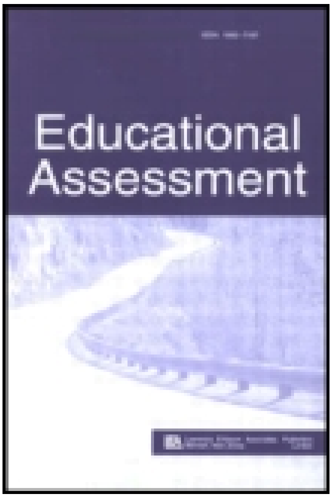

Educational Assessment
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/heda20

Semiotic Structure and Meaning Making: The Performance of English Language Learners on Mathematics Tests

Guillermo Solano-Flores [a] , Carne Barnett-Clarke [b] & Rachel R. Kachchaf [c]
[a] University of Colorado at Boulder
[b] WestEd
[c] TERC
Published online: 14 Aug 2013.

Educational Assessment, 18:147–161, 2013

ISSN: 1062-7197 print/1532-6977 online
DOI: 10.1080/10627197.2013.814515

Routledge
Taylor & Francis Group

Semiotic Structure and Meaning Making: The Performance of English Language Learners on Mathematics Tests

Guillermo Solano-Flores
University of Colorado at Boulder

Carne Barnett-Clarke
WestEd

Rachel R. Kachchaf
TERC

We examined the performance of English language learners (ELLs) and non-ELLs on Grade 4 and Grade 5 mathematics content knowledge (CK) and academic language (AL) tests. CK and AL items had different semiotic loads (numbers of different types of semiotic features) and different semiotic structures (relative frequencies of different semiotic modalities). For both linguistic groups, the percentage of items correct was higher for the AL than the CK tests. However, the score gains attributable to instruction were smaller for the AL than the CK tests. CK and AL test scores correlated more highly for non-ELLs than ELLs before instruction. This suggests that, before instruction, the meaning-making system of ELLs was less consolidated than that of non-ELLs, who dealt with the interpretive demands of CK and AL items with similar effectiveness. We discuss the importance of using a semiotic perspective in test design and in interpreting ELL student performance.

Limited proficiency in the language in which tests are administered is a major threat to the validity of measures of academic achievement for English language learners (ELLs; see, e.g., Abedi, 2004; Durán, 2011; Sireci, Han, & Wells, 2008; Valdés & Figueroa, 1994)—students in the United States who are developing English as a second language in a predominantly English-speaking society or a predominantly English-speaking school context.[1] ELL students have to learn the content of the curriculum in English and develop English as a second language at the same time (Abedi, 2011).

[1]For the purposes of this article, we prefer to use a definition that focuses on the linguistic and social aspects of English proficiency that are relevant to learning and testing. This approach is different from the legal definition of ELLs provided by the No Child Left Behind Act (2001), and which depends on the use of measures of English proficiency whose validity has been questioned (see Abedi, 2007).

Correspondence should be sent to Guillermo Solano-Flores, School of Education, University of Colorado at Boulder, 249 UCB, Boulder, CO 80309. E-mail: guillermo.solano@colorado.edu

The mandatory inclusion of ELL students in large-scale assessment programs in the United States has motivated a wide variety of efforts intended to analyze and minimize language-related factors as a source of test invalidity. These efforts focus on the linguistic complexity of test items (e.g., Martiniello, 2008; Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006) and on academic language—the form of language used in the disciplines and in the instructional contexts of those disciplines (Bailey & Butler, 2003; Halliday, 1978; Scarcella, 2003)—as critical to the effective instruction and valid assessment of ELLs (see Butler, Lord, Stevens, Borrego, & Bailey, 2004; Cummins, 1981; Guerrero, 2004; Hakuta, Butler, & Witt, 2000; Stevens, Butler, & Castellon-Wellington, 2000).

In this article, we use a semiotic perspective to examine the performance of ELLs on mathematics tests. According to this semiotic perspective, meaning in the disciplines is conveyed through multiple ways of representing information (see Gutiérrez, Sengupta-Irving, & Dieckmann, 2010; Kress & van Leeuwen, 2001; Lemke, 1998; O'Halloran, 2011; Moschkovich, 2007; Schleppegrell, 2010; Solano-Flores, 2010). Among others, these forms of representation include features such as grammatical forms and words (see Bailey & Butler, 2003; Chamot & O'Malley, 1994; Halliday, 1978; Scarcella, 2003); discursive forms (e.g., ways of expressing disagreement, ways of posing problems; Adger, Snow, & Christian, 2002; Hawkins, 2004; Solano-Flores, 2006, 2010); and visual devices such as symbols, diagrams, equations, and graphs (Lemke, 2003).

Underlying students' ability to gain access to the content of items is a *meaning-making system*, a term that we use in this article to refer to the set of interpretive resources students use in combination to integrate the information represented through multiple semiotic features according to discipline-specific conventions. Because academic language is specific to the context of a discipline (see Halliday, 1978), the notion of meaning making is used here only with reference to the context of instruction and assessment of that discipline. We view meaning making as shaped by culture, proficiency in the language of testing, first-language influences, socioeconomic factors, and opportunity to learn (Solano-Flores & Nelson-Barber, 2001). For example, lack of familiarity with the contextual information used in mathematics items may cause students to make sense of items in ways not intended by their developers (see Trumbull & Solano-Flores, 2011).

In the investigation reported here, we examined the relationship between the semiotic features of mathematics items and the performance of ELL students on those items. Our first research question was

RQ1: Can test items on a given content area be distinguished by virtue of their semiotic features?

To answer this question, we compared the semiotic features of mathematics items created to assess content knowledge and items created to assess academic language.

Then we asked,

RQ2: How is the performance of ELL and non-ELL students different or similar on content knowledge and academic language items?

To answer this question, we examined the performance of ELL and non-ELL students on the two types of mathematics items.

Because academic language encodes the knowledge of a discipline, we did not expect content knowledge and academic language items to have perfectly distinguishable sets of features. Rather, we expected to observe different semiotic structures—different relative frequencies of semiotic features. Also, because language is an important part of the interpretive resources that conform an individual's meaning-making system, we expected the two types of items to pose different sets of challenges to students from the two linguistic groups.

METHODS

Test Materials, Participants, and Data

The data and test materials used in this study originated from a broader study that evaluated the effectiveness of Math Pathways & Pitfalls (Barnett-Clarke & Ramírez, 2010), a supplementary mathematics instructional program that focuses on rational number concepts in Grades 4 to 7 and pays special attention to ELLs (Heller, Hanson, Barnett-Clarke, & Darling, 2010). In this broader study, students at each grade level experienced 15 hr of instruction using Math Pathways & Pitfalls lessons in place of 15 hr of their regular mathematics curriculum.

Math Pathways & Pitfalls emphasizes concepts identified as most difficult in the literature on mathematics learning and addresses errors typically made by students when they learn these concepts (Behr, Lesh, Post, & Silver, 1983; Carraher, 1996; Moss & Case, 1999; Parker & Leinhardt, 1995; Wearne & Hiebert, 1989). The program promotes basic mathematics academic language-related skills such as translating across representational forms, knowing the meaning of mathematical vocabulary, distinguishing between formal and informal use of mathematical vocabulary, distilling core mathematical meaning from verbal or written communication, and communicating through the use of mathematical language (Barnett-Clarke & Ramírez, 2006). Grade 4 and Grade 5 lessons for this study focused respectively on the meaning of and operations with fractions and decimals (Heller et al., 2010).

We used 132 four-option, multiple-choice items from the content knowledge and academic language mathematics tests for Grades 4 and 5 that were part of the instruments created for the broader study (Heller, Curtis, Rabe-Hesketh, & Verboncoeur, 2007). These tests were administered before and after instruction in the year 2006 and were developed in alignment with the mathematical content and academic language goals of the lessons (Barnett-Clarke & Ramírez, 2006).

All items had similar appearances and formats. On average, the number of words in an item was higher for the AL items (about 31 and 29, respectively for Grade 4 and Grade 5 items) than CK items (about 15 and 19, respectively for Grade 4 and Grade 5 items).[2] The

[2]We use this very rough indicator of linguistic complexity only to supplement the interpretation of results of this study. We did not attempt to measure linguistic complexity based on grammatical structure, use of passive voice, and other features. To produce dependable indicators of linguistic complexity, large amounts of text need to be analyzed (see Bruce, Rubin, & Starr, 1981; Harrison, 1999)—a condition that is difficult to meet with relatively small numbers of items and with items that have few words. In addition to having small amounts of text, the items used in this investigation contained multiple features (e.g., symbols, graphs) that would be difficult to characterize as text and analyze according to traditional approaches to examining linguistic complexity. In addition, unlike the items used in recent investigations that have examined linguistic complexity and the performance of ELLs on tests (e.g., Martiniello, 2008, 2009; Shaftel et al., 2006; Wolf & Leon, 2009), the items used in this investigation were not all word problems.

(a)

Content Knowledge

A puppy weighed 2.48 kilograms in June and 5.72 kilograms in November. How many fewer kilograms did the puppy weigh in June than November?

A 3.12 kilograms
B 3.24 kilograms
C 3.28 kilograms
D 3.42 kilograms

Academic Language

The number $8\frac{7}{9}$ is called ____________ because it has a whole number and a fraction.

A an improper fraction
B a mixed number
C a lowest terms fraction
D a proper fraction

(b)

Notation-Decimals: *2.48, 5.72, 3.12, 3.24, 3.28, 3.42*
Natural/Mathematical Language-Measurement units: *kilograms*
Natural/Mathematical Language-Relations: *fewer*
Testing Register-Comparative phrases: *How many fewer kilograms ...?*

Notation-Fractions: $8\frac{7}{9}$
Mathematics Register-Types of fractions: *improper fraction, mixed number, lowest term fraction, proper fraction*
Mathematics Register-Types of numbers: *whole number*
Testing Register-Cloze questions: "...is called ______________ because..."

FIGURE 1 (a) Examples of the items used in the investigation. Both items are mock-up items with similar content, complexity, and structure as the items used in the investigation. (b) Semiotic modalities and features identified. The semiotic load (number of different types of semiotic features) in both items is 4.

content knowledge items focused more on computation and problem solving, whereas the academic language items focused more on the terms used to refer to mathematical concepts (Figure 1a). The content knowledge tests for both Grade 4 and Grade 5 comprised 35 items each (respectively, a Cronbach's alpha of .81 and .88); the academic language tests for Grade 4 and Grade 5 comprised, respectively, 32 items and 30 items (respectively a Cronbach's alpha of .81 and .84; Heller et al., 2010).

We analyzed data on the students' performance on the items from 1,343 students (722 from Grade 4 and 621 from Grade 5) from inner-city schools with a Latino population of 50% or higher. Of these 1,343 students, 718 were classified as ELLs and 625 as non-ELLs by their school districts.

Semiotic Features and Semiotic Modalities

For the purpose of our study, we used semiotic feature as our basic analytical unit. From examining the 132 items, we identified 32 types of semiotic features, which we grouped into five semiotic modalities (Table 1). Needless to say, these types of semiotic features and modalities are not universal; they are specific to the set of items used in this investigation. In what follows, we define the semiotic modalities and discuss their relevance in ELL testing.

TABLE 1
Types of Semiotic Features by Semiotic Modality

Notation
Decimals (*0.1*)
Fractions ($4\frac{1}{2}$, $\frac{1}{8}$)
Letters representing variables and values (*A, x*)
Measurement units (*cm*)
Operations ($+$, $-$, */*, *x*)
Relations ($=$, $\neq$, $>$, $<$)
Special cases of numbers (-3)
Whole numbers (*182, 0*)
Mathematics Register
Number concepts (*place value, number sentence*)
Parts of a fraction (*numerator, least common denominator*)
Parts of a number (*unit, tens, hundreds*)
Types of fractions (*proper fraction*)
Types of numbers (*whole number, integer*)
Natural/Mathematical Language
Fractions (*four and a half, one fourth, hundredths*)
Geometric shapes (*circle, square*)
Measurement units (*centimeters, minutes, miles*)
Numbers (*zero*)
Operations (*sum*)
Relations (*equivalent to, equal to, greatest*)
Testing Register
Cloze questions[a]
Comparative phrases (*How many more [. . .] than [. . .]?*)
Noun phrases (*greatest fractional part, lowest terms fraction, equivalent decimal number*)
Nominalization (*identification*)
Question phrases (*Which of the following*)
Adjective phrases (*best represents, not given, best way)*
Visual Representation
Charts
Geometric shapes (circles, rectangles)
Number lines
Number sentences
Points and lines
Proportions (circles with shaded areas showing proportions or part-whole relationships)
Venn diagrams

Note. Examples are in parentheses.
[a]Questions in the form of a statement with missing words that the student is required to replace.

Notation. Notations are signs and symbols used in mathematics to express variables, magnitude, precision, proportion, units, operations, and relationships. Personal, everyday life experience shapes how students make sense of notation conventions (Pirie, 1988). An example is interpreting $+$ as equivalent to *and*. In addition, notation conventions may vary across cultures, as is the case of the decimal point, which in some cultures is represented as a comma (see Solano-Flores, 2011).

Mathematics register. Mathematics register are terms that are specific to the field of mathematics and refer to mathematical concepts, including expressions and syntactical structures used frequently in the context of mathematics. The meaning encoded by mathematics register is mediated by culture and context within a discipline. Thus, *square* may refer to the same thing in natural language and in the context of mathematics (e.g., as a geometric shape). At the same time, within the context of mathematics, *square* may have different meanings depending on the topic at hand (e.g., geometry or algebra). Furthermore, within the same topic (say, geometry), *square* may have different meanings (as either a type of rectangle or a category of polygons different from the category, *rectangle*), depending on the culture in which mathematics is taught (see Solano-Flores, 2011).

Natural/Mathematical language. Natural/Mathematical language comprises mathematical terms that are also part of everyday language and may have same or different meanings across contexts. Natural language and academic language are conceptually distinguishable but overlapping forms of language (Aukerman, 2007; Barwell, 2005; Brenner, 1998; Khisty, 1995; Lampert & Cobb, 2003; Pimm, 1987; Rothery & Shuard, 1980; Webb, 1991; Wellington & Osborne, 2001). From a probabilistic perspective of language (see Solano-Flores, 2008), a given linguistic feature is part of the natural or academic language (or both) depending on the frequency of use in different contexts and the extent of its shared meaning in those contexts (see Bod, Hay, & Jannedy, 2003; Halliday, 1978, 2005; Solano-Flores & Gustafson, 2013).

Testing register. Testing register is terms and discursive structures of high frequency in mathematics tests. The language used in the tests of specific content areas has characteristics that are not always present in typical classroom discussions or daily conversations (Abedi & Lord, 2001). Successful test takers are familiar with the register of the content area being assessed and deal effectively with linguistic challenges that are specific to the language used in tests (e.g., a phrase such as, *None of the above* or an incomplete sentence followed by phrases, which are, respectively, the stem and the options of a multiple-choice item; Solano-Flores, 2006).

Visual representation. Visual representation is nontextual representations of shapes, position, and mathematical ideas. Different forms of visual representations in mathematics (e.g., grids or pie charts with shaded areas representing proportions) may pose different sets of challenges to students. Visual and textual forms of representation of information are viewed as interacting components (e.g., Kopriva, 2008; Kress & van Leeuwen, 2001; Solano-Flores & Wang, 2011; Wang & Solano-Flores, 2011). With few exceptions (e.g., Kachchaf, 2011), little research has been conducted to examine ELLs' interpretations of images. Although various forms of visual representation are often seen as resources that support ELLs' understanding in test items (Fichtner, Peitzman, & Sasser, 1994; O'Malley & Valdez-Pierce, 1996), there is evidence that the presence of an image can increase the cognitive load of a task (see Mayer, 2005). Also, personal, cultural, and first language may influence the way in which individuals interpret different forms of visual representations of information (see Abedi & Lord, 2001; Mayer & Sims, 1994; Nisbett & Miyamoto, 2005; Pirie, 1988; Solano-Flores, 2011).

Coding Procedure

The presence or absence of each of the 32 semiotic features was coded dichotomously (1–0) for each item. According to this system, Operations (see Table 1) was coded 1 if a given sign (e.g., $+$, $-$, /, $\times$) was observed, regardless of whether that sign appeared more than once and regardless of whether different signs were observed in the item. This coding approach (illustrated in Figure 1) has proven to be both efficient and sensitive in the review of linguistic features of science and mathematics items (see Solano-Flores, Backhoff, & Contreras Niño, 2009).

Two individuals participated as coders. As part of the training, project staff discussed with the coders each type of semiotic feature and provided sample multiple-choice mathematics items with semiotic features that the coders coded independently with the purpose of identifying coding discrepancies. These discrepancies were discussed and resolved to ensure that the coders interpreted the different semiotic features consistently.

The items were given to the coders on separate sheets that concealed any information that could influence their judgments, such as grade or type of test. The coders coded the items independently and in different, random sequences. Upon completion of independent coding, the coders compared their coding and discussed, documented, and resolved their discrepancies. Then they produced a consensus version of their coding.

Data Analysis

First, we compared the content knowledge and academic language items as to their *semiotic load*—the number of different types of semiotic features observed in each item (a maximum of 32, as this is the number of types of semiotic features identified in the pool of items). For example, the semiotic load of the two items shown in Figure 1 is 4.

Second, we examined the *semiotic structure* of the content knowledge and academic language items—the relative frequencies of semiotic features belonging to the five semiotic modalities. Then, we examined the score gap and score gain differences between ELL and non-ELL students in the pre- and posttest scores. We also examined any differences between the two linguistic groups in the patterns of correlations between the content knowledge and academic language scores.

RESULTS

Intercoder Agreement

A high Cohen's coefficient for categorical variables was obtained ($\kappa = .89$), indicating that the independent coding decisions made by the coders were highly consistent. An examination of coding discrepancies revealed certain differences in the ways in which the coders interpreted some types of semiotic features. These discrepancies were easily resolved by the coders to create a consensus record of their coding. In our analyses, we used this consensus version, which contained 607 positive coding decisions distributed across the 132 items.

TABLE 2
Mean Item Semiotic Load of Content Knowledge and Academic Language Items by Grade

Target Construct	*Grade*	*n*	*M*	*SD*
Content knowledge	4	35	4.03	1.071
	5	35	4.03	1.403
Academic language	4	32	5.34	1.450
	5	30	5.13	1.665

Differences Between Content Knowledge and Academic Language Items

Semiotic load. We examined the statistical significance of semiotic load differences between content knowledge (CK) and academic language (AL) items. We observed higher semiotic loads among AL items than among CK items for both grades (Table 2). A two-way, Type of Item (content knowledge and academic language) × Grade (Grade 4 and Grade 5) analysis of variance revealed statistically significant differences in semiotic load due to type of item ($p = .000$; partial $\eta^2 = .161$) and no statistically significant differences due to grade or the interaction of type of item and grade. These results indicate that AL items had a greater semiotic load than CK items.

Semiotic structure. We examined the relative frequencies of different types of semiotic features. As shown in Table 3, for Grade 4, semiotic features belonging to the Mathematics Register, Natural/Mathematical Language, and Testing Register categories were proportionally more frequent among AL items (respectively, 23%, 29%, and 16%) than among CK items (respectively, 19%, 13%, and 6%). In contrast, semiotic features belonging to the Notation and Visual Representation categories were proportionally more frequent among CK items (respectively, 47% and 15%) than among AL items (respectively, 26% and 5%). A similar pattern of differences was observed among Grade 5 items. A series of independent-samples *t* tests performed within each grade revealed statistically significant differences in the number

TABLE 3
Semiotic Structure of Mathematics Content Knowledge and Academic Language Items: Relative Frequencies (Rounded Percentages) of Types of Semiotic Features by Semiotic Modality and Grade

		Semiotic Modality				
Type of Item	*Grade*	*Notation*	*Mathematics Register*	*Natural/Mathematical Language*	*Testing Register*	*Visual Representation*
Content knowledge	4	47	19	13	6	15
	5	61	6	13	6	14
Academic language	4	26	23	29	16	5
	5	31	18	30	17	4

Note. Percentages across semiotic modalities may not add up to 100 due to rounding.

of semiotic components for each category across CK and AL items ($p = .000$). These results indicate that CK and AL items had different semiotic structures—they were distinguishable by virtue of the relative frequencies of their semiotic features.

In summary, the answer to our first research question (Can test items on a given content area be distinguished by virtue of their semiotic features?) is affirmative. Content knowledge and academic language items had different semiotic loads and different semiotic structures.

Performance Differences Between ELL and non-ELL Students

Performance score gap and gain scores. We examined the score differences between linguistic groups and across testing occasions for each combination of grade and type of item. We also compared the pattern of score gaps between ELL and non-ELL students and the pattern of score gains attributable to instruction on the CK and AL tests for each linguistic group.

Table 4 shows the mean scores obtained by ELLs and non-ELLs by type of item, grade, and testing occasion. For each combination of linguistic group and grade, the percentage of items correct was higher for the AL than the CK tests and was higher in the posttest than in the pretest. Because, on average, the number of words (a rough indicator of linguistic complexity) was greater for AL items than CK items, these performance differences observed do not appear to be simply due to differences in linguistic complexity. Rather, these differences can be attributed to the emphasis of the instructional program on academic language.

We performed a series of Linguistic Group (ELLs and non-ELLs) × Testing Occasion (pretest and posttest) repeated measures analyses of variance. The former and the latter were treated respectively as a between-subjects and a within-subjects factor. These analyses revealed consistent, statistically significant ($\alpha = .05$) differences due to Linguistic Group (partial $\eta^2 =$ .065, .176, .029, and .136, respectively, for CK-Grade 4, CK-Grade 5, AL-Grade 4, and AL-

TABLE 4
Mean Content Knowledge (CK) and Academic Language (AL) Test Scores and Test Score Gaps and Gains

	CK			*AL*		
	ELL[a]	*Non-ELL*[b]	*Gap*	*ELL*[c]	*Non-ELL*[d]	*Gap*
Grade 4						
Pretest	22.03 (08.05)	25.24 (14.26)	3.21	31.97 (11.83)	38.43 (15.04)	6.46
Posttest	35.18 (17.10)	41.15 (19.54)	5.97	45.49 (18.85)	54.42 (19.84)	8.93
Gain	13.15	20.11		13.52	15.99	
Grade 5						
Pretest	26.81 (12.89)	38.09 (18.52)	11.28	35.84 (14.76)	48.36 (17.20)	8.02
Posttest	39.85 (19.52)	56.25 (24.93)	16.40	44.92 (19.03)	62.42 (21.95)	17.50
Gain	13.04	18.16		9.08	14.06	

Note. Test scores computed as the percentage of items correct. Performance gaps computed as non-ELL scores—ELL scores. Performance gains computed as posttest scores—pretest scores. Standard deviations in parentheses. ELL = English language learner.

[a]Grade 4, $n = 338$. Grade 5, $n = 251$. [b]Grade 4, $n = 241$. Grade 5, $n = 233$. [c]Grade 4, $n = 333$. Grade 5, $n = 254$. [d]Grade 4, $n = 231$. Grade 5, $n = 230$.

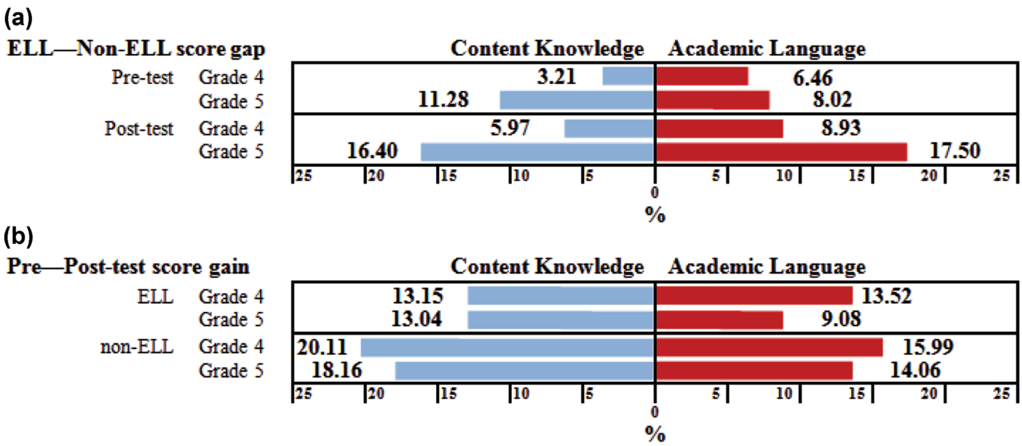

FIGURE 2 Content knowledge—academic language symmetry graph: (a) Test score gap between English language learner (ELL) and non-ELL students by testing occasion and grade. (b) Test score gain from the pretest and the posttest by linguistic group and grade.

Grade 5) and Testing Occasion (partial η^2 = .441, .314, .476, and .468, respectively, for CK-Grade 4, CK-Grade 5, AL-Grade 4, and AL-Grade 5). We did not observe any statistically significant differences due to the interaction of these factors.

Figure 2 shows, in symmetry graphs, the score gaps and gains shown in Table 4. The score gaps between ELL and non-ELL students increased from the pretest to the posttest and tended to be larger for the AL than the CK tests in each grade. (The fact that the score gap was considerably larger for Grade 5 than Grade 4 may be due to the different topics taught in the two grades.) Also, the score gains from the pretest to the posttest tended to be larger for the CK than the AL tests and were smaller for ELLs than non-ELLs.

Thus, non-ELLs outperformed ELLs on both the academic language and content knowledge tests. In terms of the percentage of items correct, both groups performed better on academic language items than content knowledge items. However, in terms of score gains, academic language items were more challenging than content knowledge items for both groups of students and even more challenging for ELL students.

Patterns of correlations between content knowledge and academic language test scores. We compared the Pearson correlations between CK and AL test scores obtained by ELLs and non-ELLs in the pretest and the posttest in both grades. This analysis allowed us to determine whether the CK and AL items posed different sets of challenges to ELLs and non-ELLs.

We observed lower correlations between CK and AL test scores for ELLs than non-ELLs in both grades (see Table 5). These differences were larger in the pretest than in the posttest and considerably larger for Grade 4 (r = .261 for ELLs and r = .635 for non-ELLs) than Grade 5 (r = .510 for ELLs and r = .706 for non-ELLs). For both linguistic groups and for both grades, the correlations were higher in the posttest than in the pretest.

TABLE 5
Pearson Correlations Between Content Knowledge and Academic Language Test Scores by Linguistic Group, Grade, and Testing Occasion

	Grade 4		*Grade 5*	
Linguistic Group	*Pretest*	*Posttest*	*Pretest*	*Posttest*
ELLs	.261 $n = 370$	.732 $n = 341$	.510 $n = 289$	.744 $n = 257$
Non-ELLs	.635 $n = 272$	.755 $n = 241$	.706 $n = 282$	.776 $n = 235$

Note. All correlations are statistically significant at the .01 level (two-tailed). ELL = English language learner.

Thus, with regard to our second research question (How is the performance of ELL and non-ELL students different or similar on content knowledge and academic language items?), the percentage of items correct was higher for AL items than CK items for both linguistic groups. However, in terms of score gains attributable to instruction, AL items were more challenging than CK items for both ELL and non-ELL students. Unlike ELL students, before instruction, non-ELL students appeared to have dealt with the interpretive demands of CK and AL items with similar effectiveness—an indication that ELLs' meaning-making system was not as consolidated as that of non-ELLs.

SUMMARY AND CONCLUSIONS

The ability to measure language-related variables in test items and to relate those measures to measures of student performance is critical to effectively evaluate the impact of language factors on student test scores (Martiniello, 2009; Solano-Flores et al., 2009). Consistent with this notion, in the investigation reported here, we examined the semiotic features of test items created to assess content knowledge and academic language and linked those features to the performance of ELLs and non-ELL students on the items.

Our findings can be summarized as follows. First, academic language items had higher semiotic loads than content knowledge items. Second, the two types of items had different semiotic structures. Third, the score gap between ELL and non-ELL students increased from the pretest to the posttest for both grades and for both content knowledge and academic language; this score gap tended to be larger for academic language than content knowledge. Fourth, although the percentage of items correct was higher for the academic language than the content knowledge tests, in terms of score gains attributable to instruction, academic language items were more challenging than content knowledge items. Score gains tended to be smaller for academic language than content knowledge and were smaller for ELL students than non-ELL students. Fifth, the interpretive demands posed by content knowledge and academic language items appeared to be different for ELL students but similar for non-ELL students. Unlike their non-ELL counterparts, prior to instruction, ELL students appeared not to have developed a consolidated meaning-making system necessary to meet the different sets of interpretive demands of the two types of items with similar effectiveness.

Our findings support concerns about prematurely including ELL students in large-scale testing before they have developed academic language proficiency levels that allow them to engage with all aspects of testing. This lack of preparedness cannot be resolved if instruction does not purportedly support these students to consolidate a meaning-making system. As a result, the magnitude of the correlation between content knowledge and academic language test scores can be interpreted as evidence of convergent or discriminant validity depending on whether the students tested are, respectively, ELLs or non-ELLs. Thus, decisions about construct validity for measures of mathematics content knowledge should not be made without examining disaggregated data for these linguistic groups.

The second implication has to do with assessment development practices. Content knowledge and academic language should be viewed as different but interrelated constructs. Rarely is academic language taken into consideration during the process of assessment development. The analysis of the semiotic structure of items appears to be a promising approach in test development. Examining the frequency of semiotic features of different modalities can help test developers to create test items more systematically, by ensuring that the academic language demands of the items are consistent with the constructs targeted.

Linking item semiotic structure to student performance allows us to understand how disciplinary knowledge and academic language are co-constructed through meaning making. Examining the structure of score gaps between linguistic groups and score gains attributable to instruction helps us to devise effective ways of assessing ELLs, beyond simply comparing score differences between linguistic groups.

ACKNOWLEDGMENTS

This research was supported in part by a grant from the U.S. Department of Education, Institute of Education Sciences (Grant No. R305K050050). The curriculum, Math Pathways & Pitfalls is based on work supported by the Stuart Foundation (Grant No. 96-82) and the National Science Foundation (Grant No. 9911374). We wish to thank Maxie Gluckman for her participation in coding and Maria Ruiz-Primo for her thoughtful comments. The opinions, findings, or conclusions expressed here do not necessarily reflect the views of the funding agencies, or the federal government.

REFERENCES

Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher, 33*, 4–14.

Abedi, J. (2007). English language proficiency assessment and accountability under the NCLB Title III: An overview. In J. Abedi (Ed.), *English language proficiency in the nation: Current status and future practice* (pp. 3–10). Davis: University of California, Davis.

Abedi, J. (2011). Assessing English language learners: Critical issues. In M. R. Basterra, E. Trumbull, & G. Solano-Flores (Eds.), *Cultural validity in assessment: Addressing linguistic and cultural diversity* (pp. 49–71). New York, NY: Routledge.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, *14*, 219–234.

Adger, C. T., Snow, C. E., & Christian, D. (2002). *What teachers need to know about language.* McHenry, IL: Center for Applied Linguistics and Delta Systems, Inc.

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

Bailey, A. L., & Butler, F. A. (2003). *An evidentiary framework for operationalizing academic language for broad application to K–12 education: A design document* (CSE Tech. Rep. No. 611). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Barnett-Clarke, C. & Ramírez, A. (2006). *Mathematics pathways & pitfalls academic language assessment framework*. Unpublished manuscript.

Barnett-Clarke, C., & Ramírez, A., with Coggins, D. (2010). *Math Pathways & Pitfalls lessons and teaching manuals* (Books 1–4). San Francisco, CA: WestEd.

Barwell, R. (2005). Ambiguity in the mathematics classroom. *Language and Education, 19*, 118–126.

Behr, M., Lesh, R., Post, T., & Silver, E. (1983). Rational number concepts. In R. Lesh & M. Landau (Eds.), *Acquisition of mathematics concepts and process* (pp. 91–126). New York, NY: Academic Press.

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

Brenner, M. E. (1998). Development of mathematical communication in problem solving groups by language minority students. *Bilingual Research Journal, 22*, 103–128.

Bruce, B. C., Rubin, A., & Starr, K. S. (1981, August). *Why readability formulas fail* (Reading Education Report No. 28). Urbana: University of Illinois, Center for the Study of Reading.

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). Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing.

Carraher, D. W. (1996). Learning about fractions. In L. P. Steffe, P. Nesher, G. A. Goldin, P. Cobb, & B. Greer (Eds.), *Theories of mathematical learning* (pp. 241–266). Mahwah, NJ: Erlbaum.

Chamot, A. U., & O'Malley, J. M. (1994). *The CALLA handbook: Implementing the cognitive academic language learning approach.* Reading, MA: Addison-Wesley.

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.* Los Angeles: Evaluation, Dissemination and Assessment Center, California State University, Los Angeles.

Durán, R. P. (2011). Ensuring valid educational assessments for ELL students: Scores, score interpretation, and assessment uses. In M. R. Basterra, E. Trumbull, & G. Solano-Flores (Eds.), *Cultural validity in assessment: Addressing linguistic and cultural diversity* (pp. 115–142). New York, NY: Routledge.

Fichtner, D., Peitzman, F., & Sasser, L. (1994). What's fair? Assessing subject matter knowledge of LEP students in sheltered classrooms. In F. Peitzman & G. Gadda (Eds.), *With different eyes: Insights into teaching language minority students across the disciplines* (pp. 143–159). Los Angeles: UCLA Publishing Center for Academic Interinstitutional Programs.

Guerrero, M. D. (2004). Acquiring academic English in one year: An unlikely proposition for English language learners. *Urban Education, 39*, 172–179.

Gutiérrez, K. D., Sengupta-Irving, T., & Dieckmann, J. (2010). Developing a mathematical vision: Mathematics as a discursive and embodied practice. In J. N. Moschkovich (Ed.), *Language and mathematics in education: Multiple perspectives and directions for research* (pp. 29–71). Charlotte, NC: Information Age.

Hakuta, K., Butler, Y. G., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* University of California Linguistic Minority Research Institute Policy Reports (University of California, Multi-Campus Research Unit), eScholarship Repository, University of California. http://repositories.cdlib.org/lmri/pr/hakuta

Halliday, M. A. K. (1978). *Language as social semiotic: The social interpretation of language and meaning*. London, UK: Hodder Arnold.

Halliday, M. A. K. (2005). *Computational and quantitative studies*. New York, NY: Continuum.

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

Hawkins, M. R. (2004). Researching English language and literacy development in schools. *Educational Researcher, 33*(3), 14–25.

Heller, J. I., Curtis, D. A., Rabe-Hesketh, S., & Verboncoeur, C. J. (2007). The effects of "Math Pathways and Pitfalls" on students' mathematics achievement: National Science Foundation final report (ED498258). Retrieved from Education Resources Information Center database: http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED498258

Heller, J. L., Hanson, T. Barnett-Clarke, C., & Darling, K. D. (2010). *The impact of* Math Pathways & Pitfalls *on students' mathematics achievement and mathematical language development: A Study conducted in schools with high concentrations of Latino/a students and English learners.* Report submitted to the U.S. Department of Education, Institute of Education Sciences, Grant R305K050050. Oakland, CA: WestEd.

Kachchaf, R. R. (2011). Exploring problem-solving strategies on multiple-choice science items: Comparing Native Spanish-Speaking English Language Learners and Mainstream Monolinguals (Unpublished doctoral dissertation). University of Colorado, Boulder.

Khisty, L. L. (1995). Making inequality: Issues of language and meaning in mathematics teaching with Hispanic students. In W. G. Secada, E. Fennema, & L. B. Adajian (Eds.), *New directions for equity in mathematics education* (pp. 279–297). Cambridge, UK: Cambridge University Press.

Kopriva, R. J. (2008). Access-based item development. In R. J. Kopriva (Ed.), *Improving testing for English language learners* (pp. 101–140). New York, NY: Routledge.

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

Lampert, M., & Cobb, P. (2003). Communication and language. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 237–249). Reston, VA: The National Council of Teachers of Mathematics.

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, NY: Routledge.

Lemke, J. L. (2003). Mathematics in the middle: Measure, picture, gesture, sign, and word. In M. Anderson & A. Sáenz-Ludlow, S. Zellweger, & V. V. Cifarelli (Eds.), *Educational perspectives on mathematics as semiosis: From thinking to interpreting to knowing* (pp. 215–234). Ottawa, Canada: Legas.

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review*, *78*, 333–368.

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

Mayer, R. E. (2005). *The Cambridge handbook of multimedia learning*. Cambridge, UK: Cambridge University Press.

Mayer, R. E., & Sims, V. K. (1994). For whom is a picture worth a thousand words? Extensions of a dual-coding theory of multimedia learning. *Journal of Educational Psychology, 86*, 389–401.

Moschkovich, J. N. (2007). Using two languages while learning mathematics. *Educational Studies in Mathematics*, *64*, 121–144.

Moss, J., & Case, R. (1999). Developing children's understanding of the rational numbers: A new model and an experimental curriculum. *Journal for Research in Mathematics Education, 30*, 122–147.

Nisbett, R. E., & Miyamoto, Y. (2005). The influence of culture: Holistic versus analytic perception. *Trends in Cognitive Science, 19*, 467–473.

No Child Left Behind Act of 2001 (NCLB), Pub. L. No. 107–110, § 115 Stat.1425 (2002).

O'Halloran, K. L. (2011). Multimodal discourse analysis. In K. Hyland & B. Paltridge (Eds.), *Companion to discourse*. London, UK: Continuum.

O'Malley, J. M., & Valdez-Pierce, L. (1996). *Authentic assessment for English language learners: Practical approaches for teachers.* Reading, MA: Addison Wesley.

Parker, M., & Leinhardt, G. (1995). Percent: A privileged proportion. *Review of Educational Research, 65*, 421–481.

Pimm, C. (1987). *Speaking mathematically: Communication in mathematics classrooms*. London, UK: Routledge & Kegan Paul.

Pirie, S. E. B. (1998). Crossing the gulf between thought and symbol: Language as (slippery) stepping-stones. In H. Steinbring, M. G. Bartolini-Bussi, & A. Sierpinska (Eds.), *Language and communication in the mathematics classroom* (pp. 7–29). Reston, VA: National Council of Teachers of Mathematics.

Rothery, A., & Shuard, H. (1980). *Children reading mathematics.* Worcester, UK: College of Higher Education.

Scarcella, R. C. (2003). *Academic English: A conceptual framework* (Rep. No. 2003–1). Santa Barbara: University of Califronia Linguistic Minority Research Institute.

Schleppegrell, M. J. (2010). Language in mathematics teaching and learning: A research Review. In J. N. Moschkovich (Ed.), *Language and mathematics in education: Multiple perspectives and directions for research* (pp. 73–112). Charlotte, NC: Information Age.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, G. (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.

Sireci, S. G., Han, K. T., & Wells, C. S. (2008). Methods for evaluating the validity of test scores for English language learners. *Educational Assessment, 13*, 108–131.

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*, 2354–2379.

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*, 189–199.

Solano-Flores, G. (2010). Function and form in research on language and mathematics education. In J. N. Moschkovich (Ed.), *Language and mathematics in education: Multiple perspectives and directions for research* (pp. 113–149). Charlotte, NC: Information Age.

Solano-Flores, G. (2011). Language issues in mathematics and the assessment of English language learners. In K. Tellez, J. Moschkovich, & M. Civil (Eds.), *Latinos/as and mathematics education: Research on learning and teaching in classrooms and communities* (pp. 283–314). Charlotte, NC: Information Age.

Solano-Flores, G., Backhoff, E., & Contreras-Niño, L. A. (2009). Theory of test translation error. *International Journal of Testing, 9*, 78–91.

Solano-Flores, G., & 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). New York: Taylor & Francis.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching, 38*, 553–573.

Solano-Flores, G., & Wang, C. (2011, April). *Conceptual framework for analyzing and designing illustrations in science assessment: Development and use in the testing of linguistically and culturally diverse populations*. Paper presented the Annual Conference of the National Council on Measurement in Education, New Orleans, LA.

Stevens, R. A., Butler, F. A., & Castellon-Wellington, M. (2000). *Academic language and content assessment: Measuring the progress of English language learners* (*ELLs*) (CSE Report 552). Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing.

Trumbull, E., & Solano-Flores, G. (2011). The role of language in assessment. In M. Basterra, E. Trumbull, & G. Solano-Flores (Eds.), *Cultural validity in assessment: Addressing linguistic and cultural diversity* (pp. 22–45). New York, NY: Routledge.

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

Wang, C., & Solano-Flores, G. (2011, April). *The use of illustrations in large-scale science assessment: A comparative study*. Paper presented at the annual meeting of the American Educational Research Association.

Wearne, D., & Hiebert, J. (1989). Cognitive changes during conceptually based instruction on decimal fractions. *Journal of Educational Psychology*, *81*, 507–513.

Webb, N. M. (1991). Task-related verbal interaction and mathematics learning in small groups. *Journal for Research in Mathematics Education, 22*, 366–389.

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

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