Educational Assessment, 20:95–111, 2015

ISSN: 1062-7197 print/1532-6977 online
DOI: 10.1080/10627197.2015.1028622

Routledge
Taylor & Francis Group

Construct Relevant or Irrelevant? The Role of Linguistic Complexity in the Assessment of English Language Learners' Science Knowledge

Brianna Avenia-Tapper and Lorena Llosa
New York University

This article addresses the issue of language-related construct-irrelevant variance on content area tests from the perspective of systemic functional linguistics. We propose that the construct relevance of language used in content area assessments, and consequent claims of construct-irrelevant variance and bias, should be determined according to the degree of correspondence between language use in the assessment and language use in the educational contexts in which the content is learned and used. This can be accomplished by matching the linguistic features of an assessment and the linguistic features of the domain in which the assessment is measuring achievement. This represents a departure from previous work on the assessment of English language learners' content knowledge that has assumed complex linguistic features are a source of construct irrelevant variance by virtue of their complexity.

There is overwhelming evidence in the literature that English language learners (ELLs) as a group score lower on content area standardized tests than their non-ELL counterparts (Kieffer, Lesaux, Rivera, & Francis, 2009). Kieffer et al. (2009) reported that the mean effect size of the difference in science achievement scores across 11 studies was high ($d = .748$). However, the cause of these score gaps, and thus how they should be interpreted, continue to be the topic of much debate. Abedi (2007) argued that "there is no evidence to suggest that these students [ELLs] have less ability to learn content knowledge than non-ELL students. Therefore, nuisance variables such as linguistic and cultural biases may mainly be responsible for such performance gaps" (p. 11).

Solano-Flores and Li (2013) also argued that "ELLs may possess the [content] knowledge being assessed but cannot demonstrate it in L2" (p. 250). Based on this perspective, much work in this area argues that content area tests may be biased against ELLs because the language of these tests unfairly adds an irrelevant hurdle for ELLs, causing construct-irrelevant variance in ELLs' scores (e.g., Abedi, 2007; Abedi et al., 2000/2005). Empirical arguments for this perspective include studies that link the presence of complex linguistic features in test items to greater relative difficulty of the items for ELLs (e.g., Martiniello, 2009; Wolf & Leon, 2009), and it is this perspective that has motivated the use of linguistically simplified tests for ELLs as a

Correspondence should be sent to Brianna Avenia-Tapper, Department of Teaching and Learning, New York University, 239 Greene Street, 6th Floor, New York, NY 10003. E-mail: bat269@nyu.edu

test accommodation (Abedi et al., 2000/2005; Abedi, Hofstetter, Baker, & Lord, 2001; Abedi, Hofstetter, & Lord, 2004; Abedi, Lord, & Hofstetter, 1998; Lotherington-Woloszyn, 1993; Rivera & Stansfield, 2001).

Kieffer et al. (2009) also consider the possibility that the performance gap between ELLs and non-ELLs is caused by different levels of linguistic ability, but in contrast to Abedi (2007), they hypothesize that the performance gaps are caused by differences in proficiency with linguistic skills that may be *necessary* for content area achievement. Kieffer et al. (2009) stated that "necessary academic language skills play a greater role in observed achievement differences between ELLs and non-ELLs than do irrelevant language skills" (p. 1190).

In this article, we draw from systemic functional linguistics to argue that linguistic features, whether or not they are complex, can constitute important symbolic tools for making disciplinary meaning, and thus complexity is not sufficient justification for the assertion that a specific linguistic feature is irrelevant to the science achievement construct. We suggest an alternative approach to evaluating the construct relevance of specific linguistic features on content area tests based on the correspondence between the language of the test and the language of what Bachman and Palmer (1996, 2010) called the target language use (TLU) domain. In the case of science tests, this refers to grade-level science talk and text. We assume that all of the linguistic features commonly used in science discourse at a particular grade level form the set of linguistic features in the TLU domain. This would include, for example, the rich array of linguistic resources that children and teachers use to make meaning in science, as well as the linguistic features common in published grade-level science text. Examining specific linguistic forms using this alternative approach allows us to judge the construct relevance of these linguistic features in content area tests to better evaluate differential item functioning (DIF)-based arguments of test bias against ELLs and to determine whether the absence of certain linguistic features of grade-level talk and text puts the test at risk for construct-underrepresentation. Although this article demonstrates the utility of this alternative approach for science tests, the method and rationale are also applicable to other content areas.

CONSTRUCT-IRRELEVANT VARIANCE AND ITEM BIAS

Construct-irrelevant variance has been defined as the situation in which "test scores are affected by processes [or content] that are extraneous to its [the test's] intended construct" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 10). In other words, if scores are affected by a variable that is unrelated to that which the test is intended to measure, we say that there is construct-irrelevant variance. For example, if a test item designed to measure students' skill with math word problems includes reference to a refrigerator, the student's familiarity with the word *refrigerator* has the potential to affect student performance on this item. Because knowledge of the meaning of the word *refrigerator* is not a component of the assessment's target construct, if the presence of this word has an effect on the item's relative difficulty, this can be described as construct-irrelevant variance (Young, Pitoniak, King, & Ayad, 2012).

If a variable unrelated to the target construct is associated with greater relative difficulty for members of a subset of the testing population, this can cause the test to be biased against individuals in that subset. If, in the preceding example, one group of students is less likely to

know the word *refrigerator*, the difficulty of that item, relative to the difficulty of the whole assessment, will likely be higher for this group of students. In this instance we would say that the refrigerator item is biased. This type of systematic score variance can cause educators to make incorrect inferences about the abilities of students in specific groups. Construct-irrelevant variance and bias weaken the validity of inferences based on scores from an assessment.

The most commonly used approach to identify bias in assessment is DIF (Camilli & Shepard, 1994). In DIF analyses, test developers first look to see if different groups of students perform differently on specific items in relation to their overall achievement.[1] If a systematic group-related difference is found for specific items when comparing members of two groups who perform at similar levels overall, this suggests that there is a variable associated with group membership that affects student performance on the specific test item (Young et al., 2012). However, the presence of DIF does not in itself constitute proof of bias (Camilli & Shepard, 1994). Camilli and Shepard (1994) stressed that the identification of bias is a two-part process. First, group differences in item difficulty are determined. Second, presence of DIF for specific items is related to presence of construct-irrelevant challenge in those items. For DIF to provide evidence of bias, it must be related to the presence of features that have been shown to be irrelevant to the target construct.

Studies investigating potential content area test bias against ELLs have evaluated linguistic complexity in test items and related this complexity in test items to the items' DIF (e.g., M. K. Lee & Randall, 2011; Mahoney, 2008; Martiniello, 2009; Shaftel, Belton-Kocher, Glasnapp & Poggio, 2006; Wolf & Leon, 2009). These studies have reported the procedures used to determine the construct-relevance of complex lexical items (vocabulary) in the test text. However, these studies do not report employing systematic procedures for judging the construct-relevance of linguistic features at the grammatical and text levels. Thus the underlying assumption in these studies is that complex grammar and text structures constitute a construct-irrelevant variable in content assessments simply by virtue of their complexity.

LINGUISTIC COMPLEXITY

In studies of linguistic complexity found on large-scale standardized tests, linguistic complexity is typically defined by naming specific linguistic features (lexical, grammatical, text level) and implying that presence of these features, or greater reliance on these features, comprises linguistic complexity. Some studies operationalize/define lexical complexity as word length and grammatical complexity as sentence length (e.g., Mahoney, 2008; Shaftel et al., 2006). In addition, complex features are described as those that "reflect an unusual or unnecessary level of linguistic sophistication" (Bailey, 2000, p. 79) or are "academic" (e.g., Wolf & Leon, 2009). Nagy and Townsend (2012) defined academic language as "the specialized language, both oral and written, of academic settings that facilitates communication and thinking about disciplinary content" (p. 92). Snow and Uccelli (2009) noted that "academic" language is not a separate

[1]Although some measures of bias compare the relative difficulty of an item between members of two different groups who have been matched on overall achievement levels, this method does not take into account the fact that the overall achievement levels are affected by bias in individual items. Therefore, some methods determine overall ability by giving more weight to the items that show less evidence of DIF than the items that show greater evidence of DIF (see Martiniello, 2009, for a full description of this method).

category of linguistic features but instead is a relative term: More academic linguistic features are features that are more often used in written language than oral language, more often used in formal than informal language, and more often used to construct expository text than narrative text. Therefore, when linguistic features are referred to as complex in studies of language-related construct-irrelevant variance on content area tests, this means that these features are some combination of long, formal, less common, and typical of expository written text, rather than oral narrative.

Linguistic features included in the descriptions of linguistic complexity in studies of content area tests can be categorized as lexical, grammatical, and text level. These categories are often used to describe content area test language (Bailey, 2000) and to describe the components of TLU domains (Bachman & Palmer, 2010). We use "lexical features" to refer to vocabulary and "grammatical features" to refer to morphological and syntactic features, or sentence-level patterns of language use. At the text level, linguistic features can be further divided into cohesion and rhetorical organization (Bachman & Palmer, 2010). See the appendix for examples of the lexical, grammatical, and text-level features discussed next.

Lexical complexity is sometimes operationalized as the presence of ambiguous or multiple meaning words (e.g., Shaftel et al., 2006), and/or morphological density (Mahoney, 2008). Measures of familiarity (i.e., whether the word is listed on a common fourth-grade word list) are also used to determine lexical complexity (Mahoney, 2008). Studies of content area test language differentiate between complex words that can be classified as technical, or discipline specific (e.g., "photosynthesis" on a science test), and those that are considered general academic vocabulary (M. K. Lee & Randall, 2011; Martiniello, 2009; Shaftel et al., 2006; Wolf & Leon, 2009). General academic vocabulary are those words that are not among the 2,000 most common words in a language but that occur frequently in a wide range of academic texts in that language (Coxhead, 2000). Wolf and Leon (2009) defined general academic vocabulary as words that are "typically used in academic settings across multiple disciplines, such as *consequently* or *based on*" (p. 143). Both technical terms and general academic vocabulary are components of "academic language." Whereas the Nagy and Townsend (2012) definition suggests that both types of vocabulary may be relevant to the target construct on content area tests, in the studies of content area tests for ELLs, technical vocabulary is typically assumed to be a component of the target construct, and general academic vocabulary is considered to be a potential source of construct-irrelevant linguistic complexity (M. K. Lee & Randall, 2011; Martiniello, 2009; Shaftel et al., 2006; Wolf & Leon, 2009).

In terms of grammatical complexity, the most commonly noted, linguistically complex features in standardized tests are sentences with an unusually large number of words; compound sentences; complex sentences, including long noun phrases and dependent clauses (such as relative, conditional, and adverbial clauses); and passive voice (e.g., Abedi, Lord, & Plummer, 1997; Farnsworth, 2008; M. K. Lee & Randall, 2011; Mahoney, 2008; Martiniello, 2009). Bailey (2000) noted "left-branching" sentences as a particularly complex grammatical feature. Some researchers look at the combined presence of these features using more holistic "linguistic complexity rubrics," which measure grammatical complexity as the cumulative effect of multiple complex grammatical features (Martiniello, 2009; Shaftel et al., 2006).

In terms of components of linguistic complexity at the text level, previous studies have considered the relative difficulty of various cohesive devices and the relationship between words and visuals in test text. Bailey (2000) identified pronoun reference as a potential source of

linguistic complexity. Wolf and Leon (2009) explored more holistic potential text-level barriers to item comprehension. In their analysis of standardized math and science tests for students in Grades 4, 5, 7, and 8, Wolf and Leon (2009) identified "the proportion (in each item) of language to non-language, the amount of language in visuals, and the extent to which the test-taker needs language knowledge in order to answer an item" (p. 143). Previous work on content area tests for ELLs, however, typically did not define text-level linguistic complexity by identifying specific styles of rhetorical organization or "conventions for sequencing units of information" (Bachman & Palmer, 2010, p. 46) that are present in the test text. These studies have not included a "text structure," "genre," or "rhetorical organization" category for analyzing linguistic complexity to examine whether these structures may introduce language-related construct-irrelevant variance for ELLs. The potential absence on such tests of text structures, genres, or styles of rhetorical organization common in the discipline has not been investigated, leaving open the possibility of construct underrepresentation.

Rethinking the Role of Linguistic Complexity in the Assessment of ELLs' Content Knowledge

Recent reconceptualizations of science standards suggest that science educators now understand specific types of language use to be an important component of science knowledge. Specific ways of using language have been identified as a target of science education. For example, the Next Generation Science Standards (National Research Council, 2013) emphasize that science students need to enact specific linguistic practices such as argumentation and explanation. The Framework for K-12 Science Education (National Research Council, 2012) states that "the focus [of science education] … is on important practices, such as modeling, developing explanations, and engaging in critique and evaluation [argumentation], that have too often been underemphasized in the context of science education" (p. 44). If some language use is relevant to the target construct on science assessments, as implied by the Next Generation Science Standards and the Framework for K-12 Science Education, then it follows that we must find a way to evaluate the relevance of specific types of language use for specific assessments. Further, if some language use is construct relevant, then reported correlations between linguistic complexity and DIF only constitute sufficient evidence of bias against ELLs if the linguistic features in question have been systematically evaluated for (ir)relevance to the target construct. DIF evidence shows only that there are systematic between-group differences in the relative difficulty of specific test items. Correlations between DIF against ELLs and linguistic complexity show that linguistically complex items are relatively more difficult for ELLs. However, to convincingly support an argument for bias against ELLs, there must also be evidence that the complex linguistic features correlated with DIF are in fact irrelevant to the target construct.

An example may prove useful in clarifying this concept. Imagine a science test with an item designed to assess student knowledge of the stages of a frog life cycle. Students' knowledge of frog life cycles is an intended target of this assessment; knowledge of frog life cycles is a component of the construct definition of this assessment. Imagine, for the sake of the example, that children growing up in rural areas have more experience playing in ponds than do children who grow up in urban areas. Consequently, on average, children in rural areas have seen more frog eggs, tadpoles, and adult frogs than have children in urban areas. Perhaps this difference in experience is reflected in statistically significant levels of difference in difficulty (relative to the

overall test scores) for the frog life cycle items between urban and rural students (DIF against urban kids). In this case, the difference between the relative difficulty of the frog life cycle items for children in rural versus urban areas would not indicate presence of bias, because the test takers' differing performances on that item would reflect differing levels of skill with a construct-*relevant* concept. Further, if, in response to cries of item bias based on DIF, the frog life cycle items were taken out of the assessment, it might close the score gap between rural and urban populations, but it would do so by taking a construct-*relevant* element, central to the target construct, out of the test. Not only would such an action reduce the construct validity of test interpretations, it would also risk causing negative washback. In such a scenario, teachers might begin spending less instructional time on frog life cycles, because this topic is not tested, which would further disadvantage those who were scoring lower in the first place. As in this example, taking construct-*relevant* items with different relative levels of difficulty for different groups out of a national standardized assessment may be more harmful than helpful.

Drawing from systemic functional linguistics, we explain why specific lexical, grammatical, and text-level features, whether or not they are complex, and despite connections to DIF against ELLs, may be construct relevant on content area tests and thus important to include in assessment for all students. Then, drawing on the concept of a target language use domain (Bachman & Palmer, 2010), we propose an approach for determining which linguistic features are central to the target construct in content area assessments.

A SYSTEMIC FUNCTIONAL LINGUISTICS (SFL) PERSPECTIVE ON LANGUAGE AND CONTENT

SFL asserts that language is integral to academic content. In her book on the applications of SFL to school language, Schleppegrell (2004) claimed that "content knowledge and skills cannot be separated from the linguistic means through which that knowledge and skill is manifested" (p. 155). In this view, the meanings that make up the accumulated sets of information we call content area knowledge are inseparable from the symbol systems (language) that allow us to know beyond direct sensory experience. In particular, theoretical knowledge, such as that which is highly prized in science, is that which, by virtue of its abstraction, must exist within some type of symbol system.

If we accept this conception of theoretical knowledge, it follows that this knowledge does not just shape the way we use language but that the way we use language shapes our theoretical knowledge. SFL theorists assert that linguistic forms at the lexical, grammatical, and text levels have evolved in tandem with scientific ideas such that specific linguistic features at each level can be understood as useful tools specifically tailored to do the communicative work needed in science. The constellation of linguistic features that we call science discourse functions to facilitate science-style meaning making (Halliday & Martin, 1996). For example, the frequency of prepositional phrases in science language both facilitates and is necessitated by science's focus on positional relationships (Lemke, 1990). Similarly, nominalizations, which occur commonly in science text, both facilitate and are necessitated by the scientific work of looking at relationships between processes, such as cycles and correlations. Nominalizations are verbs that have been reconstrued as nouns, such as *erosion*, *evaporation*, and *growth* (e.g., Fang, 2006; Halliday & Martin, 1996; Schleppegrell, 2004). This linguistic feature allows us to represent

processes as things, and from an SFL perspective, communicative use of the nominalization allows us to construct an understanding of a process as a thing. SFL theorists suggest that using the symbolic resources common in science discourse allows students to know in ways that align with the epistemology of the science discourse community. Scott (1998) articulated this well when he wrote, "In learning to talk science we must buy into and learn to work with, the conceptual tools, epistemological framing, ontological perspectives and forms of reasoning of the scientific community" (p. 75). Schleppegrell (2004) asserted that "new ways of using language ... lead to new ways of thinking and new forms of consciousness in students" (p. 18).

The discipline-specific and functional nature of the lexical, grammatical, and text-level features of science discourse suggest that understanding these features may be relevant to the target construct on tests of science knowledge and understanding. This suggestion has important practical implications, because many of the same complex linguistic features hypothesized to be sources of construct-irrelevant variance have also been identified as functional for making scientific meaning. In the next section, we explore the implications of SFL theory for the discussion of language as a construct-irrelevant variable on content area tests.

Construct-Irrelevant Linguistic Complexity on Content Area Tests From an SFL Perspective

Grammatical features. SFL proposes that grammatical features constitute a meaning-making system and thus that specific grammatical features will have varied usefulness depending on the content area about which one is communicating. This varied usefulness will be reflected in probabilities of use that differ depending on the content area of the text, just as specific vocabulary words have different probabilities of use depending on the content area of the text (Halliday & Martin, 1996). This perspective suggests that specific grammatical structures may be construct relevant, just as technical vocabulary items may be construct relevant. Previous analyses of science test language have noted the possibility that science vocabulary is relevant to the "science knowledge" construct, but these studies have not reported systematic investigation of the possibility that specific complex grammatical features may be more or less relevant to the "science knowledge" construct (e.g., M. K. Lee & Randall, 2011; Mahoney, 2008; Martiniello, 2009; Shaftel et al., 2006; Wolf & Leon, 2009). Shaftel et al. (2006) and Martiniello (2009) do note that comparative structures (e.g., ___ is greater than ____), may be grammatical elements of content area knowledge. However, there is no discussion of the process by which this evaluation was made.

Often, there is overlap between the features included in lists of complex language used on tests and features included in functional analyses of science texts. For example, "nominalization," described previously, has been named as an element of linguistic complexity (e.g., Abedi et al., 1997) and has been hypothesized to cause construct-irrelevant linguistic challenge. However, nominalization is also theorized to facilitate expression of relationships between processes therefore facilitating expression of the sort of meanings on which science focuses (Halliday & Martin, 1996). This suggests that an ability to comprehend and produce "nominalized" forms of processes may be relevant to the science knowledge construct at some grade levels and thus may challenge the assumption that this linguistic feature introduces irrelevant complexity. At the very least, the ability to comprehend nominalized forms in text is likely to be a component of the ability to read science text at some grade levels, and if the ability

to read science text at those grade levels is a component of science content mastery, then so too is the ability to understand nominalizations.

Relevant text-level features. SFL analyses investigate "*texts* rather than sentences as the basic unit through which meaning is negotiated" (Halliday & Martin, 1996, p. 22). This suggests that particular text structures are designed to construct different types of meaning and therefore may also be either relevant or irrelevant to a target construct. For example, Halliday and Webster (2004) commented that "the structure of the experiment genre symbolizes the scientific method . . . and has evolved to enable scientists to document their research" (p. 192). This genre, which includes explicit recounting of a procedure, is necessary to convince readers of an experiment's scientific rigor and to make it possible for other scientists to replicate a study (Reeves, 2005). Both of these objectives are central to scientific work. If specific text structures are organized so as to facilitate science-style meaning making, then understanding of these structures may be important to mastery of science knowledge and practice and thus may be integral to the target construct on science tests. However, as mentioned earlier, specific text structures have not been analyzed in studies of science test language.

AN ALTERNATIVE APPROACH TO IDENTIFYING CONSTRUCT-IRRELEVANT LANGUAGE: THE TLU DOMAIN MATCHING APPROACH

If the science knowledge construct includes a degree of proficiency with science language, and linguistic forms are functional for making scientific meaning, then previous complexity-based evaluations of the construct relevance of test items may be inappropriate. Instead, these evaluations may be better made based on the test language's match to the language used in science learning. Bachman and Palmer (1996, 2010), working in language assessment, devised a method for evaluating the relevance of specific task characteristics on tests. Bachman and Palmer (1996) asserted that relevance of specific task features can be judged by matching the language and characteristics of the test to the "situation or context in which the test taker will be using the language outside the test itself" (p. 18). This context is called the TLU (Bachman & Palmer, 2010). The closer the match between the test and the TLU domain, the more confident we can be that success on the test predicts whether students will be able to successfully negotiate the TLU domain. In the case of science assessment, the TLU domain is grade-level science talk and text. We can use the degree of match (at the lexical, grammatical, and text levels) between the language of the assessment and the language of the grade-level content in order to assess the relevance of the language on the assessment.

To illustrate this approach, we present a compilation of reported science talk and text features in K-12 classrooms and compare these to the linguistic features hypothesized to be a source of construct-irrelevant variance in studies of content area test language. Our compilation of linguistic forms that have been documented as common within the TLU domain was based on a wide variety of studies that were originally designed to answer a range of questions. Some of these studies attempted to operationalize the concept of academic language in order to better inform the creation of English language proficiency tests (Bailey, Butler, La Framenta, & Ong, 2004; Butler, Bailey, Stevens, Lord, & Huang, 2004; Butler, Lord, Stevens, Borrego, & Bailey, 2004). There are also studies that have investigated science language in order to better inform science instruction (Arnold, 2012; Bruna, Vann, & Escudero, 2007; Naylor, Keogh, & Downing,

2007; Pappas, 2006; Parkinson & Adendorff, 2005). Finally, much of the evidence on features of science language was gathered by researchers applying SFL theory to science text and talk in order to explore the relationship between science content and science language and to illustrate central tenets of SFL theory (Fang, 2006; Halliday & Webster, 2004; Honig, 2010; Lemke, 1990; Schleppegrell, 2004). Despite their differing objectives, these studies are all useful for describing the linguistic features of the TLU domain, specifically science education talk and text. Therefore, as we see next, these studies both allow us to demonstrate application of the TLU domain matching approach to science assessment and provide evidence that problematizes previous assumptions that complex linguistic structures must be construct irrelevant on content area tests.

Tables 1 and 2 list lexical and grammatical features that have been identified as complex and therefore as potential sources of construct-irrelevant variance. These features are drawn from previous studies of item bias against ELLs on content area tests and linguistic modification accommodation studies. Table 1 includes the lexical and grammatical forms that have also been identified in the TLU domain. Therefore, Table 1 lists linguistic features of science tests and of the TLU domain that "match" and may be considered relevant to the target construct.

The features listed in Table 1, such as dependent clauses, morphologically complex words, long noun phrases, and comparatives, are features that have been reported as common in the TLU domain of science education. If the descriptions of the linguistic features of science talk and text compiled here provide an accurate depiction of the linguistic features of the science talk and text at the grade level to be tested, then under the TLU domain matching approach, these linguistic features can be considered construct relevant and thus do not pose a threat to the validity of the score-based interpretations of an assessment for either ELL or non-ELL students. If ELLs were to perform less well on items that include these complex yet construct-relevant features, our approach suggests that there is little evidence that this lower performance is caused by construct-irrelevant variance. Association between these features and DIF against ELLs would not be sufficient evidence of item bias. Instead, our approach would suggest that these score differences are caused by real differences in ELLs versus non-ELLs' mastery of linguistic features that are a component of the assessment's target construct (Kieffer et al., 2009). It is important to point out that understanding of the linguistic features described here is *a necessary but not sufficient* component of science mastery. We fully acknowledge that the range of language used in science classroom talk and text extends far beyond the features listed here. However, because these features are necessary, they cannot be considered construct irrelevant.

Table 2 lists lexical and grammatical features identified as potential sources of language-related construct-irrelevant variance on content area tests that have not been documented in the TLU domain.

As Table 2 demonstrates, some complex features hypothesized to cause construct-irrelevant variance in the literature on test bias against ELLs may be considered irrelevant to the target construct within the TLU domain matching approach as well. These features include interrogatives that do not begin with a question word, perfect tenses, compound sentences, and slang. Both approaches find that these features are construct irrelevant and therefore may cause construct-irrelevant variance for ELLs.

As just noted, SFL asserts that text-level structures are discipline specific and functional, a position echoed by recent national science standards (e.g., Next Generation Science Standards) that include specific text structures in their description of science-learning goals. Therefore,

TABLE 1
Complex Linguistic Features Documented in the TLU Domain

Linguistic Features Identified as Potential Sources of Construct-Irrelevant Variance	*Example From Science Text*	*Who Has Documented These Features in the TLU Domain?*
Lexical features		
General (nondiscipline specific) academic vocabulary (e.g., Shaftel et al., 2006; Wolf & Leon, 2009)	Common	Bailey et al., 2004; Butler, Bailey, et al., 2004
Ambiguous, multimeaning words (Shaftel et al., 2006)	Feet	Fang, 2006
Grammatical features		
Logical connectors (Abedi et al., 2000/2005)	Also, if you break a rock in half, it is still a solid, *although* the shape has changed.	Schleppegrell, 2004; Butler, Bailey, et al., 2004
Adverbial clauses and adverbial clause connectors (Abedi et al., 2000/2005)	*When a physical change occurs,* the substance keeps its identity.	Butler, Bailey, et al., 2004; Fang, 2006
Comparatives (Mahoney, 2008)	Solids dissolve *more quickly* in warmer liquids.	Butler, Bailey, et al., 2004; Butler, Lord, et al., 2004
Passive constructions (e.g. Abedi et al., 2000/2005; Mahoney, 2008; Wolf & Leon, 2009)	A new substance *is formed.*	Hanrahan, 2006; Jackson, Meyer, & Parkinson, 2006; Schleppegrell 2004
Long sentences (Abedi et al., 2000/2005; Mahoney, 2008; Wolf & Leon, 2009)	Mixing Kool-Aid™ with water changes the color of the water, but it is a physical change because you can separate the two substances by evaporating or boiling off the water from Kool-Aid™.	Bailey et al., 2004; Butler, Bailey, et al., 2004; Fang, 2006
Long noun phrases/ prepositional phrases (e.g., Bailey, 2000; Mahoney, 2008)	*The units for measuring length in the customary system* are inches, feet, and yards.	Butler, Bailey, et al., 2004; Fang, 2006; Lemke, 1990; Schleppegrell 2004
Relative clauses (e.g. Abedi et al., 2005; Bailey, 2000; Mahoney, 2008; Wolf & Leon, 2009)	A barometer is a tool *that we use to measure air pressure.*	Bailey et al., 2004; Fang, 2006

Sentences with multiple dependent clauses (e.g. Abedi et al., 2005; Bailey, 2000)	When a liquid reaches a temperature where it spontaneously changes into a gas, we say the liquid is boiling.	Bailey et al., 2004; Fang, 2006
Multi-morphemic words/ nominalizations/ long words (e.g., Bailey, 2000; Mahoney, 2008; Wolf & Leon, 2009)	Decomposition	Bailey et al., 2004; Butler, Bailey, et al., 2004; Butler, Lord, et al., 2004; Fang, 2006; Hanrahan, 2006; Schleppegrell, 2004
Text-level features (cohesion)		
Pronoun referents (e.g., Bailey, 2000)	Some common physical changes you have seen are freezing, boiling, melting, breaking, cutting, and bending. When you do any of *these* things to an object, you are changing what the object looks like, but you are not changing what the object is made of.	Fang, 2006; Schleppegrell, 2004

Note. Examples come from P-SELL Student Book (O. Lee & Llosa, 2011–15).

TABLE 2
Complex Linguistic Features *Not* Documented in the TLU Domain

Linguistic Features Identified as Potential Sources of Construct-Irrelevant Variance	*Example*
Grammatical features	
Interrogatives that do not begin with a question word. (Abedi, Courtney, & Leon, 2003; Abedi et al., 2000/2005)	I understand that you are hungry and tired, but do you have to whine so loudly?
Perfect tenses (Abedi et al., 2005)	I had already bought the ring, so I decided to go through with the proposal.
Compound sentences (e.g., Abedi et al., 2000/2005; Wolf & Leon, 2009)	I like books and I like magazines.
Lexical features	
Words referencing American holidays (Shaftel et al., 2006)	Christmas tree
Colloquialism or slang (Abedi et al., 2000/2005; Mahoney, 2008)	Awesome

an SFL perspective suggests that evaluating test language for construct relevance should be done not only at the lexical and grammatical levels but also at the text level. Similarly, Bachman and Palmer's (2010) method for analyzing linguistic elements of a test task requires us to investigate the rhetorical organization of the test text(s). Although studies have not analyzed the rhetorical organization (text structure) of science tests, the styles of rhetorical organization used in the texts of science learning (the TLU domain) have been examined. In science learning, students are often expected to produce and comprehend explicit procedure recounts and/or research articles (Halliday & Webster, 2004; Honig, 2010; Jackson, Meyer, & Parkinson, 2006; Reeves, 2005), arguments with claim and evidence (Lemke, 1990; National Research Council, 2012; Naylor et al., 2007; Reeves, 2005), explanations (Bailey et al., 2004; Butler, Bailey, et al., 2004; Butler, Lord, et al., 2004; Halliday & Webster, 2004; National Research Council, 2012; Schleppegrell, 2004), descriptions (Arnold, 2012; Bailey et al., 2004; Butler, Bailey, et al., 2004; Butler, Lord, et al., 2004; Honig, 2010; Schleppegrell, 1998), and comparisons (Arnold, 2012; Bailey et al., 2004; Butler, Bailey, et al., 2004; Butler, Lord, et al., 2004). Therefore, if students are not expected to make meaning from argumentation and explanation texts on science tests, this may suggest that these assessments suffer from construct underrepresentation. Likewise, inclusion of text-level styles of rhetorical organization not well documented in the TLU domain could create construct-irrelevant variance in scores on these assessments. Construct underrepresentation and/or construct irrelevance related to the inclusion or exclusion of specific text structures may pose a threat to the validity of these tests.

CONCLUSIONS AND IMPLICATIONS

The TLU domain matching approach for identifying construct-relevant and irrelevant linguistic features proposed here provides evidence that challenges previous assumptions that linguistic complexity is synonymous with language-related construct-irrelevant variance in content assessments for ELLs. Analyzing science test language using the TLU domain matching approach suggests that many "complex" grammatical structures are construct relevant because they are components of science classroom discourse and important tools for construing the world

scientifically. These structures help students to articulate theories, to appropriate paradigmatic explanations, and to enact the linguistic/social practices that constitute paradigmatic participation in the discipline. This position in no way suggests that the linguistic features listed here are the only features useful in science classrooms. There is a wide array of linguistic features used in science classroom discourse across our multilingual, multicultural society. It may be possible to express scientific ideas using "simple" grammatical features just as it may be possible to explain the concepts referenced by technical vocabulary words using nontechnical language. However, just as this possibility does not negate the relevance of technical vocabulary to content area achievement, this possibility also does not negate the relevance of complex grammatical features to content area achievement. In addition, this analysis suggests that the omission of text-level structures such as argumentation and explanation on multiple-choice science tests may suggest that these assessments suffer from construct underrepresentation.

The approach just illustrated is useful for judging the construct relevance of specific linguistic features such that we can better evaluate claims of content area test bias against ELLs. This approach could also be helpful for constructing assessments, which are less likely to include construct-irrelevant linguistic features and/or less likely to suffer from construct underrepresentation. However, the specific linguistic features identified in the demonstration of the matching approach in this article would not be appropriate for direct use in assessment design. The linguistic features of the science learning TLU domain differ depending on the grade level, and thus to be able to determine the match between test and TLU features, a complete description of the TLU domain by grade level is needed. The available literature does not yet provide detailed descriptions of the TLU domain at each grade level. In addition, these grade-level domain descriptions need to differentiate between the lexical, grammatical, and text-level features that students are expected to produce and those that they are expected to understand.

Implications

The review of the literature on science talk and text suggests that understanding and using relevant grammar and text structures such as nominalization, prepositional phrases, and argumentation constitute important science learning objectives to be taught and, in turn, to be assessed on large-scale tests. Therefore, we suggest that teachers of all students, but ELLs in particular, should create regular opportunities for meaningful exposure to and use of the grammar and text structures common in grade-appropriate science language (Kieffer et al., 2009).

The SFL perspective suggests that linguistic features at lexical, grammatical, and text levels are all functional meaning-making tools. Therefore, a systemic functional perspective on language suggests that in order for students to appropriate specific linguistic features at the lexical, grammatical, or text levels, students should be expected to engage in the communicative tasks for which those features are functional. For example, if the "experiment genre" facilitates faithful communication of a detailed procedure to others so that the procedure can then be replicated, science educators should require students to communicate detailed procedures to others with replication as a goal. If specific linguistic features are useful for meeting specific communicative goals, as is suggested by SFL, then we must engage students in the communicative tasks for which these features are functional.

Implications for Test Developers

This article argues that specific linguistic knowledge is a component of content area mastery and thus that we cannot assume linguistic features are irrelevant to the target construct on content area tests without analyzing the use of the feature in the domain to which the test is intended to generalize. This implies that definitions of the science achievement construct used in assessment design should explicitly include a description of the linguistic features that are necessary for participation in grade-level written and oral discourse.

In addition, both the SFL perspective and the TLU domain matching approach (Bachman & Palmer, 2010) emphasize the need to analyze not only lexical and grammatical elements of language but also text-level organizing structures. Bachman and Palmer (2010) suggested that analyses of test tasks and TLU domains should include description of two text-level elements of language use: rhetorical organization and cohesion. As previously noted, previous analyses of the language of content area tests have investigated cohesion, but there is very little information on the relevance or irrelevance of the rhetorical organization of texts in these tests. Work is needed to investigate the styles of rhetorical organization used in these assessments and to match this organization to that common in the TLU domain. First, this work is needed so that the relevance of the rhetorical organization currently used in assessment texts can be evaluated. Second, this work is needed to determine whether organizational styles common in the TLU domain are appropriately represented in the assessments. If text-level styles of organization such as argumentation are central to the TLU domain, but absent from content area tests, this may suggest language-related construct underrepresentation.

Finally, this article weakens DIF-based arguments for bias against ELLs in content area tests and thus problematizes "linguistic simplification" test accommodations for ELLs. The argument presented in this article implies that correlations between DIF against ELLs and linguistic complexity only constitute sufficient arguments for item bias if the complex feature related to the observed DIF is not a common feature of the domain to which test results are intended to generalize. Our argument implies that linguistic features should not be eliminated from "linguistically simplified" tests on the basis of complexity alone. Instead, the strong presence or absence of the features in the domain to which the test should generalize is a better criterion for judging the relevance of a given linguistic feature. This would prevent accommodation designers from eliminating complex structures that may be critical to science understanding and practice, thus guarding against the possibility of creating accommodated tests that suffer from construct underrepresentation, which could in turn cause negative washback for ELLs.

ACKNOWLEDGMENTS

We thank Michael Kieffer for his feedback on an earlier version of this manuscript.

REFERENCES

Abedi, J. (2007). *Language factors in the assessment of English language learners: Theory and principles underlying the linguistic modification approach.* Paper developed for the U.S. Department of Education LEP Partnership. Retrieved from http://www.ncela.gwu.edu/files/uploads/11/abedi_sato.pdf

Abedi, J., Bailey, A., Butler, A., Castellon-Wellington, M., Leon, S., & Mirocha, J. (2000/2005). *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives* (CSE Rep. No. 663). National Center for Research on Evaluation, Standards, and Student Testing (CRESST) Graduate School of Education and Information Studies, University of California, Los Angeles. Retrieved from http://files.eric.ed.gov/fulltext/ED492891.pdf

Abedi, J., Hofstetter, C., Baker, E., & Lord, C. (2001). *NAEP math performance test accommodations: Interactions with student language background* (CSE Tech. Rep. No. 536). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from http://www.cse.ucla.edu/products/reports/newTR536.pdf

Abedi, J., Hofstetter, C., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research*, *74*, 1–28. doi:10.3102/00346543074001001

Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance* (CSE Tech. Rep. No. 478). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from http://www.cse.ucla.edu/products/Reports/TECH478.pdf

Abedi, J., Lord, C., & Plummer, J. (1997). *Final report of language background as a variable in NAEP mathematics performance* (CSE Tech. Rep. No. 429). Los Angeles: University of California, Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from https://www.cse.ucla.edu/products/reports/TECH429.pdf

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

Arnold, J. (2012). Science students' classroom discourse: Tasha's umwelt. *Research in Science Education*, *42*, 233, doi:10.1007/s11165-010-9195-0

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, UK: Oxford University Press.

Bailey, A. (2000). Language analysis of standardized achievement tests: Considerations in the assessment of English language learners. In *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives* (pp. 85–105). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from http://files.eric.ed.gov/fulltext/ED492891.pdf

Bailey, A. L., Butler, F. A., LaFramenta, C., & Ong, C. (2004). *Towards the characterization of academic language in upper elementary science classrooms*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from http://www.cse.ucla.edu/products/reports/r621.pdf

Bruna, K. R., Vann, R., & Escudero, M. P. (2007). What's language got to do with it? A case study of academic language instruction in a high school? English learner science class. *Journal of English for Academic Purposes*, *6*, 36–54. doi:10.1016/j.jeap.2006.11.006

Butler, F. A., Bailey, A. L., Stevens, R., Lord, C., & Huang, B. (2004). *Academic English in fifth-grade mathematics, science, and social studies textbooks*. Los Angeles, CA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from http://people.ucsc.edu/~ktellez/academicEnglish5th.pdf

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 mathematics*. Los Angeles, CA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing. https://www.cse.ucla.edu/products/reports/R626.pdf

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213–238. doi:10.2307/3587951

Fang, Z. (2006). The language demands of science reading in middle school. *International Journal of Science Education*, *28*, 491–520. doi:10.1080/09500690500339092

Farnsworth, T. L. (2008). *The construct of academic English in tests of academic achievement and its effect on student performance: A confirmatory factor analytic study* (Unpublished doctoral dissertation). University of California, Los Angeles.

Halliday, M. A. K., & Martin, J. R. (1996). *Writing science: Literacy and discursive power*. London, UK: Routledge Falmer.

Halliday, M. A. K., & Webster, J. (2004). *The language of science*. London, UK: Continuum.

Hanrahan, M. (2006). Highlighting hybridity: A critical discourse analysis of teacher talk in science classrooms. *Science Education*, *90*, 8–43. doi:10.1002/sce.20087

Honig, S. (2010). What do children write in science? A study of the genre set in a primary science classroom. *Written Communication*, *27*, 87–119. doi:10.1177/0741088309350159

Jackson, L., Meyer, W., & Parkinson, J. (2006). A study of the writing tasks and reading assigned to undergraduate science students at a South African university. *English for Specific Purposes*, *25*, 260–281. doi:10.1016/j.esp.2005.04.003

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

Lee, M. K., & Randall, J. (2011). *Exploring language as a source of DIF in a math test for English language learners* (NERA Conference Proceedings 2011, Paper 20). Retrieved from http://digitalcommons.uconn.edu/nera_2011/20

Lee, O., & Llosa, L. (2011–15). *Promoting science among English language learners (P-SELL) scale-up* (Discovery Research K-12; NSF Grant DRL 1209309). New York University, New York.

Lemke, J. L. (1990). *Talking science: Language, learning, and values*. Norwood, NJ: Ablex.

Lotherington-Woloszyn, H. (1993). Do simplified texts simplify language comprehension for ESL learners? In M. L. Tickoo (Ed.), *Simplification: Theory and application* (Anthology Series 31; pp. 140–154). Singapore: SEAMEO Regional Language Centre.

Mahoney, K. (2008). Linguistic influences on differential item functioning for second language learners on the National Assessment of Educational Progress. *International Journal of Testing*, *8*, 14–33. doi:10.1080/15305050701808615

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational Assessment*, *14*, 160–179. doi: 1080/10627190903422906

Nagy, W., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly*, *47*, 91–108.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press. Retrieved from http://www.nap.edu/openbook.php?record_id=13165&page=1

National Research Council. (2013). *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press.

Naylor, S., Keogh, B., & Downing, B. (2007). Argumentation and primary science. *Research in Science Education*, *37*, 17–39. doi:10.1007/s11165-005-9002-5

Pappas, C. C. (2006). The information book genre: Its role in integrated science literacy research and practice. *Reading Research Quarterly*, *41*, 226–250. doi:10.1598/RRQ.41.2.4

Parkinson, J., & Adendorff, R. (2005). Science books for children as a preparation for textbook literacy. *Discourse Studies*, *7*, 213–232. doi:10.1177/1461445605050367

Reeves, C. (2005). *The language of science*. New York, NY: Routledge.

Rivera, C., & Stansfield, C. W. (2001). *The effects of linguistic simplification of science test items on performance of limited English proficient and monolingual English-speaking students*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

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

Scott, P. (1998). Teacher talk and meaning making in science classrooms: A Vygotskian analysis and review. *Studies in Science Education*, *32*, 45, doi:10.1080/03057269808560127

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

Snow, C. E., & Uccelli, P. (2009). The challenge of academic language. In D. R. Olson & N. Torrance (Eds.), *The Cambridge handbook of literacy* (pp. 112–133). New York, NY: Cambridge University Press.

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

Wolf, M., & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educational Assessment*, *14*, 139–159. doi:10.1080/10627190903425883

Young, J., Pitoniak, M., King, T., & Ayad, E. (2012). *Guidelines for accessibility for English language learners* (Smarter Balanced Assessment Consortium, Measured Progress/ETS Collaborative). Retrieved from http://www.smarterbalanced.org/

APPENDIX

TABLE A1
Linguistic Features and Examples

Linguistic Features	*Examples*
Lexical	
Technical or discipline-specific vocabulary	In science, words such as "magma," "chrysalis," and "larva."
General academic vocabulary	"classify," "alphabetize," "paragraph"
Ambiguous, multimeaning words	"variable," "matter," "force"
Morphologically dense words	"insoluble," "precipitation," "deforestation"
Grammatical	
Compound sentences	"We held the control variables constant, and we manipulated the test variable." "Girls" scores were higher in the treatment condition, but boys" scores remained the same."
Long noun phrases	"The rate of change," "the ten milliliter beaker," "the amount of liquid in the glass"
Complex sentences with dependent conditional clause	"*If we raise the temperature*, the rate of change increases." "*If we apply greater force,* the object will move farther."
Complex sentences with dependent adverbial clause	*"When the temperature drops to zero,* water freezes." *"Although whales live in the water,* they are classified as mammals."
Complex sentences with dependent relative clause	"The plant *that received direct sunlight* lived longer." "Animals *that live in the tundra* have adapted to a cold climate."
Passive voice	*"* In the control condition *plants were watered* once a day." "Many lakes *have been polluted*."
Left-branching sentences	"*Unlike the New England salamander,* the hellbender has five toes on its rear feet."
A left-branching sentence includes information before the subject of the sentence.	*"With windspeeds up to 381 miles per hour,* a F5 tornado can turn cars into missiles."
Textual features (cohesion)	
Pronoun reference	"Substances *can change from one state of matter to another* by heating or cooling. *This* is called a physical change." "The common barron caterpillar *looks very similar to the leaves where it lives*. *This* allows the caterpillar to avoid being seen by predators. Students must infer to which idea the pronoun "this" refers.