociolinguistics raises awareness of the tremendous heterogeneity of linguistic groups of ELLs or speakers of non-standard forms of English. Taking this heterogeneity into account is critical to reasoning probabilistically about linguistic groups and the use of languages. Due to differences in their cultural experiences, each student has a unique set of strengths and weaknesses in English and in their first language. As a consequence, the effectiveness of item accessibility resources is, to a large extent, a function of how likely it is to function as intended for the majority of target students.

Social Semiotics

Social semiotics—the study of how meaning is conveyed and constructed through various forms of representation of information, and the ways in which this process is mediated by culture (Halliday, 1978)—allows examination of the features of items and of the features of item accessibility resources as alternative forms of representation of information.

Under this framework, culture comprises different semiotic resources that social groups use to interpret their experiences with the world and with each other. Pre-established broad semiotic resources such as language and art affect how individuals interpret the world and interact with each other. Yet members of a society are constantly modifying and changing their culture as they interact with other meaning making systems.

According to Halliday (1978), the structures of semiotic resources evolve as a result of the meaning making functions they serve within a culture. Each semiotic system has three main functions: ideational, interpersonal, and textual. In order to communicate successfully, members of a culture must have a shared understanding of the meaning making potential of the different semiotic resources that are used to create, maintain, and negotiate their reality—a reality that is socially constructed (Eco, 1984; O'Halloran, 2005).

Culture can be viewed in terms of the totality of its interrelated semiotic modes. Different semiotic modes communicate meaning in different ways (Lemke, 1998). For example, images employ spatial and simultaneous ordering principles to construct meaning, whereas meaning through oral or printed language is ascertained temporally and sequentially (Hull & Nelson, 2005). However, there is a synergistic relationship between the different modes in multimodal texts such that the modes work separately and integratively to construct meaning (Royce in Caple, 2008). On the other hand, multimodal texts also place constraints on a given individual's capacity to process information. Guichon and McLornan (2008) encourage designers to be cognizant of tasks that make students process multiple sources of information. However, while images can be used to support students' comprehension of the text of items, students also use the text of items to make sense of the images (Solano-Flores et al., in press).

In order to successfully interpret items and item accessibility resources, test takers need to be able to process and integrate different meaning making systems. To address the notion that multiple semiotic modes work together to communicate meaning, test developers need to take a comprehensive approach when designing items and item accessibility resources. This comprehensive approach takes into consideration the multiple features of items and their interaction. Furthermore, test items should be piloted with representative samples of different segments of the target population to ensure that students from different linguistic and cultural groups interpret the items and the item accessibility resources similarly and as intended by test developers.

Language as a Source of Measurement Error

The perspective of language as a source of measurement error establishes a link between language variation and error variance in testing. It combines the view of language as a process of communication mediated by context and the methods from generalizability theory—a psychometric theory of measurement error (Cronbach et al., 1972; Shavelson & Webb, 1991).

Assessment can be viewed as the process of communication between the student and an assessment system (Solano-Flores, 2008). Because assessment takes place through language, all assessments are, to some extent, measures of language proficiency. Accordingly, error in the measurement of academic achievement is, to a large extent, due to language variation (Solano-Flores & Li, 2013). Facets of this language variation include language proficiency and dialect. From this perspective, item accessibility resources are intended to minimize measurement error due to these sources.

Addressing language variation in assessment can be thought of as an effort to address the misalignment between the linguistic features of a test and the features of the variety of a language used by a given population (Solano-Flores, 2006). Figure 2.2 illustrates this misalignment. Dots represent language features, such as vocabulary, spelling, pronunciation, discursive structure (among many others), and the frequencies with which those features appear in a standard and a non-standard form of a language. The dots in the intersection represent the linguistic features of a test that are shared by users of a standard and a non-standard dialect. The greater this intersection, the fewer linguistic challenges are posed to users of the non-standard dialect.

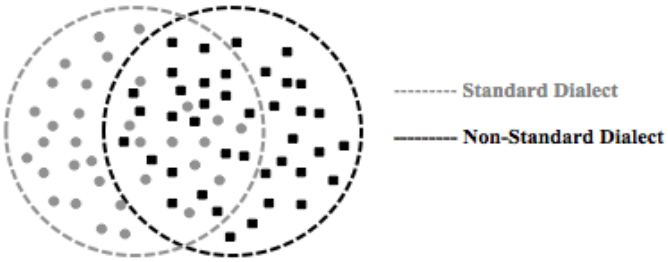

Figure 2.2. Linguistic (mis)alignment. Adapted from *Language, dialect, and register: Sociolinguistics and the estimation of measurement error in the testing of English language learners* (Solano-Flores, 2006).

When the process of test development only takes into consideration the population of students who are users of a standard form of a language, its linguistic features are not sensitive to non-standard forms of the language. The gray dots outside the intersection show the linguistic features of a test that are part of the standard dialect but are not shared with the non-standard dialect.

Among other strategies, alignment can be improved through item review and proper sampling. First, test development teams should consist of multidisciplinary groups (including, for example, translators, sociolinguists, content and measurement specialists, and bilingual teachers). These professionals should review and adapt items considering language variation across multiple linguistic groups throughout the entire process of test development. Second, items can be viewed as unintended samples of the linguistic features of a language or dialect. Due to the tremendous variation in English proficiency among ELLs, larger samples of items may be needed to test these students in order to properly minimize measurement error due to language. Third, representative samples of students from different linguistic cultural/ethnic, socioeconomic, and linguistic groups, including ELL students, should participate in the pilot stages of assessment development, so that test developers have the opportunity to refine the linguistic features of test items based on information obtained from these groups.

The view of language as a source of measurement error allows devising both judgment-based and empirically-based procedures for developing and evaluating items and item accessibility resources. Of special importance for culturally and linguistically diverse populations is the notion of item microanalysis, which examines how item meaning making can be influenced by the interaction of structural and pragmatic features of items and the characteristics of different cultural groups (Solano-Flores & Trumbull, 2003). Microanalytical approaches allow examination and refinement of both items and item accessibility resources.

Test Translation Error

The theory of test translation error—which examines how test translation is shaped mainly by the fact that languages encode experience differently—contributes to the notion of aggregated effect of test translation error. Multiple negligible translation errors, which individually would not affect student performance, may have a substantial, detrimental effect when they concur in an item (Solano-Flores, Backhoff, and Contreras-Niño, 2009). This knowledge underscores the importance of examining the features of item accessibility resources holistically.

Linguistic features of items are interconnected. Consequently, one test translation error can affect several translation error dimensions, such as those concerned with syntax, grammar, semantics, and

register. As a result of this interconnectedness, there is a tension between several different translation error dimensions.

Because languages encode meaning in different ways (Nettle & Romaine, 2002), translation error is inevitable. While test developers cannot totally eliminate translation error, they can minimize the frequency and severity of errors in translated items through a process of review. In this process, multiple users of the target language and professionals with different backgrounds examine the translated items and decide by consensus regarding their acceptability.

Solano-Flores, Backhoff, and Contreras-Niño (2009) represent translated items as entities (see the dots in Figure 2.3) within a probabilistic space. In this probabilistic space, there is an area of acceptability and an area of objectionability. A translated item is more likely to be objectionable when it has many mild translation errors, few severe translation errors, or many severe translation errors. Research findings indicate that highly objectionable translated items tend more than acceptable translated items to be responded to incorrectly by students. The most objectionable translated items are those with translation errors concerning semantics, register, and grammar and syntax.

The theory of test translation error enables test developers to view full-text translation and L1 glossaries from both a multidimensional and a probabilistic perspective. Optimal translation addresses the confluence of multiple possible errors along several dimensions. Also, optimal translation effectively takes into consideration the target language as it is used by its speakers.

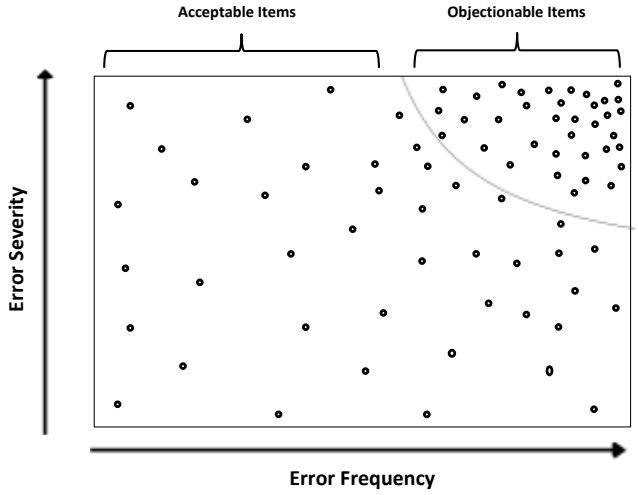

Figure 2.3. Areas of acceptability and objectionability of translated test items. Adapted from *Theory of test translation error* (Solano-Flores, Backhoff, & Contreras-Niño, 2009).

The multidimensional nature of translation error signals a need for professionals of multiple disciplines (e.g., translators, linguists, teachers, and content and assessment experts) to participate in the process of test translation and test translation review. This approach is in contrast with the traditional approach to test translation, which relies on the expertise of a limited number of translators.

Key Issues on Language Variation and Item Accessibility

According to the theoretical perspectives discussed above, how effectively language variation and item accessibility is addressed depends on the ability of test developers and test translators to: (1) consider the interaction of linguistic, cultural, social, and design factors in the development of item accessibility resources; (2) view language, language varieties, and linguistic groups as dynamic and fluid (rather than static), overlapping categories; and (3) develop tests and review their linguistic features as two interacting components of the same process.

Interaction of Linguistic, Cultural, Social, and Design Factors

Item language features do not act in isolation. Rather, item language features shape students' access to the content of items through their interaction with both features related and features not related to language. For example, issues about glossing (e.g., *What is the rationale justifying glossing for this constituent, not this other?*) are often related to features not directly related to the original text. The presence or absence of certain features seemingly unrelated to language (e.g., the visual display of the tools that students need to use to enter their responses or the characteristics of an illustration accompanying an item) may in fact be a reason in support or against glossing a given constituent or the way in which that constituent is glossed.

Because language variation is a reflection of cultural diversity, culture imposes subtle differences in the meanings and connotations of multiple constituents. If those differences are not properly taken into consideration, item accessibility resources may end up being ineffective. Glossing *vitamins* in English by defining the term as *substances that are good for the health* is potentially distracting for students from inner-city areas, in which *substance* is typically used to refer to an illegal drug. The ability of these students to properly interpret the word in context should not be underestimated. Still, the term has strong undesirable connotations and educators in inner-city environments would avoid it in their teaching and would certainly omit it in a test item.

Socio-economic differences are highly associated with certain linguistic groups. As a consequence, different linguistic groups may differ on the kinds of access they have to certain sets of experiences, opportunities, and resources. A case in point is the limited access that students from certain linguistic groups may have to computers, the limited amount of time they spend online, and the limited familiarity they have with computer tools and icons used in the context of formal instruction. Ensuring usability in the design of item accessibility resources addresses language variation not only through specific linguistic features, but also through proper consideration of socio-economic factors highly associated with language factors.

Owing to the interaction of linguistic, cultural, social, and design factors, test development and test translation should be viewed as intimately related activities, rather than separate stages in the process of assessment development. Ideally, tests should be developed taking into consideration, at all stages of the process, the fact that tests will be translated and adapted. Also, the process of development of item accessibility resources should inform the process of test development. Test developers and test translators need to work in combination if they are to effectively address language variation and item accessibility.

Language, Language Varieties, and Linguistic Groups as Dynamic Categories

Terms used to refer to natural languages (e.g., *English, Spanish*), varieties of languages (e.g., *Standard English, African American Vernacular English, academic language*) or linguistic groups (e.g., *English language learners, native users of Tagalog*) denote discrete categories. Yet the boundaries distinguishing them are not entirely clear. The ability to distinguish commonalities and

differences between these categories is critical to designing effective item accessibility resources. Two given natural languages may share multiple features. For example, English and Spanish have a common basic sentence structure (subject-verb-object).

Natural and academic languages (e.g., English and English academic language) are rather arbitrary categories that should be interpreted cautiously (see Aukerman, 2007). In this document, they are used to refer to the contexts in which language is used (e.g., in a classroom conversation, in a test on an academic subject), not to fixed properties. Whether a constituent is *natural* or *academic* is shaped by the context in which language is used. The higher frequency with which it is used in texts and speech in certain disciplines is what makes it "academic."

ELL students, as emergent bilinguals, are not totally unfamiliar with English. Indeed, many ELL students have a well-developed set of basic conversational skills in English. What makes them "English language learners" is not the incipient ability to communicate in English, but their incipient ability to communicate in English in academic contexts, and their incipient cultural experience through English or in cultural contexts in which English is used as the language for communication. Also, because they have multiple schooling histories and multiple forms of exposure to English, ELLs may vary tremendously in their English skills in different language modes (speaking, listening, reading, and writing).

Finally, regarding dialect, a given form of English is distinguished from another not only owing to a specific set of distinctive words, grammatical forms, and pronunciation features, but also because of differences in the frequencies with which certain common features are used in combination. For example, one of the two sentences shown below appeared in a Smarter Balanced Mathematics item; the other is a version of the same item created by educators with the intent to reflect the ways in which the students in their communities use English:

> *Drag the numbers to the boxes and the symbols to the circles to create an equation to show how much money she has left to spend.*
>
> *Move the numbers to the boxes and the symbols to the circles to make an equation that shows how much money she has left.*

While the two sentences can be understood by any English speaking student, each would be more likely to be said, written, or understood by users of a different English dialect. It is the combination of linguistic features, rather than the features themselves, that often appears to be critical to properly addressing language variation.

Because languages are vast domains and because each item has a unique set of linguistic and contextual features, it is difficult, in many cases, to anticipate which constituents should be glossed. Thus, information provided by educators on the ways in which language is used in their schools and communities is among the multiple sources of information that need to be used in combination to properly identify the constituents that need to be translated or adapted.

Test Development and Test Translation as a Continuum

Given the confluence of concepts and reasonings from multiple theories, the process of test development and the process of test translation need to be thought of as interacting, rather than separate stages. This is a critical condition that needs to be met if the intersection of language variation and item accessibility is to be properly addressed.

The implication of this view is that test development and test translation teams need to work in a coordinated manner throughout the process of test development and the process of the development of item accessibility resources. Both teams need to be multidisciplinary groups

comprised of members with diverse content knowledge, experience, and skills (e.g., content, measurement, and curricula specialists, sociolinguists, bilingual teachers, translators) and from diverse cultural and linguistic backgrounds.

3. BASIC CONCEPTS IN LANGUAGE VARIATION AND ITEM ACCESSIBILITY

As part of their attempts to meet the Common Core State Standards, assessment programs need to develop tasks of various levels of complexity and a wide variety of item formats (multiple-choice, open-ended, performance tasks, etc.). This wide variety of items is intended to tap into different kinds of knowledge and to contribute to producing better indicators of student academic achievement.

Since language is the medium through which tests are administered, the potential for better indicators of academic achievement comes with a high price tag—the increase of linguistic demands for test takers. Many items currently used in large-scale assessment programs are rich in text and contextual information (e.g., stories, fictitious characters) intended to make them meaningful. The linguistic features of these items (vocabulary, forms of speech, idiomatic expressions, discursive structure, etc.) may not necessarily be part of the knowledge and skills being assessed or are not equally familiar to all segments of the student population. Also, the situations used in the contextual information provided by these items may not be equally familiar to students from all cultural backgrounds. These challenges are serious for students from underrepresented groups, especially those who are developing English as their second language and those who are users of non-standard forms of English.

As an effort to ensure fair, valid testing for these students, the Smarter Balanced Assessment Consortium has devised a series of accessibility resources. These accessibility resources are tools for providing linguistic support to gain access to the content of items without giving their responses away (Measured Progress & National Center on Educational Outcomes, 2014; Measured Progress & Educational Testing Service, 2012). Delivered through the same platform used to administer Smarter Balanced tests online, these accessibility resources are intended to minimize the effect of multiple linguistic features that are not relevant to the knowledge and skills assessed. Ultimately, these accessibility resources are intended to contribute to more valid interpretations of Smarter Balanced test scores.

Smarter Balanced offers two broad types of resources intended to ensure item accessibility—stacked translations and glossaries. Stacked translations are full-text translations of items into the native, first language (L1) of English language learners. Glossaries are lists of synonyms, definitions, or paraphrases of words, terms, idiomatic expressions, etc., that are not part of the specialized language of the content area assessed.

This conceptual framework is intended to provide test developers and test translators with reasoning and strategies for addressing language variation in ways that ensure item accessibility for all students. More specifically, the conceptual framework is intended to support test developers and test translators in their efforts to develop effective item accessibility resources.

This chapter defines basic concepts that are used throughout the conceptual framework or that provide support for the ideas presented in the document. Since this conceptual framework addresses the intersection of language variation and item accessibility, the basic concepts presented are grouped according to those broad areas.

Language Variation

The term, *language variation* is used in three main ways, to refer to: (1) different levels and forms of proficiency among users of English as a second language, (2) dialect variation within a given language, and (3) the set of differences and commonalities between natural and academic language. These forms of language variation are discussed in this section. Also, the section

discusses the concepts of translation, constituent, and glossary, which are critical to reasoning about item accessibility and language variation.

English Language Learners and English Language Proficiency

English language learners (ELLs) are students who are not entirely proficient in English—the language of testing (Abedi, 2008). Also called *emergent bilinguals*, ELLs can be defined as students who are developing English as a second language while they continue developing their first language (Garcia & Kleifgen, 2010).

The alternative term, *emergent bilingual*, calls for attention to the fact that limited proficiency in English is not an indication of a deficit, but the result of a natural process in which two languages are being developed. Indeed, the linguistic skills an ELL has in their two languages, taken together into consideration, can exceed the linguistic skills of a monolingual student of the same age (Oller, Pearson, & Cobo-Lewis, 2007).

English language learners are often grouped into one large subgroup and treated as a homogeneous population. However, ELLs are a heterogeneous group of second language learners comprising students with different native languages, different patterns of English proficiency, and different background knowledge and experiences (Valdés & Figueroa, 1994). Issues of fairness and validity are especially serious in the testing of ELLs because, in addition to having to demonstrate their content knowledge in a language that they are still developing (Escamilla, 2000; Hakuta & Beatty, 2000; Kopriva, 2008), definitions of "English language learner" are inconsistent across states (Linquanti & Cook, 2013).

Dialect

Dialects are mutually intelligible varieties of language (as in *Standard English*, *Southern U.S. English*, and *African American Vernacular English*) that differ from each other on features such as pronunciation and the frequency of use of certain forms of speech, vocabulary, etc. While the term, *dialect* is sometimes used in a derogatory form to refer to a variety of a language, everybody uses dialects. Even the most prestigious form of a language (e.g., Standard English) is a dialect. However, some dialects are more prestigious than others (Wolfran, Adger, & Christian, 1999) because dialects are the result of differences in factors such as socio-economic status, geographical region, and ethnicity.

Contrary to common misconceptions, dialects are not corrupted or low forms of a language. Research in sociolinguistics has shown that all the dialects of a given language have comparable levels of sophistication and are equally governed by rules and conventions (Wardaugh, 2002).

Register, Academic Language, and Natural Language

A *register* is a form of language associated with different social practices and the people who engage in such practices (Agha, 2003; Halliday, 1978). Each language has multiple registers, but not all speakers of a language are familiar with all of the registers. For instance, within English, football players, lawyers, or truck drivers use their own registers to refer in a meaningful way to concepts and issues that are specific to their activities.

Registers develop and evolve as a consequence of specialization. As a consequence, disciplines such as Mathematics, English Language Arts, and other disciplines have their own registers. *Academic language* is the term used to refer to the register used in textbooks and by teachers in promoting conceptual understanding of a subject within the academic context of a discipline (Butler et al., 2004; Scarcella, 2003; Schleppegrell, 2004). While academic language is usually associated

with technical vocabulary, it also comprises other aspects of language such as grammatical forms or ways of asking problems, arguing, expressing disagreement and socializing through language that are more frequent in the context of a discipline than in any other context.

For instance, the combination of the interrogative and the conditional mood in a sentence, as in

How many flowers can Jane buy with $7.50 if each flower costs 76 cents?

is a form of posing problems frequently used in Mathematics textbooks and tests.

In the context of education, the term *natural language* is commonly used to refer to the everyday, non-specialized, or colloquial language. For example, natural English is used to refer to the form of English used by persons in multiple contexts.

The distinction between natural language and academic language, without considering the context in which social interaction takes place is, in many cases, a matter of judgment (Aukerman, 2007). While it is easy to view *angiosperm* as part of the academic language used by botanists or in textbooks or science units on plants, a word like *unit* may be more difficult to characterize because it belongs to the language used both in academic and non-academic contexts with the same meaning (see Wellington & Osborne, 2001). Furthermore, phrases such as *on the other hand*, *nonetheless*, and *in contrast* may be even more difficult to characterize because they have the same meaning, but are used more frequently in formal, academic contexts than non-academic contexts.

As with other forms of language variety, an effective way of reasoning about the differences between natural and academic language consists of adopting a probabilistic perspective (Solano-Flores, in press). This perspective takes into consideration that, in addition to the meaning they are intended to convey, certain words, terms, expressions, discursive forms, etc., appear more frequently or are more likely to be used in academic than non-academic contexts.

One of major concerns regarding fairness in the testing of ELLs is that natural and academic language do not develop at the same pace (Cummins, 1981), which limits the opportunities for these students to benefit from instruction (Lee, 2005). Thus, reasoning critically about the nuances of the concepts of academic language and natural language helps test developers to address the fact that meaning is shaped by context. This reasoning contributes to enrich the process of development of item accessibility resources.

Translation

Translation can be defined as the activity (and the product of that activity) intended to communicate meaning, mainly in the form of text, in a language that is not the language in which the text was originally created. Effective translation not only considers the characteristics of the languages but also the characteristics of the users of that language, which may vary across social contexts.

Current thinking in the field of translation holds that perfect translation is not possible because languages are cultural products which evolved in different societies over time and under different sets of circumstances (Greenfield, 1997: Nettle & Romaine, 2002). Therefore, languages encode different sets of experiences and meet different sets of needs.

The impossibility of perfect translation is relevant to taking appropriate actions to ensure test equivalence across languages. Translation error is inherent to translation; it can be minimized but not eliminated. Accordingly, the quality of the translation of a test should be judged based on disconfirming evidence as well as on confirming evidence that a translation is adequate (Solano-Flores, Backhoff, & Contreras-Niño, 2009).

While translation usually evokes full text translation, this is but one of the many forms in which translation can be offered to support students who are not yet proficient in the language of testing. An alternative form of translation consists of making available to students glossaries of specific terms identified as likely to pose a challenge for them to gain access to an item.

The analysis of the different forms of translations and translation procedures are beyond the scope of this conceptual framework. However, they are discussed in a related document, developed with the intent to support the actions of test translators who develop translations for Smarter Balanced Assessments (Solano-Flores, 2012).

Constituent

The term, *constituent* is used in this conceptual framework to refer to a string of words in an item. Thus, in the sentence,

Put all your money in a big brown bag[1]

brown, *bag*, and *brown bag* are three different constituents.

Reasoning about constituents allows test developers and test translators to focus on the functions of sets of words used in combination. An expression such as, *a jack of all trades* needs to be interpreted as a whole in order to be able to make sense of it. Accordingly, in the sentence

Rainy days and Mondays always get me down[2]

the three words in *get me down* act as a whole to convey meaning about the emotional impact of an event. The translation of the three words as a block helps students to make sense of the entire sentence more effectively than a translation of each word separately.

A constituent is, to some extent, the unit based on which item accessibility resources can be designed. The concept allows identification of the most important constituents to gloss in a given item.

Glossary

The term, *glossary* is used in this conceptual framework to refer to a device which offers an alternative textual representation of a constituent identified as likely to pose a challenge for students to gain access to an item. This alternative textual representation may consist of a definition, a paraphrase, a synonym, or a list of synonyms in the same language in which the glossed constituent appears in the text of the item or, in the case of ELL students, in their L1.

While glossaries typically appear at the end of a document, the technology for computer-based testing makes it possible to make glossaries available *in situ*, next to the glossed constituents, in the form of a boxes that appear on the computer screen next to them when the student clicks on it.

Proper consideration of the three facets of language variation (variation due to language proficiency, to different dialects within a language, and to the commonalities and differences between natural and academic language) is critical to effective glossing. Careful consideration of these facets helps test developers to decide which constituents in an item are to be glossed and how they have to be glossed.

Item Accessibility

The term, *item accessibility* is defined here as the condition in which a test is administered to minimize or eliminate factors irrelevant to the construct assessed, to support students in

demonstrating their knowledge and skills. This section discusses several concepts relevant to addressing item accessibility: construct, semiotic modality, cognitive load, design, affordance, visibility, and usability. After discussing these concepts, a formal definition of item accessibility resource is given. Then, the section ends with the definition of design and user interface, concepts that are critical to the development of high-quality item accessibility resources.

Construct

Construct is the term commonly used in the field of educational measurement to refer to the knowledge or skill a test is intended to measure. A construct is not visible; it is an abstraction about the nature and organization of that knowledge or skill (Cronbach, 1990). Verbal fluency, number sense, and reading comprehension are examples of constructs. The notion of construct is critical to validity and fairness in testing. No serious claims can be made about the validity of interpretations of the scores of a test if the performance of students is influenced substantially by factors that are unrelated to the knowledge and skills that the test is intended to measure (Messick, 1989). Those extraneous factors are commonly referred to as *construct-irrelevant* factors.

Proficiency in the language in which a test is administered is one of the most serious threats to the validity and fairness of a test (AERA/APA/NCME, 1999). Ultimately, efforts to ensure item accessibility are aimed at minimizing construct-irrelevant language factors.

Semiotic Mode

Semiotic mode is a term used to refer to any resource used to represent information and convey meaning (Kress & Van Leewen, 2001). Although *semiotic* evokes images and signs, the modern use of the term refers to the meaning conveyed through text, images, genre, and even the medium (e.g. printed text, e-mail) through which information is represented.

A verbal description of a process, a formula describing that process, the graph of that process, and a gesture made with the hand to show the shape of the line in that graph are different representations of the same thing using different sets of semiotic modes. Each form of representation uses a set of conventions for representing information and assumes that the recipient of that information is familiar with those conventions (Eco, 1984; Iedema; 2003). Also, each form of representation has a unique set of advantages and disadvantages concerning accuracy and efficiency (Hull & Nelson, 2005).

While text and images can be thought of as broad semiotic modes, any resource for representing information can be a semiotic mode, depending on the context in which communication takes place (van Lier, 2004). A photograph, a cartoon, a silhouette, and a line drawing of the same object can be thought of as different semiotic modalities because they convey different meanings, emphasize different aspects of the object, and lend themselves to different sets of interpretations. Moreover, even a specific font or typeface may become a semiotic mode when there is meaning associated with it, as shown by these three versions of the same sentence, intended to communicate different meaning:

But I must be moving on[3]

*But I **MUST** be moving on*

But I "must" be moving on

Implicit or explicit social conventions shape how meaning is created through the combination of multiple semiotic modes (Caple, 2008). For example, how effectively a formula represents information depends on the proper use of multiple conventions concerning the position of the

decimal point in numbers, the relative position of the numbers and symbols, and the alignment of signs with respect to those numbers, among many other features.

While the representation of information through multiple broad semiotic modalities (e.g., textual and visual) supports understanding (Guichon & McLornan, 2008), it also imposes additional cognitive demands. The brain processes the information provided in the presentation formats separately and then integrates it (Mayer, 2001). In order to avoid a cognitive overload, especially for ELL students, the use of text and images in combination should be carefully planned (Schnotz, 2005).

Since languages (such as English) are convention-governed systems for communication, a text in English, the translation of that text, and the way in which the translation of a constituent is shown (e.g., as a "pop-up" text, to the right of or below the constituent, with the same or different font as in the constituent in the original text) can be viewed as ways of representing the same information using different sets of semiotic modes.

Cognitive Load

Cognitive load is the amount of information that a person's working memory needs to process simultaneously before being able to make sense of it and complete it. The brain is limited in its capacity to efficiently handle multiple pieces of information. Cognitive overload takes place when those limits are reached (Sweller, Van Merriënboer, & Paas, 1998).

Cognitive overload can take place when, in addition to processing the information needed to make sense of an item, a student needs to interpret the display of the information on the screen of the computer and understand how their responses need to be entered. For ELL students and students who are users of non-standard forms of English, these sources of cognitive overload add to those that stem from having to interpret text in a language with which they are not as familiar as English native speakers.

Cognitive overload can potentially affect the ability of a student to perform effectively on a test. Hence the importance of optimizing the design of user interfaces and to provide appropriate support through translations and glossaries.

Affordance

In general, *affordances* are possible actions that persons perceive they can take with an object (Gibson, 1977). The concept of affordance applies to both real and virtual objects. Thus, a plate and a handle on a door invite the user to push and pull the door to open it, respectively. Likewise, an "OK" button and a "Cancel" button next to each other on the screen of a computer invite the user to make a decision by clicking on one of the buttons.

Affordances are not fixed characteristics of an object, regardless of their potential users. Rather, they are the result of the interaction of the characteristics of objects and each individual's personal experience. This experience is shaped by culture. Depending on an individual's experience, different combinations of semiotic modes (as in the three versions of the sentence, *I must be moving on,* shown above) may or may not be effective in communicating the desired meanings. Likewise, how to use the tools provided in a computer-administered test to enter the answers to items may or may not be obvious depending on the student's personal experience.

When real or virtual objects are designed without properly taking cultural diversity into account, the characteristics that may be affordances for some individuals may end up being challenges for others. Thus, as part of the activities for their design, computer-administered tests, translations, and

glossaries should be attempted with representative samples of different linguistic and cultural groups.

Visibility

Visibility is the property of an object that makes it easy to use and understand (Norman, 1988). Visibility signifies the proper mappings between the user's intended actions and what appears to be possible (Norman, 1988). For example, a stove usually has four burners that are controlled by four separate knobs. There is a one-to-one correspondence between the different controls and their functions. Furthermore, the knobs are usually labeled in some fashion indicating which knob controls which burner (Figure 3.1).

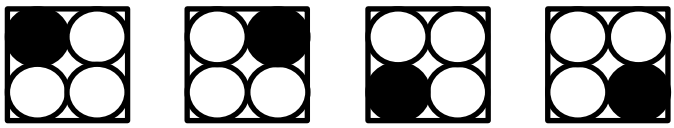

Figure 3.1. Visibility and mapping in the control labels of a stove with four burners. (Adapted from Norman, 1988).

The design of objects should make it easy for the user to see what actions can be taken by making functions visible, using correct mapping, showing and/or limiting alternatives, and providing feedback (Lidwell, Holden, & Butler, 2003).

In the context of testing, there should be a visible and appropriate mapping between the visible features of tools, what the test takers infer to do with these tools. For example, if students see a tool that mimics the appearance of a calculator, then the tool should have the function of a calculator. Furthermore, each control (e.g., number or operation key) should only have one function.

Each item accessibility resource should be visible and distinct from other resources. For instance, there should be a way in which students can visually distinguish, before engaging in reading, between the text of an item provided in English and the text provided in L1 in a stacked translation.

Usability

Usability is the ease with which an object can be used, given the function it is intended to serve and the intentions of the user (Norman, 1988). A critical aspect of usability is the amount of time needed to learn to use that object through the interaction with it—how intuitively it can be operated or how its features are self-explanatory. The use of an object's affordances and natural mappings decreases the amount of time users need to learn the functions of an object.

The concept of usability is especially important in the context of computer-administered testing, in which students interact with tools made available to them to obtain information and enter their responses. If students need to spend considerable time trying to understand how to perform those actions, or if it is necessary to invest too much time training them to perform those actions, probably the usability of those tools is poor.

Item Accessibility Resource

Technically, in terms of the concepts discussed in this section, an *item accessibility resource* can be defined as an alternative form of representation of the information contained in a constituent

through a selected set of semiotic modalities with the intent to ensure that students make sense of an item as intended.

Specifically with regards to glossaries, the effectiveness of these item accessibility resources depends not only on the accuracy with which constituents are translated, defined, or rephrased. It also depends on the ways in which multiple semiotic modalities are used to represent the information and on the extent to which the use of those semiotic modalities constitute affordances—rather than challenges—for the majority of the students and do not excessively increase the cognitive load of a task.

Design

The term, *design* is used here to refer to the series of activities oriented to determining, in ways that are systematic and scientifically defensible, the optimum arrangement of characteristics of items and item accessibility resources.

User Interface

The term, *user interface* refers to the space or environment in which the student interacts with a computer. An example of an interface is the display on the computer screen showing a pad on which the student clicks to enter a number in response to a problem and the box showing the numbers entered. Another example of an interface is the highlighting of a word used with the intent to indicate that a glossary is available and the box containing the glossary that appears when the student clicks on the word.

4. PRINCIPLES FOR THE DESIGN OF ITEM ACCESSIBILITY RESOURCES

The last decades have witnessed a significant progress in the field of design. Mainly informed by the cognitive sciences, this field has generated important ideas for designing objects that enable their users to use them properly and with ease. Some objects may be simple and may not have changed substantially through history. These simple objects tend to serve a limited range of functions and their shape makes evident to the user how they are to be used. For example, the angle and form of the handle of an iron invites to hold and push down, whereas the angle and form of the handle of a kettle invites to lift, hold, and pour.

Other objects are considerably more complex. They appeared very recently in human history and keep evolving as new technologies arise. These complex objects tend to serve a wide range of functions and their appearances do not make evident to the user how they are to be used. In addition, these complex objects interact with the user; the actions the user takes are based on the information the objects provide. An example of this type of object is the smartphone. The device can perform multiple functions and the user cannot tell from its shape how to operate it. Successfully and efficiently placing calls, taking pictures, sending e-mails, checking the weather, and watching a video, among many other functions, depends on how easy it is for the user to interact with the device (Saffer, 2014)—to both interpret the information displayed on the screen and perform a relatively small set of actions, such as clicking, scrolling, and signaling.

From a design perspective, items administered by computer belong to the category of complex objects. Properly designing their features (directions, wording, illustrations, buttons, cascade menus, dragging options, physical arrangement of text, and shape and size of response boxes, among many others) is critical to minimizing cognitive load—the working memory the user needs to process the information provided by the object.

The cognitive load imposed by an item should be mainly related to the knowledge and skills targeted by the item, not related to figuring out what the item is about or how to enter an answer. This principle, which is applicable to any student, is especially critical to fairly and validly testing English language learners (ELLs) and, in general, students from ethnic/cultural minorities and low socioeconomic groups. These students are likely to have less access to computers than mainstream students (e.g., they should not be assumed to have a computer at home) and less experience taking tests online.

This chapter discusses design principles that are critical to properly addressing language variation and increasing the accessibility of Smarter Balanced assessment items. The chapter also describes tools that test developers, test translators, and assessment systems can use to support the implementation of those strategies throughout the entire process of test development. Eight aspects of design are considered: consistency, symmetry, contrast, meaningfulness, constraints, customization and design from scratch, pragmatic suitability, and standardization.

Consistency

Consistency is the condition in which elements (e.g., components of a user interface or text constituents) with similar functions and meanings have also similar appearances across the items of a test. It also refers to the condition in which the appearance of these elements is in accord with the students' previous experience.

User Interface

A good user interface contains features that are simple and work together such that there is a correspondence between what users need to do and what appears to be possible to do (Norman,

1988). The elements of a user interface should be similar to elements with similar functions with which the user has prior experience interacting. Likewise, elements with different functions should have different appearances.

Figures 4.1 and 4.2 show two versions of an interface of a Mathematics item which asks the student to enter a number. To many students, the interface shown in Figure 4.1 may look very much like a calculator. That is because humans interpret and understand new experiences based on mental representations of previous experiences. Because of this tendency, students (especially those who could have difficulty understanding the directions due to limited proficiency in English) may believe that the task involves performing a calculation.

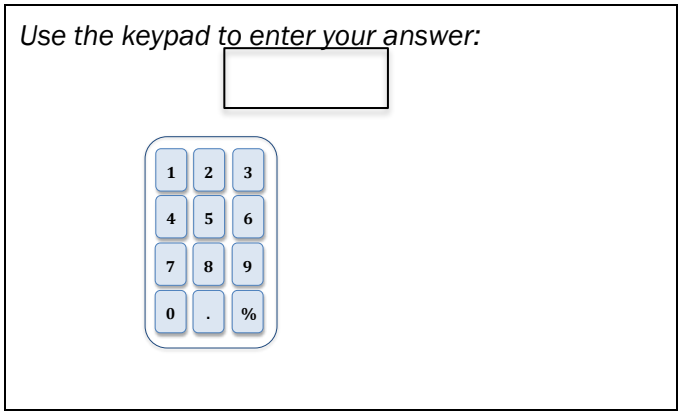

Figure 4.1. An interface for entering a numeric response.

The interface shown in Figure 4.2 has an alternative design in which the number keys are arranged in a reversed order—from the bottom to the top, which is the same arrangement used in the number keypads of computers. The alternative design also displays the window that shows the student's response to the right of the keypad.

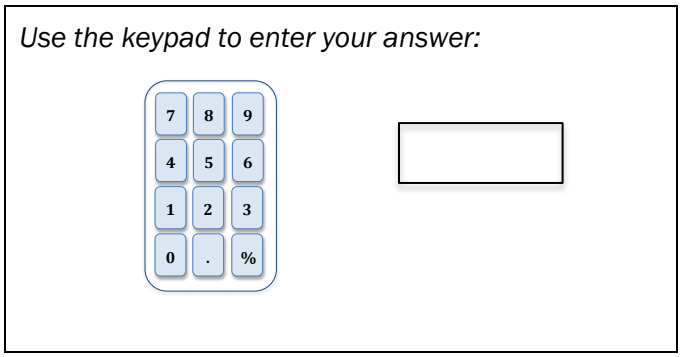

Figure 4.2. An alternative interface for entering a numeric response.

Test developers and test translators can benefit from creating a *book of interface features*. This is a document that establishes the characteristics that the user interfaces should have for the different

actions students need to take as part of their interactions with computers in computer-administered tests. For example, the interface for entering numbers using a keypad, the interface for constructing a graph, the interface for arranging sentences in a logical sequence, the interface for entering open-ended responses, etc., should always be the same across items.

Needless to say, the book of interface features needs to be developed with the participation of multiple professionals. Also, during their development, the interfaces need to be tried out with students from diverse linguistic and cultural groups. The reason is that, in part due to linguistic and cultural influences, individuals vary tremendously in the ways they make sense of their experiences. As a result, not all of them may interpret items and interfaces in the ways test developers intend.

Text

When the same information has to appear repeatedly in the form of text, the constituents should be the same. Sizable amounts of text are devoted in tests to provide *operational information*—directions for students on the actions they are expected to take to complete the items and to provide their responses. This is especially the case with computer-based testing, in which directions are provided to students with the intent to ensure they interact properly with the computer. The operational information on actions of the same kind should always be presented in the same way across items.

Suppose that there are several items in the same assessment, each with different wording for the same action:

> *Enter your response in the box below.*
>
> *Use the box below to enter your response.*
>
> *In the box below, type your response.*

Only one of these versions of the same set of directions should be used across items. Additionally, the glossing and translation of these directions should always be consistent across items. Ensuring consistency in the ways in which directions are provided to students contributes to maximizing item usability. To ensure consistency, test developers need to identify the types of actions students are asked to take and the different ways in which they need to enter their responses across all the items in an assessment.

A *book of wording forms,* like the one illustrated in Table 4.1, can be created to establish the exact ways in which different forms of operational information should be worded. This document allows standardization in the process of development of items and maximizes usability. Needless to say, the table lists only a few of the multiple directions that a test may need to provide to test takers.

Table 4.1

Fragment of a Hypothetical Book of Wording Forms: English and Spanish

Function	Wording Form	
	English	Spanish
Type in an answer	*Enter your answer in the response box.*	*Teclea tu respuesta dentro del cuadro de respuestas.*
Multiple choice (mutually exclusive)	*Select the correct answer.*	*Selecciona la respuesta correcta.*
Select the correct answer(s) (non-mutually exclusive)	*Click in the box to select the correct answer. You may click more than one box.*	*Haz clic en el cuadro de la respuesta correcta. Puedes hacer clic en más de un cuadro.*
Transitional words for multi-step directions	*First,.. Second,... Third,... Then,... Next,...*	*En primer lugar,... En segundo lugar,... En tercer lugar,... Entonces,... A continuación,...*

Symmetry

Symmetry is the condition in which entities with the same importance receive the same level of consideration and emphasis. An important set of entities in the context of language variation and item accessibility comprises English and L1—ELL students' first language.

Symmetry is especially important in the case of stacked translations. Differences in language spread, political power, social status, and even the availability of professionals with expertise in different areas related to linguistic diversity in testing may unintentionally produce a double standard in the rigor with which languages are treated in a test. For example, the time frame for review and piloting with sampled students may be proportionally tighter and less realistic for test translation than the development of items in English.

This difference produces and reflects an asymmetrical relationship between the two languages. When languages are treated asymmetrically, one language receives more attention over the other. This asymmetry may affect the effectiveness of the accessibility resource and, ultimately, the validity of a test.

An example of *asymmetry* is the case in which the same amount of space is assigned to the English and L1 versions of an item. While, in appearance this strategy is fair, it does not take into account the fact that languages vary considerably on the amount of screen (or paper) area they take to represent the same information. For example, the Spanish translation of a text in English takes about 20 percent more characters. These differences stem from the fact that languages may vary considerably in their properties, such as the average number of characters in a word, the average number of words in a sentence, the frequency or availability of acronyms, and the level of explicitness with which information needs to be provided (see Smith, 2012). Also, different writing

systems (e.g., alphabetic, logographic, syllabic) impose different minimum sizes needed to be able to read text with ease. These differences are particularly important in the testing of young students. If they are not properly taken into account, accommodating the text of translated items in the user interface may compromise the integrity of the format of the L1 version (e.g., by reducing its font size).

Symmetry issues are difficult to address without careful planning. Critical to effective planning is the interaction between test developers, test translators, and professionals of different backgrounds throughout the entire process of assessment development.

Contrast

Contrast is the condition in which elements with different functions are represented differently. It helps users to notice important differences between elements.

Highlighting

Highlighting is a way of producing contrast in text through the use of bold letters, italics, underlining, typeface (e.g., Times New Roman, Garamond), and color (Lidwell, Holden, & Butler, 2003). In assessment, highlighting can be used to denote different forms of information provided in the text of items. For example, highlighting can be used to differentiate text that provides operational information from text that is specific to the items of a test.

Highlighting can also be used to denote constituents deemed critical to properly understanding a statement and which may require students to pay special attention. Such is the case of terms that qualify or modify meaning (e.g., *sometimes, must*), add precision to information (e.g., *mostly, almost*), or link and compare ideas (e.g., *however, unlike*).

According to a common design practice, it is safe to say that no more than ten percent of the text of an item should be highlighted. If highlighting is too frequently used, it loses its value as an information device. The use of different fonts is not recommended for highlighting because their differences are difficult to appreciate.

Coloring should be used sparingly in text. When used, clearly distinct colors should be used and those colors should not disadvantage color-blind students. Nor should the correctness of the students' responses depend on the ability to distinguish color tone differences. An additional consideration that speaks to the caution with which highlighting must be used is the fidelity with which computers show the color tones intended by the test developers. Even computers of the same brand and model may vary in the fidelity with which they show different color tones.

A *book of highlighting codes*, like the one shown in Table 4.2 norms the use of highlighting codes. Notice that, in addition to emphasizing contrast between types of constituents of different functions, the book ensures consistency in the use of different highlighting codes across the items of a test. Needless to say, the table shows only a few of the many possible ways in which highlighting can be used systematically to support students.

Table 4.2

Fragment of a Hypothetical Book of Highlighting Codes

Highlighting Code	Use it to…	Example
Bold	ensure clarity	Find **another** fraction equal to 4/12.
Italicize	provide operational information	*Enter* your answer in the box below.
Underline	denote constituents in passages that are referred to in the items	Uncle John was looking forward to seeing his friend, after so many years.

Contrasting in Stacked Translations

Stacked translations are a case of accessibility resources that deserves special consideration concerning contrast. Contrary to what intuition dictates, the different appearances of two language versions of the same item may not be obvious at first glance (Figure 4.3). This is especially the case for students who lack a strong set of metalinguistic or reading skills which enable them to identify with ease the point at which the L1 version of an item ends and the English version of the same item starts.

Using different colors and typefaces is not recommended for contrasting different language versions of items; the resulting visibility differences may favor one language version over the other, which violates the principle of symmetry, discussed above. The adequacy of different forms of emphasis (e.g., bold, italics, underline) to contrast two language versions of items is not recommended either, as they are already well established as semiotic resources to convey certain meanings (e.g., relevance, authorship).

Figures 4.4 and 4.5 show simple and potentially effective contrasting approaches. The dotted line and the box are intended to support the user to distinguish at a glance when one language version ends and the other starts. Needless to say, the effectiveness of these contrasting approaches should be tested empirically by observing and interviewing users.

¿Qué número es igual a 10^4?

Which number is equal to 10^4?

Figure 4.3. An item and its stacked translation.

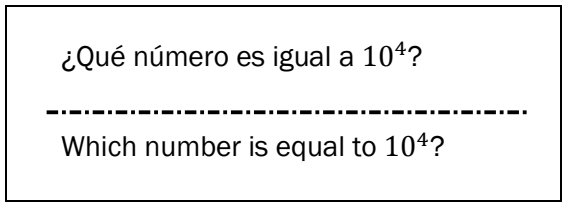

Figure 4.4. Contrasting the Spanish version and the English version of the same item.

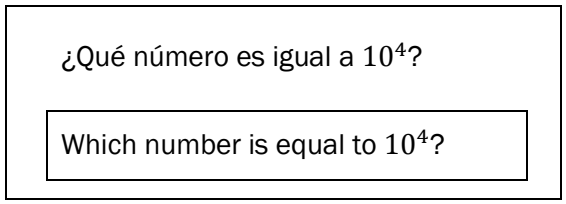

Figure 4.5. A second approach for contrasting the Spanish version and the English version of the same item.

Meaningfulness

Meaningfulness is the condition in which students are likely to relate the content of test items and the contextual information they provide to their personal experience. Meaningfulness is important in the current assessment scene, in which context-rich items are used in alignment with new standards and testing practices that require situating problems and tasks in meaningful contexts. Thus, many English Language Arts test items ask questions about passages provided as stimulus materials, and Mathematics items present short stories and concrete situations to frame problems.

Culture and Meaningfulness

A challenge for proper design of context-rich items is the selection of topics and contexts that are equally meaningful to all the students. Literature on issues of testing and fairness has documented concerns that the contexts used in many items privilege mainstream, upper- and middle class students because they tend to reflect their views, values, and life styles (Solano-Flores & Li, 2009).

Suppose that, as part of a reading comprehension task in an English Language Arts assessment, students have to respond to a set of items after reading a passage. The passage is the story of a girl who finds that a couple of birds have put their nest in the holiday wreath hanging on the door at her house's porch. This story is more likely to be meaningful to white, Christian, upper-class, suburban students than any other group of students simply because it contains elements that are more familiar to their everyday lives and to which they can easily relate their own experiences. The cognitive load of the items asking questions on the story can be considerably higher for students who are not as familiar with Christmas and holiday wreaths and with living in houses with porches. As a result of these differences, these students need to pay extra attention not only to answer the questions, but also to make sense of the information provided by the story.

Of course, just because an item uses a context that depicts the life of a given cultural/ethnic or socioeconomic group, does not necessarily mean that the item is not accessible to all students. However, issues of fairness arise when a substantial proportion of the items in a test reflect the lifestyle, values, and traditions of a single cultural/ethnic or socio-economic group.

Equitable Meaningfulness

Information on how students reason when they respond to an item and how this reasoning is influenced by culture is seldom available for many items. The reason is that this information is expensive to collect, as it has to be obtained through cognitive interviews and focus groups (Ercikan, 2002). Even when empirical procedures exist for examining the differential functioning of items across main linguistic or cultural groups, these procedures cannot always be used with all the items in a test and during the test development stage (Allalouf, 2003).

In the absence of empirical data, test developers and test translators need to use formal strategies to judge equitable meaningfulness in a test. *Equitable meaningfulness* is the extent to which the contextual information used in the items of a test portrays or reflects a wide variety of contexts and situations to which students from multiple ethnic/cultural and socioeconomic backgrounds can relate.

A sampling perspective enables test developers and test translators to ensure equitable meaningfulness. Context-rich items necessarily reflect sociocultural activity. Thus, rather than attempting to create context-rich but culture-free items, cultural aspects need to be addressed by ensuring that, overall, different types of contexts and different cultural/ethnic or socioeconomic groups are represented across all the context-rich items of an assessment.

Table 4.3 provides a template of what can be called, *equitable meaningfulness analysis*. It presents a simple classification of topics or situations reflected or implied by items. Needless to say, this is an illustration; other classifications can be produced. In the example, two factors are considered: social context and equitable meaningfulness. The former refers to whether the topic or situation presented or implied by an item is representative of a rural, urban, or suburban (or indistinct or undefined) context. The latter refers to how likely the topic or situation presented or implied by the item reflects the everyday lives and views of students. Three meaningfulness categories are identified. Categories A and B describe topics or situations that are accessible to many students. Category C represents topics or situations that are likely to privilege a given cultural/ethnic or socioeconomic group. Topics or situations within category C favor specific groups and marginalize students from other groups.

Table 4.3

Equitable Meaningful Analysis: Topics and Situations Used in the Context-Rich Items of a Test by Social Context and Meaningfulness Likelihood

	Equitable Meaningfulness		
	Likely		Unlikely
	A.	B.	C.
Social Context (Locale)	The topic or situation reflects the everyday lives of many students, regardless of cultural/ethnic or socioeconomic group	The topic or situation reflects the everyday lives of individuals from a specific cultural/ethnic or socioeconomic group, but is familiar to individuals from multiple cultural/ethnic or socioeconomic groups	The topic or situation reflects the everyday lives of individuals from a specific cultural/ethnic or socioeconomic group, and is unfamiliar to individuals from multiple cultural/ethnic or socioeconomic groups
Rural	Acceptable	Acceptable	Objectionable
Urban	Acceptable	Acceptable	Objectionable
Suburban	Acceptable	Acceptable	Objectionable
Indistinct or undefined	Acceptable	Acceptable	Objectionable

Ideally, all the items in an assessment which use contextual information should be distributed proportionally across the cells in Columns A and B. No context-rich item should appear in Column C.

Needless to say, a team of test developers and translators need to determine whether a topic or situation is classified as belonging to any of the three categories of item meaningfulness. The reason is that individual decisions may be based on inaccurate perceptions of what is or is not meaningful to students of a given cultural/ethnic or socioeconomic group.

Design Constraints

Design constraints are limits imposed on certain parameters that cannot be exceeded in the design of items or their L1 translations. These constraints are intended to minimize unnecessary variation in the linguistic features of items.

Text Length

Depending on school grade, content area, and the knowledge and skills being assessed, test developers and test translators should establish a range in the amount of text used in items. The reason is twofold. First, there is a minimum amount of text needed to word an item or a passage in a way that is accessible to students. Second, there is a maximum amount of text beyond which the validity of items can be threatened due to excessive reading demands. While excessive text length is always an issue in assessment development, it is a special source of concern in the testing of English language learners.

A possible criterion for establishing text length specifications is based on the total number of words contained in the item or on the amount of time that the average student takes to read the item aloud at a normal, reasonable pace. These specifications may be different for Mathematics and English Language Arts items, even within the same given grade, and for items and passages used as stimulus material for English Language Arts items. They should be established by a team of developers with expertise from different fields (for example, in the case of English Language Arts, teachers of this subject, reading specialists, assessment developers, and professionals with expertise in biliteracy).

Grammatical Features

Linguistic features other than text length can be included as part of the set of constraints for the design of items and their stacked translations. These features may include, among many others: double negation, passive voice, nominalization, embedded clauses, and the combination of conditional and interrogative sentence moods. Also, the set of constraints for grammatical features may not be the same for items in English and their translations. For example, passive voice may occur more frequently in some languages than others.

A specific set of text length and grammatical constraints needs to be determined for items or passages to be used in each grade and subject. The use of readability indexes—such as the Flesch-Kincaid index of readability available in Word—as proxy measures of linguistic complexity is discouraged in this conceptual framework. Since these indexes are developed for specific populations of readers and are applicable to specific kinds of text (Harrison, 1999), they are unlikely to produce dependable indicators of linguistic adequacy.

Restricted-Use Constituents

ELL students and users of non-standard English often struggle with homonyms (sets of terms with the same spelling but different meanings). An example is the word, *table,* which has a meaning in everyday life contexts—a piece of furniture—and a different meaning in academic contexts—an arrangement of data in rows and columns. Another example is the word, *root*, which has two meanings in different academic contexts—the part of a plant that attaches it to the ground and the value of the variable for which a polynomial is equal to zero. Terms like these can lead students to incorrect interpretations of items and should be treated as *restricted-use terms*.

A *book of restricted-use constituents* contributes to minimizing the chances for students to make incorrect interpretations of items due to polysemic terms. The document consists of a list of all the terms that have multiple meanings and establishes the one use that should be given across all items in an assessment.

An example of a book of restricted-use constituents is shown in Table 4.4. Of course, the example contains only a fraction of the many terms that it could include. Notice that the list includes both academic terms and operational terms.

Table 4.4

Fragment of a Hypothetical Book of Restricted-Use Constituents: Correct and Incorrect Use

Constituent	Context of Use in the Assessment	Example of Use	
		Correct	Incorrect
Table	Mathematics: A set of data arranged in rows and columns	*Use the **table** below to solve the problem.*	*The family ate dinner at the **table**.*
Line	Mathematics: A straight or curved continuous extent of length without breadth	*Write the equation corresponding to the **line**.*	*The students had to walk in a **line** to the cafeteria.*
Lesson	English Language Arts: The important point of a story	*What is the **lesson** of the story?*	*The teacher created a wonderful **lesson** plan.*
Character	English Language Arts: A person in a novel, play, or movie	*Who was the main **character** of the story?*	*Running away was not in keeping with her **character**.*
Click	Directions relevant to computer-based testing: Select or choose.	***Click** on the correct answer.*	*She heard the **click** of the door.*
Enter	Directions relevant to computer-based testing:: Type or key information in a computer	***Enter** the quotient.*	*She **entered** the kitchen.*

Fictitious Character Names

As part of the information they provide, many context-rich items use fictitious characters. The names of those characters may pose unnecessary linguistic challenges to ELL students or users of non-Standard English when they are uncommon or when they have features that make them likely to be confused with other words.

A *book of fictitious character names* facilitates the work of test developers by providing a restricted set of names for characters to be used across all the items generated by an assessment system. Across the different items and over time, students become familiar with the characters. Table 4.5 shows an example of a book of character names.

The document facilitates the work of test developers and test translators who, instead of searching for a name each time they need a character, simply select a name randomly from the table. In addition, the document helps test developers to keep a count of the number of times that each

character has been used. Moreover, it helps test developers to ensure a balance in the percentage of times a given character is right or wrong in problems formulated as a dilemma (common in Mathematics assessments), in which students are asked to decide which of two characters is right and which is wrong, and to justify their selection.

Table 4.5

A Hypothetical Book of Fictitious Character Names

Female Names		*Male Names*	
Destiny	*Rachel*	*Tyler*	*Eli*
Kiara	*Rebecca*	*Brandon*	*Abraham*
Alissa	*Sarah*	*Christian*	*David*
Yuki	*Diana*	*Hiro*	*Carlos*
Indira	*Maria*	*Manzur*	*Eduardo*
Meyumi	*Rosa*	*Cheng*	*Roberto*
Molly	*Aida*	*Steve*	*Abdul*
Claire	*Adela*	*John*	*Yousef*
Emily	*Shakira*	*Mike*	*Ahmed*

Each of the names included in the book of fictitious names should meet the majority of the following conditions:

(1) common in the U.S. or in the cultures in which they originated

(2) easy to pronounce

(3) simple spelling

(4) short

(5) representative of names used in different cultural/ethnic groups

(6) likely to be recognized by users of other languages

(7) without homonyms in English

(8) no phonetic or pronunciation similarity with other words in English

(9) typically used for one gender

In addition to increasing the usability of items and to facilitating the work of test developers, the book of fictitious character names ensures a fair representation of genders and cultural/ethnic groups.

Customization and Design from Scratch

Customization is the practice of modifying the characteristics of text to fit certain requirements. *Design from scratch* is the creation of original text when customization is not feasible.

Many English Language Arts items use text from selected literary work as passages which students need to read to respond to items. An argument in favor of using original literary work is the authenticity of the materials with which students ideally should be familiar. However, in the context of testing, this practice poses important challenges when literature appreciation or related skills are not part of the knowledge and skills the items are intended to assess. Those passages may contain multiple linguistic features that may be irrelevant to the targeted constructs, such as archaic or infrequently used vocabulary or narrative styles. This unnecessary linguistic complexity threatens the validity of interpretations of scores for all students, especially ELL students.

Finding passages from existing original texts that fit certain item specifications but do not have additional, unnecessary features is virtually impossible. Under such circumstances, test developers may need to either customize existing texts or design new passages from scratch. For practical reasons (including copyright issues), designing new passages from scratch may be a more viable strategy.

Designing passages to the specific needs of the assessment enables test developers to have better control of the linguistic features of the passages used as stimulus materials in English Language Arts items. Features that can be controlled include: vocabulary frequency, grammatical complexity, text length, sentence length, nominalization, and the use of embedded clauses.

Regardless of whether passages are customized or designed from scratch, steps should be taken to allow control of the topics and situations used in the passages in order to ensure equitable meaningfulness. This work can be done systematically with the aid of a *passage sampling matrix*.

Table 4.6 shows an example of a passage sampling matrix. The letter *n* in each cell represents the numbers of passages in an assessment that represent a combination of social locale and equitable meaningfulness.

Table 4.6

Passage Sampling Matrix: Percentages of Different Types of Topics Represented in an Assessment

	Equitable Meaningfulness	
Social Context (Locale)	A. The topic or situation reflects the everyday lives of many students, regardless of SES or cultural/ethnic group	B. The topic or situation reflects the everyday lives of individuals from a specific SES cultural/ethnic group, but is familiar to individuals from multiple SES or cultural/ethnic groups
Rural	*n*	*n*
Urban	*n*	*n*
Suburban	*n*	*n*
Indistinct	*n*	*n*

Pragmatic Suitability

Pragmatic suitability is the condition in which the features of items are appropriate in context, thus ensuring that students interpret them as intended by its developers.

While, typically, the process of assessment development includes stages in which the wording of items is reviewed, this review focuses on determining whether students are likely to interpret the items as intended, not on the unintended ways in which students may interpret them. Yet research has shown that looking for confirming and disconfirming evidence that the wording of items is adequate yield different, but complementary, kinds of information about the ways in which the wording of items needs to be improved (Basterra, Trumbull, & Solano-Flores, 2011; Solano-Flores, Backhoff, & Contreras-Niño, 2013). This notion is particularly important for ELL students and users of non-standard forms of English, among whom unintended interpretations of text may be more likely to occur.

Wording pragmatic suitability analysis is the activity intended to examine the appropriateness of the ways in which items are worded based on both the specific contexts of the items and the characteristics of the target population. This analysis should be conducted on the original English version of an item, on its stacked L1 translation, on its English glossaries, and on its L1 glossaries. The analysis is based on the question: *Which constituents may be misinterpreted by students due to the context presented by the item or to unanticipated linguistic similarities?*

Suppose that the first sentence of a Mathematics item and the glossary for one of its words read as shown in Figure 4.6.

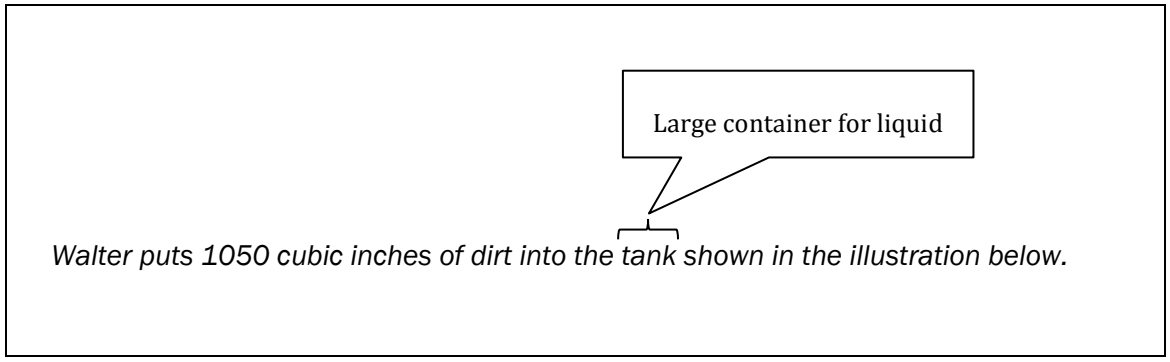

Figure 4.6. The sentence of a Mathematics item showing the definition glossary of one of its constituents. Adapted *from: Smarter Balanced Assessment Consortium (2013): Student Interface Practice and Training Tests:* Mathematics, Grade 5, Item No. 16. *https://login4.cloud1.tds.airast.org/student/V42/Pages/LoginShell.aspx?c=SBAC_PT&v=42* .

The reasoning of a team of reviewers performing the wording pragmatic suitability analysis for this item could be summarized as follows:

Since the context of the item involves a tank, the name of the character, *Walter*—which has a phonetic and formal resemblance to *water*, can create confusion for some ELLs.

The use of a tank as part of the context used to frame the problem may be inappropriate in this item because what is put in it is dirt (soil), not a liquid. Tanks are associated with liquids, not solids. This can create confusion among students.

The definition of "tank" provided by the glossary is the definition, "large container for liquid," which is in contradiction with the context presented by the item.
The inconsistency unnecessarily increases the cognitive load of the contextual information provided.

Suppose that a passage used in an English Language Arts item reads as shown in Figure 4.7.

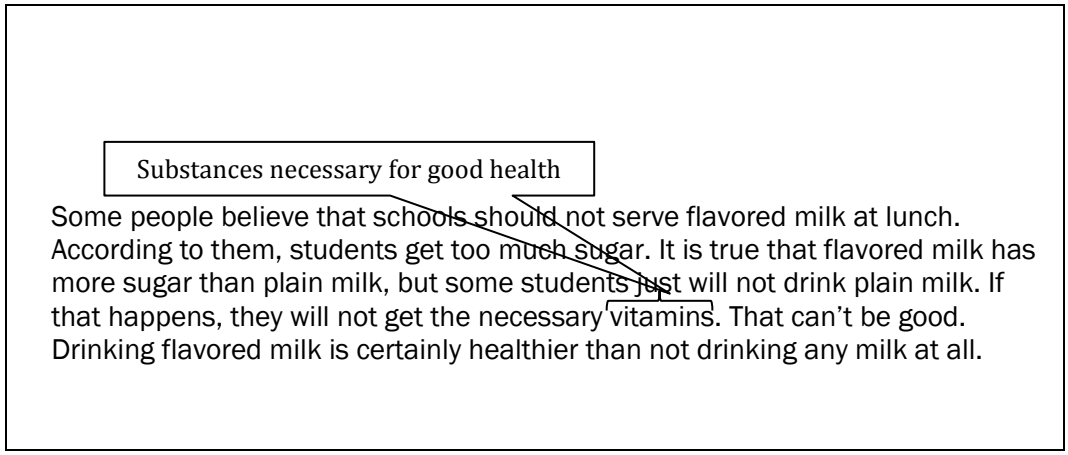

Figure 4.7. Portions of the passage for a set of English Language Arts items. Adapted from*: Smarter Balanced Assessment Consortium (2013): Student Interface Practice and Training Tests:* (English Language Arts, Grade 3, Item No. 18). *https://login4.cloud1.tds.airast.org/student/V42/Pages/LoginShell.aspx?c=SBAC_PT&v=42* .

The reasoning of a team of reviewers performing the wording pragmatic suitability analysis for this item could be summarized as follows:

> The English glossary defines vitamins as substances necessary for good health. In addition to being vague, the definition of vitamins as substances is not recommendable because, in inner-city environments, *substance* is a term more frequently used to refer to drugs.
>
> The term, *flavored milk* is not used among students. Students used terms such as *chocolate milk*, *strawberry milk*, *white milk*, or *plain milk*.
>
> The use of constituents that are not used by the target population imposes an unnecessary cognitive demand on students when interpreting such items.

Standardization

Standardization refers to the set of actions taken with the purpose of ensuring that all items of the same type have the same format and are developed with the same procedure. An outcome of standardization is that all items belonging to the same type of problem consistently have the same organization and appearance. Another outcome of standardization is that it enables test developers to generate items efficiently.

As a result of standardization, students can transfer their experience across all those items that have the same format. This minimizes cognitive load because students can focus, for example, on

solving a problem without simultaneously trying to figure out how they need to enter their responses. Minimizing cognitive load is especially relevant for students who already have an increased cognitive load inherent to being tested in a second language or a second dialect.

A powerful standardization strategy is the use of item shells. An *item shell* can be defined as a blueprint for creating items of the same type (Haladyna & Shindoll, 1989). An item shell can be thought of as both a document that specifies the structural properties of items and an authoring environment for test developers (Solano-Flores, Trumbull, & Nelson-Barber, 2002).

An effective shell establishes, with a high level of detail:

1. the physical arrangement of the components of the item (e.g., directions for students, contextual information, stimulus material, statement of a problem, space to provide a response)
2. the syntactical structure of the language used in the item
3. a set of generation rules

Figure 4.8 shows a shell for generating items involving addition of whole and fraction numbers. Figure 4.9 shows an item generated with the shell. This item is one of many possible items with the same content, complexity, and appearance that can be generated. Notice that, while a shell is a generic description, it can be very specific about the nature of information provided (e.g., the number of digits to the right of the decimal place) and the characteristics of the language used (e.g., number of sentences, range of number of words in a sentence). Indeed, it can contain sentences that must not vary across items. This enables test developers to have good control of the linguistic features of the items, thus complying with the principle of consistency, as discussed before.

Shell for Generating Items Involving Addition of Whole and Fraction Numbers

One 10-15 word long sentence describing a situation familiar to all students. The problem involves buying several elements to assemble something.

With the information in the table, find the total cost of the party.

Enter your response in the box below:

Type of Element	Unitary Cost	Type of subtotal
Name of Element 1	Fraction number	(>1; <9)
Name of Element 2	Whole value	1
Name of Element 3	Mixed value	1

Make sure to show all fractions to the hundredths. Do not show the zero for Element 1. In the column for unitary cost, In the Unitary cost column, show the word "each" (no abbreviations) next to each unitary cost.

Figure 4.8. Hypothetical shell used to generate items involving addition of whole and fraction numbers. Times New Roman font denotes text that has to appear as shown in all the items generated with the shell. Script typeface denotes specifications for test developers.

Your class is organizing a party for 20 people.

With the information in the table, find the total cost of the party.

Enter your response in the box below:

Treat	Cost	Needed per person
Cup of lemonade	.15 each	2
Fruit	1.00 each	1
Hot dog	1.35 each	1

Figure 4.9. An item generated with the shell shown in Figure 4.8.

5. DEVELOPMENT OF GLOSSARIES

A *glossary* is an accessibility resource that makes available to students alternative representations of the information communicated by a constituent in the text of an item with the intent to ensure that students understand the constituent appropriately and are able to make sense of the item as intended by the test developers. A *constituent* is a string of words that act together as a whole to encode meaning in a specific way. The term can apply to a word, a term, an idiomatic expression, or a grammatical construction.

Thanks to the information technology used to administer Smarter Balanced assessments by computer, glossaries can be offered as "pop-up" glossaries. When a student clicks on a constituent flagged as glossed, a box on the screen appears showing the glossary next to it.

A glossary is intended to minimize the threats to the validity of an item that stem from students' limited proficiency in the language in which tests are administered or from the lack of exposure to or lack of familiarity with certain features of the language used in tests, such as vocabulary, idiomatic expressions, or discursive structures.

Alternative forms of expressing the meaning encoded by a constituent include: a definition of the constituent, a synonym, a list of synonyms, or a paraphrase. These alternative forms of expressing a constituent can be provided in English, the language in which the constituent appears in the item, or, for English language learners (ELLs, emergent bilinguals), in the students' first language (L1).

In Smarter Balanced assessments, as with other assessment systems, such as NAEP [the National Assessment of Educational Progress], glossaries in L1 are not allowed for English Language Arts items. The reason for this restriction is that, being that English language is the knowledge assessed; it would be very difficult to provide this accessibility resource without giving away the responses to the items. However, English glossaries are provided in the stimulus materials used in many English Language Arts items. These stimulus materials are passages, tables, illustrations, etc., that provide information that students need to use to respond to the items. In Smarter Balanced assessments, English glossaries are made available for both English Language Arts and Mathematics to all students, regardless of their English proficiency status. In addition, L1 glossaries are provided in the Mathematics items to students classified as ELLs.

This chapter discusses the reasoning and methods for test developers and test translators to use to inform their decisions on which item constituents should be glossed and how these glossaries should be created. Unless stated otherwise, the discussion refers to both English and L1 glossaries.

Glossability Analysis

The term, *glossability* is used here to refer to the importance of glossing a constituent; how critical glossing a constituent is to supporting students to make sense of an item. Thus, the term, *glossability analysis* refers to the activity of examining the multiple constituents of an item and systematically determining which constituents should be glossed. Glossability analysis allows optimization of the number of glossaries offered in a given item.

This section discusses the glossability of constituents according to two aspects: (1) the linguistic demands of constituents and their relationship with the characteristics of what is called academic language and natural language and (2) the morphological and semantic correspondence of constituents within and across languages.

Linguistic Demands of Constituents in Natural and Academic Language

A notion widely held among test developers is that, in order to avoid compromising the assessed content of test items, linguistic support should not be provided for academic language constituents. Underlying this notion is the claim that, intrinsic to a knowledge domain is the command of the specialized language used to encode the facts, processes, reasoning, and methods that it comprises. A related notion is that the use of natural (everyday) language in tests contributes to minimizing unnecessary linguistic demands that test items may impose.

The reasoning has serious limitations. One limitation is the view of academic language and natural language as clearly distinguishable categories of language. Another limitation is the implicit assumption that a constituent's condition of "academic" or "natural" is fixed, regardless of the contexts in which it is used. Yet another limitation is the assumption that natural language imposes fewer linguistic demands.

A probabilistic reasoning allows a more defensible identification of glossable constituents. A probabilistic approach models language as a function of multiple variables, rather than a set of fixed categories. (Bod, Hay, & Jannedy, 2003). This approach recognizes randomness and the convergence of multiple historical, social, and contextual factors that shape how language is used at a given time in a specific situation (Solano-Flores & Gustafson, 2013). According to this probabilistic approach, there is no clear cut approach to distinguish what belongs to the category of "academic language" and what belongs to the category of "natural language" based only on the formal properties of constituents (Solano-Flores, in press). Recognizing these limitations, in the context of testing, and for the purpose of designing item accessibility resources, regarding a constituent as "academic" or "natural," should consider whether it:

- originated within a discipline to refer to a specific concept
- is likely to be used in academic or everyday life contexts
- is likely to be used by multiple or few speech communities
- is likely to be used in multiple or few everyday life contexts

Table 5.1 presents seven types of constituents (labelled A-G for short reference) that result from the combinations of these factors. The table identifies and illustrates seven types of academic language constituents (A and B), four types of natural language constituents (D, E. F, and G), and one hybrid type (C) in which the characteristics of academic and natural language converge.

The table also identifies the overall frequency of use of each type of constituent. *General use* and *restricted use* refer respectively to how likely a constituent is to be used (i.e., said, heard, written, read, and understood) in multiple contexts by multiple or few speech communities. The table rates the glossability of each type of constituent for Mathematics and English Language Arts. Notice that the types of constituents do not have the same levels of glossability in the two content areas.

Constituents of Types D, E, and F are of general use, mainly because they are used by multiple speech communities. Examples of speech community are:

- white, high-income students who are native English speakers and live in a suburban area
- black, low-income students who are native English speakers and live in an inner-city area
- Latino, low-income students who are ELLs, native Spanish speakers, and live in a rural area

Table 5.1

Examples of Types of Academic Language and Natural Language Constituents in Items From Two Content Areas: Use in Society and Glossability

	Academic		Hybrid	Natural			
				General use across social classes, ethnic/cultural groups, dialects, etc.		Specific to a social class, ethnic/cultural groups, dialects, etc.	
	A. Technical, created for a specific concept within a discipline, used **mainly** in academic contexts	B. Semi-technical, not originated within a discipline but used **mainly** in academic contexts	C. Non-technical, used by **multiple** speech communities in **both** academic and **multiple** everyday contexts	D. Used by **multiple** speech communities in **multiple** everyday contexts	E. Used by **multiple** speech communities in **few** everyday contexts	F. Used only by **few** speech communities in **multiple** everyday contexts	G. Used only by **few** speech communities in **few** everyday contexts
Use and Glossability							
Examples	*polygon*	*symbol*	*classify*	*appropriate*	*acknowledge*	*picture album*	*pitched a tent*
Overall Frequency of Use in the Society	Restricted	Restricted	General	General	General	Restricted	Restricted
Glossability							
in Mathematics items	Null	Null	Low	Low	Medium	High	High
in English Language Arts items' passages	High	High	Medium	Null	Low	High	High

Types A and B are of restricted use because they are likely to be used only in academic contexts—for instance, in textbooks and classroom conversations on topics related to a discipline (e.g., Mathematics). Types F and G are also of restricted use, although for a different set of reasons—they are a reflection of social stratification and social differences by virtue of which they are used only by the members of certain speech communities.

The table also shows four levels of glossability for constituents in Mathematics items and in passages used in English Language Arts items—Null (or not allowed), Low, Medium, or High. According to the table, what should be considered as highly glossable depends not only on the type of constituent and whether it is of restricted or general use in the society, but also on the content area in which the constituent is used.

For Mathematics items, Type A and Type B constituents are unglossable because they are intrinsically associated with the content the items are intended to assess. It is assumed that being familiar with these types of constituents is part of possessing the knowledge or the skills being assessed. Types C and D are regarded as having a Low level of glossability because they are likely to be used by multiple speech communities. Type E constituents are regarded as having a Medium level of glossability. Although these constituents are used by multiple speech communities, students are likely to have limited exposure to them because they are used in a limited number of everyday contexts. Types F and G constituents are regarded as highly glossable because they are likely to be used only by a limited number of speech communities (and, in the case of Type G constituents, only in a limited number of contexts).

Some types of constituents have different levels of glossability for English Language Arts and for Mathematics items. The reason is twofold. First, unlike Mathematics, the language used in English Language Arts items is part of the constructs measured, rather than being an extraneous factor. Second, unlike Mathematics, the academic language from other disciplines used in the passages of English Language Arts items is an extraneous factor, rather than part of the constructs measured.

Because language is relevant to the constructs measured by English Language Arts items, Type D, Type E, and Type C constituents are rated respectively as of Null, Low, and Medium glossability, which reflects the level of expectations regarding the use of language (e.g., knowledge of sophisticated vocabulary, familiarity with certain grammatical constructions).

Academic language from disciplines other than English Language Arts is an extraneous factor in English Language Arts items because these items use passages taken from text from other knowledge domains (e.g., archaeology, music, social sciences) which contain constituents that are irrelevant to this content area. Accordingly, Type A and Type B constituents are rated as highly glossable for English Language Arts.

Also rated as highly glossable for English Language Arts items are Type F and Type G constituents. These constituents are likely to appear, for example, in text that portrays selected cultural groups and situations. They are of restricted use in a society and favor members of speech communities who belong to those groups and who have been exposed to those situations.

Morphological and Semantic Correspondence of Constituents

In addition to the linguistic demands of constituents in English, glossability analysis addresses the morphological and semantic correspondence of constituents within and across languages. Needless to say, the glossability of a constituent in English varies by language, as each language has its own set of morphological and semantic correspondences with English.

Table 5.2 shows four cases of morphological and semantic correspondence between constituents and their glossability. Within the same given language, when a constituent has a synonym (i.e., a constituent with a different spelling and/or pronunciation but the same meaning), it is unlikely to mislead students' interpretations of an item. Its glossability is rated Low, although it should not be assumed that all students are familiar with all the synonyms of a given term.

Also within the same given language, when a constituent has a homonym (i.e., a constituent with the same spelling and/or pronunciation but a different meaning), it is likely to mislead students' interpretation of an item. Thus, its glossability is rated High.

Across languages, when the translation of a constituent is a cognate (i.e., a constituent with a similar spelling or pronunciation and a similar meaning), the constituent is unlikely to mislead ELL students' interpretations of an item. Its glossability is rated Low, although it should not be assumed that all students are equally able to identify cognates. For example, some bilingual, English-Spanish individuals may be surprised to realize that the terms "television" and "televisión" are similar not only in meaning, but in pronunciation and spelling, even if they use both words when they speak respectively in English and in Spanish. This is a common occurrence among bilingual individuals.

Also across languages, when the translation of a constituent is a false cognate (i.e., a constituent with similar spelling and/or pronunciation but a different meaning), the constituent is likely to mislead students' interpretations of an item. Thus, its glossability is rated High.

Table 5.2

Cases of Morphological and Semantic Correspondence: Within and Across Languages

	Constituents		Correspondence		
Relationship	Type	Example	Morphologial	Semantic	Glossability
Within Language (English)	English synonyms	*longed* *wished*	Different	Same	Low
	English homonyms	*table* (furniture) *table* (Mathematics)	Same	Different	High
Across Languages (English-Spanish)	English and a Spanish cognate	*quadrilateral* *cuadrilátero*	Similar	Same	Low
	English and a Spanish false cognate	*patron* (customer) *patrón* (boss)	Similar	Different	High

The combination of these four cases of morphological and semantic correspondence between constituents yield a complex set of relations between constituents and their possible translations that needs to be considered in the design of L1 glossaries. Synonyms and homonyms not only occur in English, but also in any of the languages into which tests are translated. As a consequence, ideally, when a translation of a cognate or a false cognate is made, the existence of synonyms and homonyms for these translations should also be taken into consideration, as Figure 5.1 shows.

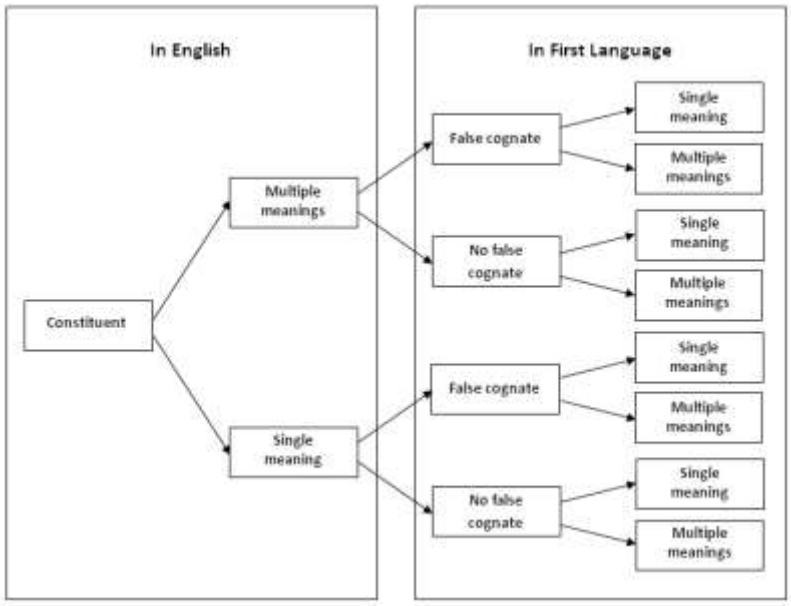

Figure 5.1. Morphological and Semantic Correspondence Within and Across Languages.

Glossary Design

This section discusses several concepts and principles for designing glossaries: (1) types of glossaries, (2) notion of semantic space, (3) selection of optimal constituents, (4) consistency of features, (5) glossing consistency, and (6) glossing density.

Types of Glossaries

As explained at the beginning of this chapter, a glossary is an accessibility resource that represents the information encoded by a constituent in alternative forms of representation. These alternative forms of representation of information may consist of synonyms (words or terms with similar meanings), definitions (statements of the meaning of concepts), and paraphrases (rewordings).

Table 5.3 compares these three types of glossaries as to sensitivity to context—how consistent the glossary can be with the context in which the constituent appears in the item—and usability—how easily students can understand it.

Table 5.3

Glossary Types: Capabilities

Type of Glossary	Capabilities: Sensitivity to Context	Capabilities: Usability	Example: Glossed Constituent	Example: Glossary
English				
Synonym	High	High	*fortunate*	*Lucky*
Paraphrase	High	High	*a jack of all trades*	*a person who knows something about everything*
Definition	Low	Low	*carnival*	*Event with food, music, and booths*
L1 (Spanish)				
Synonym	High	High	*fortunate*	*Afortunado*
Paraphrase	High	Medium	*a jack of all trades*	*un aprendiz de todo*
Definition	Low	Low	*carnival*	*Fiesta que precede al miércoles de ceniza*

For English and L1 synonyms and for English and L1 paraphrases, context sensitivity is rated High because these glossaries can communicate meaning in accord with the context in which constituents appear in items.

For English synonyms and paraphrases and L1 synonyms, usability is rated High because these glossaries impose few cognitive demands for students to figure out how the provided information can be used to make meaning of items. For L1 paraphrases, usability is rated Medium because the equivalence of idiomatic expressions across languages is not always perfect.

English and L1 definitions are rated Low, as definitions are decontextualized statements of the meanings of concepts—of which a glossed constituent is an instance. Also for both English and L1 definitions, usability is rated Low because definitions have a high level of abstraction. To benefit from definitions, students need to reason how they apply to the specific contexts in which constituents originate. In addition, definitions tend to pose unnecessary reading challenges due to their unique set of linguistic features (e.g., long sentences, multiple embedded clauses).

In the example, the definition in L1 has an additional problem—it is the definition of a false cognate. In Spanish, *carnaval* is typically used to refer to a party before Ash Wednesday, not to any kind of

event with food, music, and booths. While definitions may be relatively easy to create (e.g., by taking definitions from English-L1 dictionaries), errors of this kind are very likely to occur.

Semantic Space

The notion of semantic space is useful to design glossaries in a rigorous, systematic way, beyond simply showing synonyms or paraphrases of a constituent. A *semantic space* can be defined as a set of concepts or words interconnected by the meaning they share (Masucci et al., 2011). The concepts sharing meaning are called, *interpretants*.

As shown in Figure 5.2, thesauruses are lists of synonyms, and they are shown as tree-like representations of semantic spaces, such as those used with data visualization technology and network theory. They enable the user to understand the meaning of a concept or word from examining the interpretants that circumscribe its meaning. For the purpose of simplicity, this document focuses on thesaurus-like representations of semantic spaces. These representations are compatible with the ways in which glossaries are currently provided by Smarter Balanced. However, the principles for their design readily apply to tree-like representations.

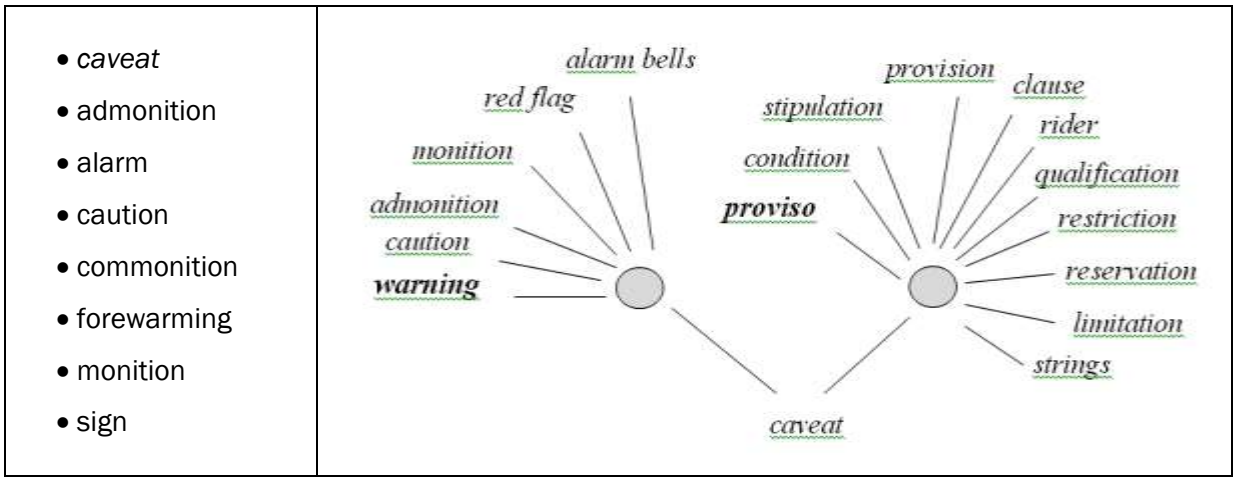

Figure 5.2. Semantic space representations of the word, caveat: List and tree. Adapted respectively from *Thesaurus.com* and *Wordflex.*

Unlike thesauruses, pop-up glossaries provided in Smarter Balanced assessments cannot always show all the interpretants of a given glossed constituent due to the small size of the boxes in which the glossaries are displayed. Most importantly, because the number of elements that the user needs to process simultaneously in working memory plays a key role in the complexity of a task, displaying many interpretants may unnecessarily increase the cognitive load inherent to understanding the meaning of the glossed constituent. Since cognitive load is shaped by a user's ability and practice with the topic at hand (Sweller, 1994), it would not be appropriate to assume that all users are equally able to interpret glossaries. Thus, there is a limit to the number of interpretants that a glossary can include (say, no more than four or five interpretants, or seven or eight words) if a glossary is to be an effective support. Hence the importance of selecting the optimal set of interpretants from the semantic space and according to the context of the text of the item in which the constituent appears.

This optimal set of interpretants can be established in three steps: (1) specifying the semantic space of the glossed constituent; (2) selecting the interpretants in which the meaning of the glossed constituent is circumscribed according to the context in which it appears in the text of the item; and (3) sequencing the selected interpretants according to the strength of their meaning relationships with the glossed constituent.

1. *Specification of a Semantic Space.* This step consists of assembling a list of all the words, terms, or phrases which, as a whole, test developers and test translators regard as circumscribing the meaning of the glossed constituent. In the case of L1 glossaries, these words, terms, or phrases are in L1.

   While commercially available thesauruses in English or in English-L1 dictionaries can be used in support of this activity, they should be used judiciously, only as resources. Those documents are decontextualized; they may not include interpretants that share meaning with the glossed constituent in the context in which the constituent appears.

   The limitations of English-L1 dictionaries can be especially serious; they may provide interpretants that are too prescriptive and which reflect a specific variety of L1 (mainly, a standard version of that L1) that may be the most prestigious variety of L1, but not the most common among the users of that L1 in the U.S. For instance, due to the unique set of characteristics of the varieties of Spanish in the United States (see Lipski, 2008; Moreno-Fernández, 2009), using a Spanish thesaurus published in Spain as a main source of reference may lead to creating ineffective glossaries.

   Given the limitations of commercially available thesauruses and dictionaries, test developers and test translators play a critical role in specifying a semantic space. These professionals need to ensure that this initial list of interpretants reflects the language used by their students in their schools and communities.

   In specifying a constituent's semantic space, test developers and test translators may need to decide about the inclusion of words, terms, or phrases that are informal or colloquial. An important fact to take into consideration in these decisions is that glossaries are intended to be linguistic supports, not academic dictionaries or thesaurus entries. Their intent is to ensure that students understand the meaning of constituents, rather than promoting the use of formal language. This can be accomplished only by capitalizing on the language with which students are familiar.

2. *Selection of Interpretants.* This step consists of selecting, from the specified semantic space, a small set of interpretants for inclusion in the glossary. This selection needs to be made according to both the characteristics of the item (i.e., the context in which the constituent appears in the text of the item) and the knowledge of the ways in which students use language (e.g., the interpretants with which the majority of the students are likely to be familiar).

   The complexity of this task cannot be overestimated, as there may be a different set of reasons to include or not to include each potential interpretant. For example, an interpretant can be considered a good synonym to the glossed constituent, but, because it is rarely used, students may be unlikely to be familiar with it. Thus, in addition to semantic proximity, an important aspect to consider in the decision process for the inclusion of an interpretant is the extent to which the word or phrase is used by multiple speech communities in multiple everyday contexts.

3. *Sequencing of Interpretants.* This step consists of listing the selected interpretants, from left to right, in descending order of the strength with which the meaning of each is related to the meaning of the glossed constituent. For example, using their knowledge of the use of language among students and in the students' schools and communities, test developers and test translators may list the selected interpretants from the most likely to the least likely to be said, heard, written, read, and understood in multiple contexts by multiple speech communities.

   Several criteria can be used to sequence the interpretants, including frequency of use, precision, and number of users of different L1 dialects, as Table 5.4 shows. An important factor to consider in the development of L1 glossaries is the tremendous linguistic heterogeneity of ELLs within the same broad linguistic group. In the table, the sequence in which different interpretants are listed in the L1 glossary for *straw* reflects the proportion of users of three Spanish dialects in the U.S.—Mexican, Colombian, and Iberian Spanish.

Table 5.4

Examples of Sequences of Interpretants Displayed by Glossaries: Language and Main Sequencing Criterion

Type of Glossary	Constituent	Criterion	Sequence in the Glossary
English	*Archaic*	Frequency of use	*old, ancient, obsolete*
English	*Symbol*	Precision	*sign, figure, logo*
L1	*Straw*	Number of users of L1 dialects in the U.S.	*popote, pitillo, pajilla,*

Selection of Optimal Constituents

An *optimal constituent* is a constituent that encodes the maximum meaning possible in the minimum number of words. Identifying optimal constituents is critical to effective glossary development.

Suppose that in the sentence

At a lemonade stand, each cup of lemonade cost 24 cents

stand is identified as a word that may pose a challenge to ELL students. Logical reasoning leads to thinking about the best word into which *stand* can be translated. However, while *lemonade* is not identified as a word posing a challenge to students, a better constituent to translate is *lemonade stand. Lemonade stand* is an optimal constituent because it encodes more meaning than *lemonade* and *stand* separately. As a consequence, a glossary for *lemonade stand*, as a whole, is more meaningful and may require the display of fewer interpretants than the glossary for *stand*.

Consistency of Features

The usability of both English and L1 glossaries can be increased by ensuring that the features of the interpretants included in a glossary are consistent with the features of the glossed constituent. Most of these features are grammatical, such as number (plural or singular), tense (past, future, etc.), case (nominative, accusative, genitive), gender (male, female), etc. Other features whose consistency may need to be preserved in the glossing may be format, such as the use of italics, capitalization, etc.

As Table 5.5 shows, each constituent has a specific set of features which need to be preserved in its glossary in order to increase usability.

Table 5.5

Examples of Interpretants that are and are not Grammatically Consistent with the Glossed Constituent

			Interpretant	
Type of Glossary	Glossed Constituent (Underlined) and Sentence in which it Appears	Critical Grammatical Feature	Consistent	Inconsistent
English	*John entered his answer.*	*tense*	*clicked*	*click*
English	*He bought some utensils for his trip.*	*number*	*tools*	*tool*
English-L1	*The collector's house was beautiful.*	*genitive*	*casa del coleccionista*	*coleccionista*

Contrary to common intuition, grammatical consistency is possible across languages. Moreover, bilingual individuals are used to code switching between languages within the same sentence in ways that do not violate the grammatical rules of either language. This characteristic, which is indicative of a strength rather than a deficit, can be followed in the design of L1 glossaries. Accordingly, in an effective L1 glossary, the grammatical features of any of its interpretants should allow replacing the glossed constituent (in English) by the L1 interpretant without violating the grammatical rules of either English or L1. Often, meeting this condition also involves identifying an optimal constituent.

Take as an example the sentence:

I'm not sleepy and there is no place I'm going to[4]

Suppose that *sleepy* is identified as a constituent that may pose unnecessary challenges for ELL students whose first language is Spanish.

A translation of sleepy in Spanish is *somnoliento*. If this word replaces sleepy, the grammatical integrity of the sentence in both English and in Spanish is preserved, as shown in Figure 5.3.

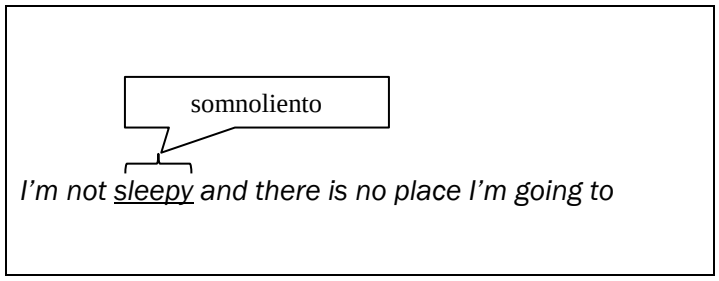

Figure 5.3. Glossing of *sleepy*.

However, *somnoliento* is uncommon in Spanish, at least among children and youth. This constituent is accurate but its frequency of use is low.

An alternative approach to address the challenges posed by *sleepy* consists of expanding the constituent to include other words. *I'm not sleepy* can be an optimal constituent to glossary. If the translation, *No tengo sueño* replaces the constituent, the grammatical integrity in the two languages is still preserved, as shown in Figure 5.4.

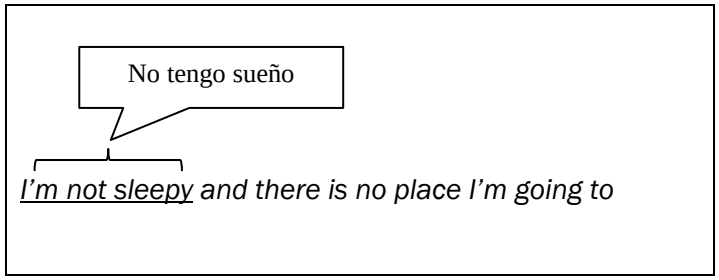

Figure 5.4. Glossing of *I'm not sleepy*.

Ensuring grammatical consistency between the constituent and the interpretants and identifying optimal constituents beyond isolated words increases the usability of glossaries.

Glossing Consistency

The same set of criteria used to determine which constituents need to be glossed should be applied consistently throughout all the items in an assessment. Glossing consistency is an indicator that glossing is performed systematically. Also, it contributes to the overall usability of all the glossaries provided in an assessment by allowing students to develop a sense of the types of constituents that are glossed across items and how they are glossed. Three forms of glossing consistency can be identified: *horizontal, vertical, and functional*.

Horizontal Consistency. If a given constituent is glossed, other constituents appearing in the same item which are concepts of the same type should also be glossed. The principle of horizontal consistency is supported by the notion of coordination in linguistics—a form of hierarchical organization which identifies linguistic units of equal importance or weight (Hudson, 1988).

Take the item shown in Figure 5.5.

Christy has $60 to buy some plants.

She buys a peach tree for $23 and a plum tree for $19.

She wants to buy one more plant.

(…)

Choose a plant she could buy with the money she still has.

Figure 5.5. Example of horizontal consistency. Adapted from: *Smarter Balanced Assessment Consortium (2013): Student Interface Practice and Training Tests:* Mathematics, Grade 3, Item No. 9. https://login4.cloud1.tds.airast.org/student/V42/Pages/LoginShell.aspx?c=SBAC_PT&v=42.

Suppose that *plum tree* is identified as a constituent to gloss. Then, *peach tree* should also be glossed. While *peach tree* may not pose challenges to students, its glossing enables students to better understand *plum tree* by making evident the contrast of concepts of the same type.

Vertical Consistency. If a given constituent is glossed, other constituents of higher glossability appearing in the same item should also be glossed. The principle of vertical consistency is supported by the notion of subordination in linguistics—a form of hierarchical organization which identifies linguistic units of different importance or weight (Hudson, 1988).

Suppose that the constituents *acknowledge* and *pitched a tent* appear in the same Mathematics item. Since the glossability of the former is Medium and the glossability of the latter is High, there is no justification for glossing *acknowledge* without also glossing *pitched a tent*.

Functional Consistency. If a constituent is glossed within a given item, the same constituent should also be glossed when it appears in other items. A principle established in the field of design (Lidwell, Holden, & Buttler, 2003), functional consistency "improves usability and learnability by enabling (users) to leverage existing knowledge about how the design (of a system) functions" (p. 56). Functional consistency ensures that all items generated by Smarter Balanced (and many other large-scale assessment systems) have comparable levels of accessibility in spite of the fact that, due to its spiral design, different sets of students are given different sets of items.

Glossing Density

Glossing density can be defined as the proportion of glossed words with respect to the total number of words in that item. Glossing density can be used by test developers and test translators as an evaluation tool. A high glossing density may indicate that an item poses too many unnecessary linguistic challenges. It also may indicate that too many of the constituents of an item have been glossed—probably unnecessarily. A considerable glossing density variation among the items of a test indicates that the items are uneven in their linguistic complexity or that the glossing functional consistency is low.

While it is a gross indicator of the linguistic complexity of items and the quality with which they are glossed, the concept of glossing density allows coordination of the work of test developers. Its importance stems from the fact that the effectiveness of accessibility resources cannot be examined in isolation, without considering a test as a whole.

6. MODEL FOR THE INCLUSION OF LANGUAGES IN ASSESSMENT SYSTEMS

The Smarter Balanced Assessment Consortium offers translations of its test items into various languages with the intent to ensure equitable access to the content of items and equitable sets of opportunities for students to demonstrate knowledge.

The main target population of this accessibility feature is the population of emergent bilinguals or English language learners (ELLs)—students who are developing English as a second language while they continue developing their first language. The translations offered are full translations or pop-up glossaries in the students' primary language (L1). Currently, full translations are available only in Spanish, while pop-up glossaries are available in Arabic, Cantonese, Ilokano (Filipino), Korean, Mandarin, Punjabi, Russian, Spanish, Tagalog (Filipino), Ukrainian, and Vietnamese. The variety of languages and dialects served and the number of items to be translated makes this one of the most ambitious test translation endeavors in history.

This chapter adds to the set of documents that formalize the procedures used by Smarter Balanced, which include a conceptual framework on test translation (Solano-Flores, 2012) and a family of guidelines on accessibility and accommodations (e.g., Measured Progress & Educational Testing Service, 2012; Measured Progress & National Center on Educational Outcomes, 2014). It provides a conceptual tool that allows systematic selection of the languages it is to serve in the future. The need for such a conceptual tool increases as Smarter Balanced enters into its operational stage and translation procedures are streamlined—which raises among states, educators, and linguistic groups, the expectations for inclusion of other languages.

Deciding which languages should be included for translation to serve ELLs is more complex than it looks at first glance. While the vast majority of ELLs are users of a few languages (e.g., Spanish, Tagalog), the proportions of users of certain national low frequency languages are tremendously high for certain states and for certain regions within those states. An unfortunate consequence of this disparity is that ELL populations from Smarter Balanced states benefit differently from the translation accessibility resources.

This document offers a model on the inclusion of languages in the group of languages served by Smarter Balanced. The model is not intended to propose specific languages to include. Rather, it is intended to support decision makers in their reasoning and inform their decisions.

The model is presented in two sections. The first section briefly discusses the concepts of relevance and viability as basic to making language inclusion decisions, and the notion of priority space as the relationship between relevance and viability. The second section offers a procedure for language selection.

Basic Concepts

According to the model, deciding which languages are to be included necessitates a consideration of the tension between two sets of factors, *relevance factors* and *viability factors*. This tension results from the high number of linguistic groups, which could benefit from test translation into their primary languages and the limited human and financial resources available to support them. Effectively addressing this tension is based on establishing which languages should be given priority over other languages.

Relevance Factors

Relevance factors contribute to making a compelling case in favor of including a language for translation, thus giving it precedence for inclusion over other languages.

Table 6.1 lists some relevance factors, grouped in three main categories: *frequency, proportionality,* and *criticality*. Needless to say, the table is not exhaustive; other relevance factors may need to be considered. As the table shows, the relevance of a language as a candidate for translation may be justified by the fact that it is the primary language of many students in the consortium states (frequency) or because it is the primary language of a high percentage of students in a given state (proportionality).

In contrast, criticality justifies the support of a language even if it has a low frequency and a low proportionality. The condition of criticality may stem from multiple factors related to social justice and history, such as marginalization and the need for measures of academic achievement for certain linguistic groups.

Table 6.1

Relevance Factors Relevant to the Inclusion of an ELL Primary Language

Frequency • *What is the sheer number of users of the language as a primary language across Smarter Balanced states?*
Proportionality • *What is the percentage of users of the language as a primary language within a given state?*
Criticality • *Is the ethnic/cultural or socioeconomic group user of the language as a primary language vulnerable or historically underrepresented?* • *Is that group particularly vulnerable due to poverty or segregation?* • *Does that group rarely benefit from social programs?* • *Are the indicators of academic achievement for that group limited?* • *Is that group among the groups with the lowest national or state academic achievement?*

Viability Factors

Viability factors contribute to making a compelling case that, if a given language is included for translation, this translation is likely to succeed and may contribute to obtaining valid measures of academic achievement of its users.

Table 6.2 lists some viability factors. As with the previous table, this table is not exhaustive, as there may be many other viability factors that need to be considered.

As Table 6.2 shows, the most obvious viability factor is cost. For the purpose of this document, cost can be understood in multiple ways (e.g., financial, political, logistical). For example, *How much money will it cost to develop a translation in a language whose speakers live in a remote area?* Or, *How difficult is it to gain access to a given speech community?*

Regarding human resources, the availability of qualified professionals should not be underestimated as a viability factor. For certain languages, it may be extremely difficult to find potential translators living outside their communities. In addition, since official certification from professional translator organizations is available only for a few cosmopolitan languages, special profiles of translators may need to be created.

Table 6.2

Viability Factors Relevant to the Inclusion of ELL Primary Languages

Sustainability • *How likely are the translators to keep doing translation work for a long time in the future?* • *Are students from that group schooled in their primary language?* • *Is there a critical mass of teachers who are users of their students' primary language?*
Human Resources • *Are there sufficient individuals with the proper qualifications needed to properly translate the tests into the language?* • *Are these individuals easy to identify and recruit?*
Cost • *How much money will have to be spent to develop translations in the language?* • *How complex is the logistics needed to stage in order to properly develop those translations?*
Dependability of Information • *How trustworthy is the existing information about the language?* • *How dependable is the information on the numbers of users of that language as a primary language?*
Fidelity of Implementation • *How likely will the language version of the test be created in accord with the existing procedures?* • *Are there ways of evaluating the proper implementation of those procedures?*

The dependability of information on a given language or its users may be a source of concern. For example, the speakers of different languages with small numbers of users may be wrongly classified as users of the same language. Especially for threatened languages, languages in remote areas, or languages used by marginalized ethnic/cultural or socioeconomic groups, the data on numbers of users may be outdated or unreliable.

Fidelity of implementation and sustainability are factors shaped by the stability of a linguistic group or its communities of users, and the critical mass of professionals that can reasonably be expected to continue performing translation tasks for many years during the operational stage of the assessment.

Priority Space

The tension between relevance and viability can be represented as a priority space, which represents the relation between relevance and viability for each language in a given set of languages.

Suppose that sufficient information is gathered from different sources concerning relevance and viability factors for a large set of languages in the U.S., and for which inclusion decisions need to be made. If the information from all those sources are properly combined and standardized, a priority coefficient can be calculated for each language as indicated by the formula,

$$P_l = r / v \qquad (0 < r \leq 1;\ 0 < v \leq 1) \qquad \text{[Eq. 6.1]}$$

in which *P* is the priority that should be given to a language, *l,* and *r* and *v* are, respectively, the standardized measures of relevance and viability.

Put simply, the priority coefficient compares how necessary it is to include the language and how feasible it is to include it. The following cases can be identified:

Case 1. P ≈ 1: Proportionally, viability is approximately equal to relevance.

Case 2. P > 1: Proportionally, viability is low, relevance is high.

Case 3. P < 1: Proportionally, viability is high, relevance is low.

Figure 6.1 represents these three cases in what can be called, a *priority space*, a bi-dimensional representation of the proportion of *r* and *v* for all the languages being considered. Languages in Case 2 and Case 3 areas are languages whose selection for inclusion may be inappropriate because they are likely to result, respectively, in a waste of resources and failure.

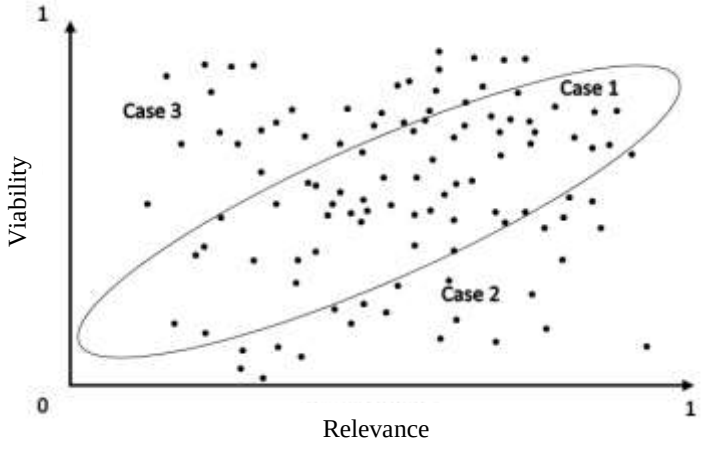

Figure 6.1. Priority space for the inclusion of languages.

The ellipse in Figure 6.1 shows the languages that are in Case 1. The languages in this area of the priority space are more likely to be cost-effective and successful selections for inclusion. However, a

step further has to be taken to refine the model because Case 1 includes languages for which $P \approx 1$ but the values of both *r* and *v* are low. Such languages are not a concern.

Building on Eq. 6.1, *P* can be computed under the restriction that, a language is considered for inclusion only when certain minimum values of *d* and *f* are satisfied:

$$P_I = r / v \qquad (0.5 < r \leq 1;\ 0.5 < v \leq 1) \qquad \text{[Eq. 6.2]}$$

Of course, the new value ranges shown here are arbitrary. Yet the minimum value of 0.5 for both *r* and *v* appears to be reasonable, as it clearly identifies those values that are in the upper half of possible values of *r* and in the upper half of possible values of *v*.

The ellipse in Figure 6.2 shows the languages belonging to Case 1, with the new set of *r* and *v* value range specifications. These languages are the languages that should be given the highest priority for inclusion. They can be called, *high priority Case 1 languages*.

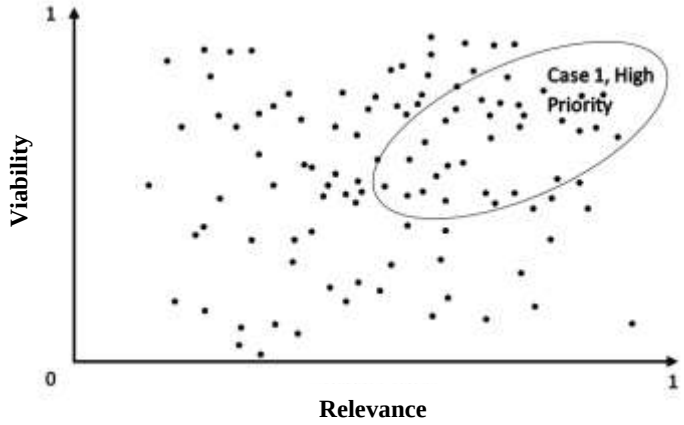

Figure 6.2. Priority space for the inclusion of languages: Refined model showing high priority languages.

Steps for Language Selection

This section presents a procedure for Smarter Balanced, decision makers, and vendors, to use in identifying languages for inclusion. The procedure addresses the challenges of making inclusion decisions for high numbers of languages, each with a unique pattern of relevance and viability.

1. ***Creation of a Language Inclusion Committee.*** A committee is established that includes Smarter Balanced staff, decision makers, vendors, sociolinguists, advocate group representatives, Smarter Balanced state representatives, experts on the assessment and instruction of ELL students, and data analysts. This committee is charged with: ensuring that appropriate and accurate information is obtained about multiple potential languages for inclusion, analyzing and interpreting that information, making inclusion decisions, and reporting those decisions.
2. ***Call for Nomination of Languages.*** Smarter Balanced and the Language Inclusion Committee invite states and stakeholders to nominate languages for inclusion in the set of languages into which Smarter Balanced are to be translated. The call establishes the kind of information that states and stakeholders are to provide in support of the languages they

nominate for support. This information includes but is not limited to the information listed in Strands 1-12 in Table 6.3.

Table 6.3.

Strands of Information Provided by Language Nominators and Analyses Performed by the Language Inclusion Committee (LIC)

Factor	Information/Analysis Strand	Source of Data
Relevance		
Frequency	1. Statistical information on the sheer numbers of users of the language as a primary language across Smarter Balanced states	Nominator
Proportionality	2. Information on the percentages of users of the language as a primary language within the state	Nominator
Criticality	3. Information on how the ethnic/cultural or socioeconomic group user of the language as a primary language is vulnerable or historically underrepresented	Nominator
	4. Information on how that group is particularly vulnerable due to poverty or segregation	Nominator
	5. Evidence that the group rarely benefit from social programs	Nominator
	6. Justification of relevance based on the limited availability of indicators of academic achievement for that group	Nominator
	7. Evidence that the group is among the groups with the lowest national or state academic achievement	Nominator
Viability		
Sustainability	8. Evidence on the availability of qualified translators that will continue the translation work for a long time in the future	Nominator
	9. Evidence that a substantial proportion of the students from that group are schooled in their primary language at least some years	Nominator

(continues)

Table 6.3.

(continuation)

Factor	Information/Analysis Strand	Source of Data
	10. Evidence that there is a critical mass of teachers who are users students' primary language	Nominator
Human Resources	11. Detailed information about the individuals with the proper qualifications needed to properly translate the tests into the language	Nominator
	12. Information about the availability of translators—how easy they are to identify and recruit	Nominator
Cost	13. Analysis of the costs of developing translations in the language	LIC
	14. Analysis of the complexity of the logistics needed to stage in order to properly develop those translations	LIC
Dependability of Information	15. Analysis of the trustworthiness of the existing information about the language	LIC
	16. Analysis of the dependability of is the information on the numbers of users of that language as a primary language	LIC
Fidelity of Implementation	17. Analysis of how likely the language version of the test will be created in accord with the existing procedures	LIC
	18. Analysis of the ways in which the implementation of the translation procedures can be evaluated	LIC

3. ***Preliminary Analysis.*** Based on the information provided by nominators (Strands 1-12, Table 6.3) and its own information sources, the Language Inclusion Committee performs a series of analyses concerning costs, dependability of information, and fidelity of implementation in support of translating each of the languages nominated (Strands 13-18, Table 6.3).
4. ***Priority Analysis.*** Using the reasoning described in the previous section, the Language Inclusion Committee transforms and summarizes the information provided by the nominators to calculate, first, a relevance (*r*) coefficient (Strands, 1-7, Table 6.3), then a viability (*v*) coefficient (Strands 8-18), then a *P* coefficient (see Eq. 6.2) for each of the nominated languages. This *P* coefficient allows the Language Inclusion Committee to locate each language in the priority space shown in Figure 6.2 and determine if it belongs in the area labeled, *Case 1, High Priority*.

5. ***Reporting.*** For reporting purposes, it can be useful to organize in one table the information on relevance and viability for all languages considered. To this end, the languages can be represented according to a limited number of levels of relevance and viability. In the example shown in Table 6.4, three levels of relevance and three levels of viability are used. Keys denote groups of languages that can be characterized by the combination of a particular level of relevance and a particular level of viability. Roughly, High Priority Case 1 languages belong in Cells 2, 3, and 6. This representation of information allows systematic analysis of subsets of languages. For example, decision makers may focus their discussion on potential inclusion of languages for those languages belonging in Cell 3.

Table 6.4

Organization of Information on Relevance and Viability on the Inclusion of Languages

	Viability		
Relevance	Low	Medium	High
High	{H, L} [1]	{H,M} [2]	{H, H} [3]
Medium	{M, L} [4]	{M, M} [5]	{M, H} [6]
Low	{L, L} [7]	{L, M} [8]	{L, H} [9]

Notes

[1]Lennon, J. & McCartney, P. "Baby, You're a Rich Man." Lyrics. Perf.: The Beatles. *All You Need is Love*. London, England: Parlophone Records, Ltd., 1967.

[2]Williams, P., & Nichols, R. "Rainy Days and Mondays." Lyrics. Perf.: The Carpenters. *Carpenters*. Santa Monica: A&M Records, 1971.

[3]Davies, R., & Hodgson, R. "Goodbye Stranger." Lyrics. Perf.: Supertramp. *Breakfast in America*. Almo Music Corp., Delicate Music, 1979.

[4]Dylan, B. (1965). "Hey, Mr. Tambourine Man." Lyrics. Perf.: Bob Dylan. *Bringing it All Back Home.* New York, NY: Columbia Records, 1965.

[5]*Caveat*. (n.d.) In *Thesaurus.com.* Retrieved June 1, 2014 from http://thesaurus.com/browse/caveat.

[6]*Caveat*. (n.d.) In *Wordflex.com* Retrieved June 1, 2014 from wordflex.com.

[7]Smarter Balanced also offers translations in American Sign Language for deaf and hard of hearing students. However, this population is not part of the scope of this document.

References

Abedi, J. (2008). Classification system for English language learners: Issues and recommendations. *Educational Measurement: Issues and Practice*, *27*(3), 17-31.

Agha, A. (2003). The social life of cultural value. *Language & Communication*, *23*(3-4), 231–273. doi:10.1016/S0271-5309(03)00012-0

Allalouf, A. (2003). Revising translated differential item functioning items as a tool for improving cross-lingual assessment. *Applied Measurement in Education, 16*(1), 55-73.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Aukerman, M. (2007). A culpable CALP: Rethinking the conversational/academic language proficiency distinction in early literacy instruction. *The Reading Teacher, 60*(7), 626-635.

Basterra, M. R., Trumbull, E., & Solano-Flores, G. (Eds.), (2011). *Cultural validity in assessment: Addressing linguistic and cultural diversity*. New York: Routledge.

Bod, R., Hay, J. & Jannedy, S. (Eds.) (2003). *Probabilistic linguistics*. Cambridge, MA: MIT.

Butler, F. A., Lord, C., Stevens, R., Borrego, M., & Bailey, A. L. (2004). *An approach to operationalizing academic language for language test development purposes: Evidence from fifth-grade science and math* (CSE Report 626), National Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation, Graduate School of Education & Information Studies, University of California, Los Angeles.

Caple, H. (2008). Intermodal relations in image nuclear news stories. In L. Unsworth (Ed.), *Multimodal semiotics: Functional analysis in contexts of education* (pp. 123–138). London and New York: Continuum International Publishing Group.

Coulmas, F. (2013). *Sociolinguistics: The study of speakers' choices. 2nd Edition*. New York, NY: Cambridge University Press.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In California State Department of Education (Ed.*), Schooling and language minority students: A theoretical framework.* Evaluation, Dissemination and Assessment Center, California State University, Los Angeles.

Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology*, *11*(44), 87–100. doi:10.1146/annurev-anthro-092611-145828

Eco, U. (1984). *Semiotics and the philosophy of language*. Bloomington, IN: Indiana University Press.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multi-language assessments. *International Journal of Testing*, 2, 199-215.

Escamilla, K. (2000). Bilingual means two: assessment issues, early literacy and Spanish-speaking children. A research symposium on high standards for students from diverse language groups: Research practice and policy. Proceedings:April 19-20, 2000. Washington D.C.

García, O., & Kleifgen, J. A. (2010). *Educating emergent bilinguals: Policies, programs, and practices of English language learners*. New York: Teachers College Press.

Gibson, J. J. (1977). The theory of affordances. In R. E. Shaw & J. Bransford (Eds.), *Perceiving, Acting, and Knowing*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Greenfield, P. M. (1997). You can't take it with you: Why ability assessments don't cross cultures. *American Psychologist, 52*(10), 1115–1124.

Guichon, N., & McLornan, S. (2008). The effects of multimodality on L2 learners: Implications for CALL resource design. *System*, *36*, 85–93. doi:10.1016/j.system.2007.11.005

Hakuta, K., & Beatty, A. (Eds.). (2000). *Testing English-language learners in U.S. schools: Report and workshop summary*. Washington, DC: National Academy Press.

Haladyna, T. M. & Shindoll, R. R. (1989). Shells: A method for writing effective multiple-choice test items. *Evaluation and the Health Professions*, *12*, 97-104.

Halliday, M. A. K. (1978). *Language as social semiotic*. London: Edward Arnold.

Harrison, C. (1999). Readability. In B. Spolsky (Ed.), *Concise encyclopedia of educational linguistics* (pp. 428-431). Oxford, UK: Elsevier.

Hudson, R. (1988). coordination and grammatical relations. *Journal of Linguistics*, *24*(2), 303-342. DOI: http://dx.doi.org/10.1017/S0022226700011816

Hull, G. A., & Nelson, M. E. (2005). Locating the semiotic power of multimodality. *Written Communication*, *22*(2), 224–261. doi:10.1177/0741088304274170

Iedema, R. (2003). Multimodality, resemiotization: extending the analysis of discourse as multi-semiotic practice. *Visual Communication*, *2*(1), 29–57. doi:10.1177/1470357203002001751

Kopriva, R. J. (Ed.) (2008). *Improving testing for English language learners: A comprehensive approach to designing, building, implementing and interpreting better academic assessments.* New York, NY: Routledge.

Kress, G., & van Leeuwen, T. (2001). *Multimodal discourse. The modes and media of contemporary communication*. London: Arnold.

Lantolf, J. P. (1996). Introducing sociocultural theory. In *Sociocultural theory and second language learning* (pp. 1–26). Oxford University Press.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. New York: Cambridge University Press.

Lee, O. (2005). Science education with English language learners: Synthesis and research agenda. *Review of Educational Research, 75*(4), 491–530.

Lemke, J. L. (1998). Multiplying meaning: Visual and verbal semiotics in scientific text. In J. R. Martin & R. Veel (Eds.), *Reading science: Critical and functional perspectives on discourses of science* (pp. 87-113). New York: Routledge.

Lidwell, W., Holden, K., & Butler, J. (2003). *Universal principles of design*. Beverly, MA: Rockport Publishers, Inc.

Linquanti, R., & Cook, H. G. (2013). *Toward a "common definition of English language learner."* Washington, DC: Council of Chief State School Officers, February 1.

Lipski, J. M. (2008). Varieties of Spanish in the United States. Washington, DC: Georgetown University Press.

Masucci A. P., Kalampokis, A,, Eguiíluz, V. M., & Hernández-Garcıía, E. (2011). *Wikipedia information flow analysis reveals the scale-free architecture of the semantic space*. PLoS ONE 6(2): e17333. doi:10.1371/journal.pone.0017333

Mayer, R. E. (2001). *Multimedia learning*. New York: Cambridge University Press.

Measured Progress & Educational Testing Service (2012). *Smarter Balanced Assessment Consortium: General accessibility guidelines*. April 16, 2012. http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CBOQFjAA&url=http%3A%2F%2Fwww.smarterbalanced.org%2Fwordpress%2Fwp-content%2Fuploads%2F2012%2F05%2FTaskItemSpecifications%2FGuidelines%2FAccessibilityandAccommodations%2FGeneralAccessibilityGuidelines.pdf&ei=eivuU-nIIoiWyASirIK4Bw&usg=AFQjCNECeF3Cb3EN2eZt51exeXqXHWPEFA&bvm=bv.73231344,d.aWw

Measured Progress & National Center on Educational Outcomes (2014). *Smarter Balanced Assessment Consortium: Accessibility and accommodations framework*. http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=0CCMQFjAB&url=http%3A%2F%2Fwww.smarterbalanced.org%2Fwordpress%2Fwp-content%2Fuploads%2F2014%2F02%2FAccessibility-and-Accommodations-Framework.pdf&ei=MBXuU7f4ENKAygT05IKgBA&usg=AFQjCNGGXEiowp4tIzGTMLy1mIJQfmrgEw&bvm=bv.73231344,d.aWw

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). (pp.13–103). Washington, DC: American Council on Education & National Council on Measurement in Education.

Moreno-Fernández, F. (2009). Dialectología hispánica de los Estados Unidos. En H. López-Morales (Coord.), Enciclopedia del español de los Estados Unidos (pp. 200-221). Madrid: Spain.

Nasir, N. S., & Hand, V. M. (2006). Exploring sociocultural perspectives on race, culture, and learning. *Review of Educational Research*, *76*(4), 449–475. doi:10.3102/00346543076004449

Nettle, D., & Romaine, S. (2002). *Vanishing voices: The extinction of the world's language*s (pp. 78–98). New York, NY: Oxford University Press, Inc.

Norman, D. A. (1988). *The design of everyday thing*s. New York: Basic Books.

O'Halloran, K. L. (2005). *Mathematical discourse: Language, symbolism and visual images*. Continuum International Publishing Group.

Oller, D.K., Pearson, B. Z., & Cobo-Lewis, A. B. (2007). Profile effects in early bilingual language and literacy. *Applied Psycholinguistics 28*, 191–230.

Rogoff, B. (2003). Development as transformation of participation in cultural activities. *The Cultural Nature of Human Development* (pp. 37-62). Oxford University Press.

Saffer, D. (2014). *Microinteractions: designing with details*. Sebastopol, CA: O'Reilly.

Scarcella, R. C. (2003). *Academic English: A conceptual framework*. Report 2003-1. Santa Barbara, CA: University of California Linguistic Minority Research Institute.

Schnotz, W. (2005). An integrated model of text and picture comprehension. In. R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning*. (pp. 49-69). New York: Cambridge University Press.

Schleppegrell, M. J. (2004). *The language of schooling: A functional linguistics perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Smarter Balanced Assessment Consortium (2013): *Student Interface Practice and Training Tests:* English Language Arts, Grade 3, Item No. 18. *https://login4.cloud1.tds.airast.org/student/V42/Pages/LoginShell.aspx?c=SBAC_PT&v=42.* Retrieved, June 1, 2014.

Smarter Balanced Assessment Consortium (2013): *Student Interface Practice and Training Tests:* Mathematics, Grade 3, Item No. 9. *https://login4.cloud1.tds.airast.org/student/V42/Pages/LoginShell.aspx?c=SBAC_PT&v=42.* Retrieved, June 1, 2014.

Smarter Balanced Assessment Consortium (2013): *Student Interface Practice and Training Tests:* Mathematics, Grade 5, Item No. 16. *https://login4.cloud1.tds.airast.org/student/V42/Pages/LoginShell.aspx?c=SBAC_PT&v=42.* Retrieved, June 1, 2014.

Smarter Balanced Assessment Consortium (2014). Usability, accessibility, and accommodations guidelines. Prepared with the assistance of the National Center on Educational Outcomes. August 1, 2014. Retrieved June 1, 2014. http://www.smarterbalanced.org/wordpress/wp-content/uploads/2014/03/SmarterBalanced_Guidelines_091113.pdf

Smith, R. (2012). Distinct word length frequencies: distribution and symbol entropies. *Glossometrics, 23,* 7-22.

Solano-Flores, G. (2006). Language, dialect, and register: Sociolinguistics and the estimation of measurement error in the testing of English language learners. *Teachers College Record*, *108*(11), 2354–2379. doi:10.1111/j.1467-9620.2006.00785.x

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

Solano-Flores (In press). Probabilistic approaches to examining linguistic features of test items and their effect on the performance of English language learners. *Applied Measurement in Education*, special issue: *Levels of Analysis in the Assessment of Linguistic Minorities*.

Solano-Flores, G., Backhoff, E., & Contreras-Niño, L. Á. (2009). Theory of test translation error. *International Journal of Testing*, *9*(2), 78–91. doi:10.1080/15305050902880835

Solano-Flores, G., Contreras-Niño, L.A., & Backhoff, E. (2013). The measurement of translation error in PISA-2006 items: An application of the theory of test translation error. In Prenzel, M., Kobarg, M., Schöps, K., & Rönnebeck, S. (Eds.), *Research in the Context of the Programme for International Student Assessment* (pp. 71-85). Springer Verlag.

Solano-Flores, G., and Gustafson, M. (2013). Assessment of English language learners: A critical, probabilistic, systemic view. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *Improving Large Scale Assessment in Education: Theory, Issues, and Practice.* (pp. 87-109). Taylor & Francis: Routledge.

Solano-Flores, G., & Li, M. (2009). Generalizability of cognitive interview-based measures across cultural groups. *Educational Measurement: Issues and Practice, 28* (2), 9-18.

Solano-Flores, G., & Li, M. (2013). Generalizability theory and the fair and valid assessment of linguistic minorities. *Educational Research and Evaluation*, 19(2-3), 245-263.

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher, 32*(2), 3-13.

Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2002). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing, 2*(2), 107-129.

Solano-Flores, G., Wang, C. Kachchaf, R., Soltero-Gonzalez, L., & Nguyen-Le, K. (In Press). Developing testing accommodations for English language learners: Illustrations as visual supports for item accessibility. *Educational Assessment*.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*, 295-312.

Sweller, J., Van Merriënboer, J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review* 1(3): 251–296.

Valdés, G., & Figueroa, R. A. (1994). *Bilingualism and testing: A special case of bias*. Norwood, NJ: Ablex.

Valdés, G., Quinn, H., & Lee, O. (2013). Science and language for English language learners in relation to Next Generation science Standards and with implications for Common Core State Standards for English Language arts and Mathematics. *Educational Researcher, 42*(4), 223-233.

van Lier, L. (2004). *The ecology and semiotics of language learning: A sociocultural perspective*. Dordrecht: Kluwer Academic Publishers.

Wardhaugh, R. (2002). *An introduction to sociolinguistics* (4th ed.). Oxford, UK: Blackwell Publishing.

Wellington, J., & Osborne, J. (2001). *Language and literacy in science education*. Buckingham, UK: Open University Press.

Wolfram, W., Adger, C. T., & Christian, D. (1999). *Dialects in schools and communities*. Mahwah, NJ: Lawrence Erlbaum Associates.