[1] Differential Performance of English Learners on Science Assessments: The Role of Cognitive Complexity

[2]George E. DeBoer, Cari F. Herrmann-Abell, and Sarah Glassman
AAAS Project 2061

Paper presented at the 2016 NARST Annual International Conference

Baltimore, MD
April 17, 2016

Introduction

This paper describes the development and validation of measures for scoring the cognitive complexity of middle school science test items. These measures were then used in a study of factors related to the differential performance of English Learners (ELs) on science assessments compared to native English speakers. The measures were used to analyze the cognitive complexity of 476 middle school assessment items drawn from almost 1000 items previously developed by the American Association for the Advancement of Science (AAAS). In a separate part of the study, we focus separately on the development, validation, and reliability of measures of **linguistic** complexity and the results of applying those measures to the same set of test items (Trumbull, Nelson-Barber, & Huang, 2016). The full study looks at linguistic and cognitive complexity together as predictors of student performance. Preliminary results of that combined analysis are presented here as well.

The Problem

Students whose primary language is not English do not perform as well as English-speaking students on tests that assess their understanding of the school curriculum (Martiniello, 2008). Although some of this difference may reflect real differences in students' knowledge, it is also possible that the tests are not fairly evaluating what students know. The magnitude of the differences in performance between EL and non-EL that we found in this study can be seen in Table 4.

Understanding the factors that affect EL student performance is becoming even more important with the advent of *Common Core State Standards* and *Next Generation Science Standards* (NGSS), which place greater emphasis on the use of language and reasoning skills so that students can engage effectively in science discourse and sense making (Lee et al., 2013). The work is also consistent with this meeting's theme of educational equity and justice for all. Given that ELs are such a rapidly growing segment of the U.S. student population, their underperformance is a particularly significant issue to address.

[1] This research was supported by National Science Foundation Grant DRL-1348622 awarded to AAAS (George E. DeBoer, principal investigator). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

[2] The authors acknowledge the contributions of Sharon Nelson-Barber and Chun-Wei Huang at WestEd, Jo Ellen Roseman and Bernard Koch at AAAS, Jill Burstein at ETS, and Elise Trumbull, consultant.

Importance

Analyzing the effect of cognitive and linguistic complexity together should help us find out the relative importance of each of these factors on student performance, and if strategies to reduce the language load of test items (Haladyna & Downing, 1989; Haladyna, Downing, & Rodriguez, 2002; Kehoe, 1995) oversimplifies test items so much that reducing linguistic complexity is achieved at the expense of assessing rigorous content. This work should also provide educators with knowledge they need to make targeted help available to underperforming students in particular areas of the science curriculum and to assess them fairly.

Methods

Assessment Items Used

About half of the nearly 1000 items (available at assessment.aaas.org) that had previously been developed by AAAS Project 2061 with a grant from the National Science Foundation were analyzed in this study. The 1000 items assess student understanding of 17 topics in middle school science and were field tested on both middle and high school students from across the country. The 476 items reported on here are from nine of those topics. 1000-2000 students responded to each item during field testing. Items were aligned to precise target learning goals, and item construction followed rigorous item development procedures (DeBoer et al., 2008).

Differential Item Functioning

The Mantel-Haenszel differential item functioning (DIF) procedure (Holland & Thayer, 1988) was used to identify items that showed a statistically significant DIF score when comparing students whose primary language is not English to students whose primary language is English. This produced a subset of items that could be carefully scrutinized for cognitive and linguistic features that might explain some of the differential performance of EL and non-EL students.

Cognitive Complexity Rubric Development

The development of rubrics for scoring cognitive complexity was informed by a number of previously developed frameworks. We began by using the hierarchy of knowledge categories – declarative, procedural, schematic, and strategic (Li, 2001; Shavelson et al., 2003). We found, however, that multiple-choice items (such as ours) rarely assess procedural knowledge (knowing how to use a scientific procedure or technique) or strategic knowledge (choosing which of several approaches to use to solve a problem). An approach more consistent with the multiple-choice format was needed. We also considered two-tiered systems that looked at both the type of knowledge and type of thinking required (Hess et al., 2009; Oosterhof et al., 2008). Kopriva and Winter (2012), for example, used the declarative, procedural, schematic, and strategic categories along with cognitive demand categories of recall, application, reasoning, and extended thinking. Our scoring system takes into account both of these dimensions—a knowledge dimension and a mental processing dimension.

The Knowledge Dimension. Our knowledge dimension includes declarative knowledge (facts, rules, and definitions) and schematic knowledge (interconnected ideas). As we began to develop and test the rubrics, we found that we agreed on how to categorize items that required knowledge of one simple fact (declarative knowledge), but we often disagreed when more than one fact was involved. Should multiple facts be thought of as complex declarative knowledge or as simple schematic knowledge? We also found that counting the number and interconnectivity of discrete pieces of knowledge was difficult to do reliably. As Knaus and colleagues (2011) pointed out, “one expert may parse a required task into two ‘easy’ steps, where another would identify the same task as one ‘medium’ step” (Knaus et al., 2011, p. 555). Their solution was to create a rubric that allowed raters multiple ways to reach the same numeric score. We ultimately created a five-level rubric broadly divided into declarative and schematic categories. As with their rubric, a middle-level score can be assigned to an item that one reviewer thinks is assessing knowledge of multiple discrete facts (complex declarative knowledge) and another thinks is assessing a simple set of interconnected ideas (simple schematic knowledge).

Scoring the Knowledge Dimension

1. Declarative Knowledge**:** Included here are knowledge of facts, rules, procedures, principles, definitions, and standard conventions such as atomic models, food web diagrams, force arrows, etc. Declarative knowledge is scored between 1 and 3 depending on the number and complexity of discrete ideas tested.
2. Schematic Knowledge: This category includes organized bodies of knowledge such as mental models, schema, and theories. Schematic knowledge involves interrelationships among ideas and can be used to understand and explain phenomena, make predictions, draw conclusions, or solve problems. Just as with declarative knowledge, schematic knowledge can be simple and it can be complex. Schematic knowledge is scored between 3 and 5 depending on the number of ideas tested and how interconnected they are.

In deciding between declarative and schematic knowledge, raters are asked to consider:

- What students are being asked to do with the knowledge (simple recognition often suggests the knowledge is declarative)
- Misconceptions in the answer choices (being able to reject misconceptions may require a mental schema)
- How the idea is typically taught in school (as a fact to memorize or as mental schema)

Example of Declarative Knowledge

Item CE72-4 simply requires knowledge that both plant and animal cells perform basic functions such as obtaining energy from food.

CE72-4

Which of the following kinds of cells perform basic functions such as obtaining energy from food?

A. Plant cells, but not animal cells
B. Animal cells, but not plant cells
C. Both plant cells and animal cells
A. Neither animal cells nor plant cells

Example of Schematic Knowledge

Item CL115-1 expects students to have knowledge of the relationship between the sun and earth and how that relationship affects the angle at which the sunlight strikes the earth during the year.

CL115-2

What is TRUE about the place where sunlight strikes the earth's surface at a 90° angle during the months of May and October?

A. *The place is a little farther north each day in May and a little farther south each day in October.*
B. The place is a little farther south each day in May and a little farther north each day in October.
C. The place is a little farther north each day in May, but it stays in the same place each day in October.
D. The place does not change at all during May or October.

The Mental Processing Dimension. Mental processing focuses on the complexity of thinking needed to answer a test question, conceptually distinct from but overlapping with the complexity of the knowledge itself. The two broad mental processing categories in our rubric—recognizing knowledge and applying knowledge—are influenced by Webb's (2002) depth of knowledge (DOK) framework. Webb differentiates between items that require students to recall knowledge, items that require students to apply knowledge, and items that require reasoning.

In our rubric, Webb's "recall" becomes "recognition," and Webb's "applying knowledge" and "reasoning with knowledge" are placed under the single category "application." The two broad categories of **recognition** and **application** are then subdivided based on additional features that influence mental processing, such as the presence of symbolic representations, the need to form a mental image of the problem, and the abstractness of the content (Ferrara, et al., 2010; Shavelson et al., 2003; Wang, et al., 2014). Recognition is divided into three subcategories, and application is divided into two subcategories. Each of these five subcategories has two scoring levels. The higher score for each subcategory overlaps with the lower score of the subcategory above it. Therefore, just as with the knowledge scores, there are different ways for raters to reach the same numeric score for mental processing.

Scoring the Mental Processing Dimension

At the lower levels, students are simply asked to recognize true statements of fact or principle, or examples that fit certain definitions or rules. As we move up the scale, students are asked to use their knowledge in more sophisticated ways and in less and less familiar contexts. The five subcategories of the mental processing dimension and their corresponding score ranges are:

1. Recognize a true statement of fact, principle, definition, or rule. (Score 1-2)
2. Recognize a correct instance or example of a fact, principle, definition, or rule. (Score 2-3)
3. Recognize a correct instance or example of a fact, principle, definition, or rule when additional mental processing is required, e.g. such as having to interpret a diagram. (Score 3-4)

4. Apply knowledge in real world contexts or to solve abstract problems. (Score 4-5)
5. Apply and reason with knowledge in less familiar contexts or in a more sophisticated way. (Score 5-6)

Examples of Items Demonstrating Different Levels of Mental Processing

1. Recognize a true statement of fact, principle, definition, or rule. (Score 1-2)

   In Item CE72-4, students need to recognize that both plant cells and animal cells perform basic functions.

CE72-4

Which of the following kinds of cells perform basic functions such as obtaining energy from food?

D. Plant cells, but not animal cells
E. Animal cells, but not plant cells
F. Both plant cells and animal cells
G. Neither animal cells nor plant cells

2. Recognize a correct instance or example of a fact, principle, definition, or rule. (Score 2-3)

   In Item SC46-5, students need to recognize that solubility (how much of the substance dissolves in water) is an example of a characteristic property.

SC46-5

Which of the following is a characteristic property of a pure substance?

A. How much of the substance dissolves in water
B. How much space the substance takes up
C. What the temperature of the substance is
D. What the width of the substance is

3. Recognize a correct instance or example of a fact, principle, definition, or rule when additional mental processing is required. (Score 3-4)

In Item IE38-8, students have to interpret a real-world scenario. They are expected to recognize that competition between wolves and bears, wolves and wolves, and bears and bears are consistent with the scientific principle of intra- and inter-species competition.

IE38-8

Wolves and bears are living in the same national park in Canada. They both eat elk that are in the park. Which of the following statements is TRUE?

A. The wolves compete with other wolves, the bears compete with other bears, but the wolves and bears do not compete with each other for the elk.
B. The wolves do not compete with other wolves, the bears do not compete with other bears, and the wolves and bears do not compete with each other for the elk.
C. The wolves compete with other wolves, the bears compete with other bears, and the wolves and bears compete with each other for the elk.
D. The wolves do not compete with other wolves, the bears do not compete with other bears, but the wolves and bears compete with each other for the elk.

Item BF130-1 also expects students to recognize a true fact, but in this case they also have to form a mental image of something that is not easily observed. They need to imagine sugars traveling through the digestive track to respond correctly.

BF130-1

Which of the following statements is TRUE about simple sugars getting from the digestive tract to cells of the brain and the cells of the skin?

A. The circulatory system carries simple sugars to cells of both the brain and the skin.
B. The circulatory system carries simple sugars to cells of the brain but not to cells of the skin.
C. Simple sugars get to cells of both the brain and the skin, but these molecules are not carried by the circulatory system.
D. Simple sugars do not get to cells of the brain or the skin.

4. Apply knowledge in real world contexts or to solve abstract problems. (Score 4-5)

At this level, the mental processing moves from recognition to application. The application of knowledge can involve solving problems, explaining phenomena, making predictions, explaining relationships between variables, drawing conclusions, evaluating claims and reasons, or analyzing situations to determine what else must be true given the conditions that are described. Application can also involve interpreting graphs, tables, diagrams, symbols, and standard conventions to help solve a problem. The application can be to real world situations or to abstract problems. At this level, the real world contexts should be relatively familiar and uncomplicated.

Item CL52-2 is an example of an item where students need to use knowledge to explain a real world phenomenon. This item asks students to use their knowledge of weather and climate to explain the phenomenon of a town being cold in the winter and hot in the summer. They use the knowledge that the air temperature at a place depends on the intensity of sunlight shining on the town and the number of hours the sun shines.

CL52-2

The air in a town is very cold in the winter and very hot in the summer. Which of the following statements explains this difference in temperature?

A. The air is colder in the winter because the sunlight shining on the town is less intense and the sun shines for fewer hours in the winter than in the summer.
B. The air is colder in the winter because the sunlight shining on the town is less intense in the winter. The amount of time that the sun shines on the town is the same in the summer and winter.
C. The air is colder in the winter because the sun shines on the town for less time in the winter. The intensity of the sunlight shining on the town is the same in the summer and winter.
D. Both the intensity of the sunlight shining on the town and the amount of time the sun shines on the town are same in the summer and winter. The differences in temperature are caused by other factors.

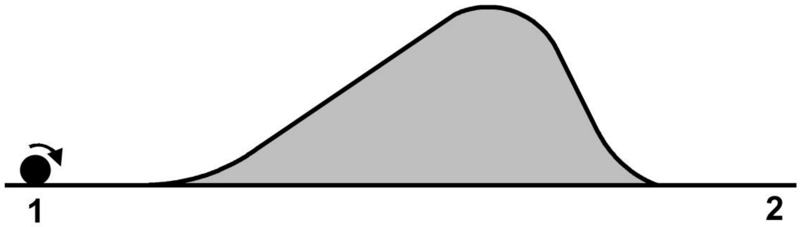
5. Reason with knowledge in less familiar contexts or in a more sophisticated way. (Score 5-6)

   At this level, students are expected to use their knowledge to figure something out. Typically, these contexts will be more complicated or less familiar than for the previous category, so it is less likely that the knowledge that is needed or how to use it will be obvious or straightforward to the students. The reasoning may also involve multiple mental processes so that students might have to both make a prediction and justify that prediction, or they may have to decide which knowledge applies to the problem they are given.

For Item NG90-2, students have to analyze the diagram to see that Points 1 and 2 are at the same height, make a prediction about the speed of the ball, and use science ideas about energy to justify their prediction.

NG90-2

Imagine a ball on a track where no energy is transferred between the ball and the track or between the ball and the air around it. It is going fast enough at Position 1 so that it will go over a hill on the track and past Position 2. Position 1 and Position 2 are at the same height.



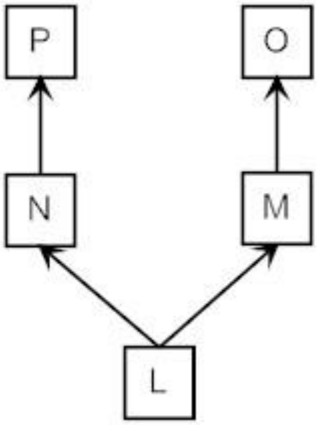
Will it be going faster, slower, or at the same speed at Position 2 compared to Position 1 and why? (Remember that no energy is transferred between the ball and the track or between the ball and the air around it.)

A. Faster, because new energy in the form of motion energy (kinetic energy) was made when the ball went down the steep side of the hill
B. Slower, because motion energy (kinetic energy) was used up when the ball went up the long side of the hill
C. The same speed, because the amount of motion energy (kinetic energy) that the ball has remained the same the entire time it was moving along the track
D. The same speed, because the total amount of energy in the system (ball and track) did not change as the ball moved along the track

Item IE28-5 is another example of a level 5 item. It asks students to make a prediction about the population of *N*s after a change in the ecosystem. They use the food web diagram in the stem to make a prediction, and then they explain their prediction using science ideas about interdependence in ecosystems. The abstractness of using letters to represent organisms adds additional mental processing.

IE28-5

The diagram below shows the feeding relationships between populations of organisms *L, M, N, O,* and *P*. The arrows point from the organisms being eaten to the organisms that eat them.



Using only the relationships between the organisms shown in the diagram, what will happen to the number of *N*s if most of the *O*s are killed and why?

A. The number of *N*s will decrease because the number of individuals in all of the populations of organisms in this diagram will decrease when the number of *O*s decreases.
B. The number of Ns will decrease because there will be more Ms to eat the Ls, so fewer Ls will be available for the Ns to eat.
C. The number of *N*s will stay the same because there will be no effect on the number of individuals in the populations of organisms below the *O*s in the diagram.
D. The number of *N*s will stay the same because *O*s and *N*s are not connected by an arrow in the diagram.

In Item EN30-2, another level 5 item, students must make a mental connection between the disease resistance context in the item and the science idea of natural selection in order to explain what happens to the population of bacteria. Knowing which science idea to use is not obvious in this item, and this adds to the mental processing needed to answer correctly.

> EN30-2
>
> Which of the following correctly describes what happens when a population of bacteria becomes resistant to an antibiotic? Note: a bacterium is one individual in a group of bacteria.
>
> A. During treatment with an antibiotic, each individual bacterium tries to become resistant to the antibiotic. Only some are able to willingly become resistant, and these individuals survive to pass this trait to their offspring.
> B. During treatment with an antibiotic, all of the bacteria gradually become more resistant to the antibiotic the more they are exposed to it. They all survive and pass this trait to their offspring.
> C. During treatment with an antibiotic, a population of bacteria usually dies. Sometimes by chance, all members of the population become resistant at once, survive, and pass their resistance to their offspring.
> *D. During treatment with an antibiotic, only those individual bacteria that already have a trait that helps them survive the effects of the antibiotic will live. Their offspring in the next generation will also have this trait.*

Score reliability. The development of the scoring rubrics involved multiple rounds of trial testing. After rubric development was completed, the same three raters scored all of the items, topic by topic. Each set of items (the items for a single topic) was coded over a period of several weeks independently by the three raters, with the entire scoring process spread out over about five months. Each rater coded two-thirds of the items for a topic so that each item was scored by two of the three raters. Score reliability was estimated by counting the number of scores that matched exactly and the number of scores that matched within one. If the ratings differed by no more than one point, the average of the two was assigned to the item. If the raters disagreed by more than one point, the item was then coded by the third reviewer. If the third rater's rating agreed with either of the first two, that score was reported. Otherwise, if there were three different ratings, a reconciliation meeting was held to decide on the final rating. Krippendorff's alpha (2004) was calculated on as an estimate of coding reliability.

Measures of Linguistic Complexity

With the help of an automated sentence parser, an index of syntactic complexity was calculated for each sentence. The number of leaves (words), branches, nodes (where two branches come together or where a branch ends), and levels (where a new node appears off an existing branch) were used as indicators of the complexity of a sentence (cf., Klammer, Schulz, & Della Volpe, 2000). By dividing the number of levels (where a new node appears off an existing branch) by the number of nodes (where two branches come together or where a branch ends) we derived a ratio of syntactic complexity ranging from 0 to 1 for each sentence. A smaller ratio means greater syntactic complexity (Solano-Flores, Trumbull, & Kwon, 2003). Item-level complexity scores were calculated from these sentence scores by counting the number of sentences in the item with a syntactic complexity index equal to or less than .200. Item length was measured by the total number of words (leaves) in the item.

We also used the ETS Language Muse natural language processing system (Burstein et al., 2012) to analyze our 476 items for the frequency of occurrence of 35 linguistic features. In the ETS system, linguistic features are organized into sentence, vocabulary, and discourse structure categories. Examples of features include: long prepositional phrases, complex noun phrases, passives, complex verbs, etc.

Results and Discussion

Frequency distributions for student test scores, two measures of cognitive complexity, number of complex sentences, and item length are shown in Figures 1-5.

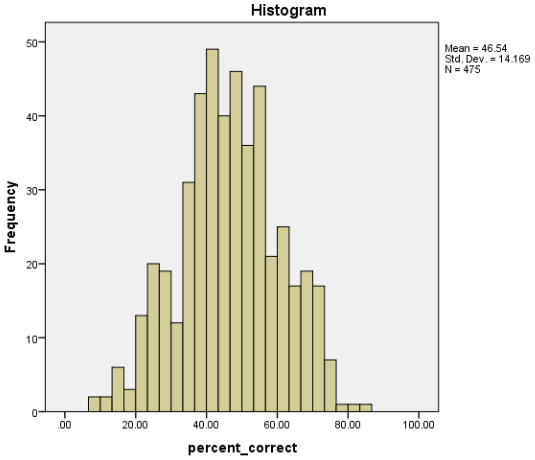


Figure 1. *Distribution of Percent Correct Scores*

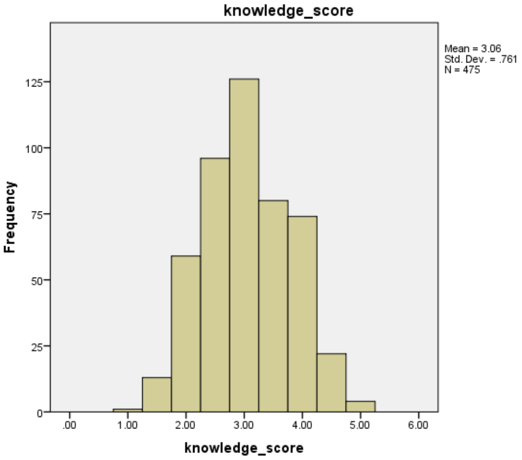


Figure 2. *Distribution of Knowledge Complexity Scores*

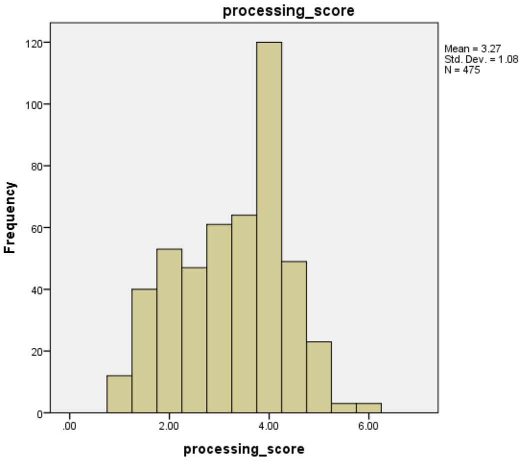


Figure 3. *Distribution of Mental Processing Complexity Scores*

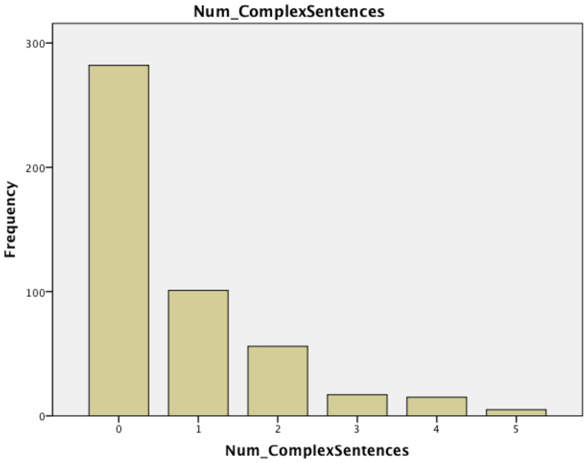


Figure 4. *Distribution of the Number of Complex Sentences*

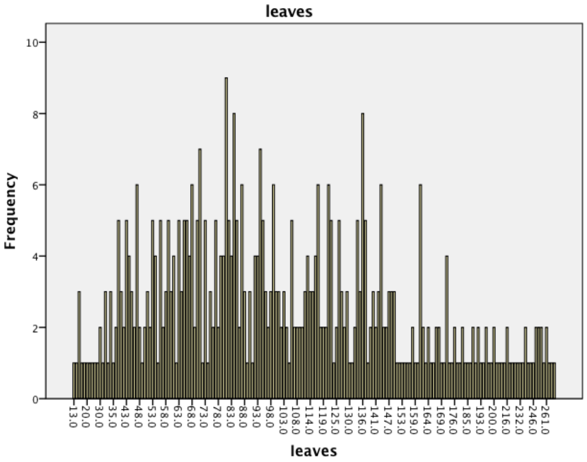


Figure 5. *Distribution of the Number of Leaves (Words)*

Results of Cognitive Complexity Analysis

Reliability. For the cognitive complexity measures, estimates of inter-rater reliability showed that over 90% of ratings matched within one point on each complexity scale. When Krippendorff’s alpha (2004) was calculated as an estimate of coding reliability among the three raters, results following reconciliation showed a reliability of 0.63 for complexity of knowledge and .77 for complexity of mental processing (Table 1).

Table 1
Krippendorff’s Alpha Reliability Estimate for Complexity of Knowledge and Mental Processing Scales

	Krippendorff’s Alpha	
	Knowledge	Mental Processing
Initial rating (2 raters per item)	0.55	0.69
After 3rd rater if initial 2 raters differed by more than 1 point	0.57	0.71
After reconciliation when 3rd rater was not an exact match to either of the first two ratings	0.63	0.77

Correlations between Knowledge Complexity, Mental Processing Complexity, and Item Difficulty. Knowledge and mental processing measures were moderately inter-correlated. For all topics combined, the correlation between the two was .567 and significant at the .01 level (Table 2). Control of Variables had the highest inter-correlation (.928). Overall, the correlations between measures of cognitive complexity and item difficulty were statistically significant but small (-.268 and -.222). For five topics (Control of Variables, Interdependence in Living Systems, Matter and Energy, Energy Transformation, Transformation, and Conservation, and Weather and Climate) both dimensions of cognitive complexity were significantly correlated with difficulty; For two topics (Evolution and Natural selection and Atoms and Molecules), only knowledge complexity was significantly correlated with item difficulty; for two topics (Plate Tectonics and Forms of Energy), the correlation was only with mental processing; and for one topic (Forms of Energy), there was no correlation between either measure of cognitive complexity and item difficulty (Table 2).

Table 2

Correlation Coefficients between Cognitive Complexity Scores and Student Performance

Topic[++]	N Items	Mean Knowledge (K)	Mean Mental Processing (MP)	% Correct	r (K x MP)	r (K x % Correct)	r (MP x % Correct)
AM	44	2.84	3.11	49.72	.564**	-.256	-.521**
CV	21	4.05	4.14	51.96	.928**	-.767**	-.795**
EN	39	2.97	2.86	45.65	.785**	-.131	-.318*
IE	45	3.21	3.60	60.86	.727**	-.693**	-.369*
ME	44	2.89	2.52	40.97	.677**	-.430**	-.366*
EG	91	2.53	3.27	46.07	.129	-.190	.044
NG	95	3.50	3.92	41.55	.639**	-.366**	-.409**
PT	35	2.84	2.77	43.63	.563**	-.213	-.140
WC	62	3.15	2.91	46.53	.492**	-.499**	-.403**
Total	476	3.06	3.27	46.64	.567**	-.268**	-.222**

$**p < .01$, $*p<.05$

[++]AM (Atoms and Molecules), CV (Control of Variables), EN (Evolution and Natural Selection), IE (Interdependence in Ecosystems), ME (Matter and Energy in Living Systems), EG (Forms of Energy), NG (Energy Transformations, Transfer, and Conservation), PT (Plate Tectonics), WC (Weather and Climate)

Multiple Regression Analysis. When the two dimensions of cognitive complexity were used in a multiple regression analysis to predict overall student performance, R^2 was small (.079) but significant for items from all topics combined and significant for six of nine individual topics. It was not significant for the Evolution and Natural selection topic, the Forms of Energy topic, or the Plate Tectonics topic (Table 3).

Table 3

Multiple Regression Analysis: Predicting Student Performance from Two Measures of Cognitive Complexity

Topic	N Items	Mean Knowledge (K)	Mean Mental Processing (MP)	% Correct	B_K	B_{MP}	R^2
AM	44	2.84	3.11	49.72	1.10	-8.03**	.273**
CV	21	4.05	4.14	51.96	-8.20	-17.38**	.638**
EN	39	2.97	2.86	45.65	5.96	-8.44*	.138
IE	44	3.21	3.60	60.86	-20.54**	3.82	.518**
ME	44	2.89	2.52	40.97	-6.83**	-2.12	.195**
EG	91	2.53	3.27	46.07	-3.75	0.74	.041
NG	95	3.50	3.92	41.55	-3.67	-4.17**	.186**
PT	35	2.84	2.77	43.63	-3.18	-0.42	.046
WC	62	3.15	2.91	46.53	-7.42**	-2.42	.282**
Total	475	3.06	3.27	46.64	-3.88**	-1.36	.079**

Results of DIF Analysis. Although there were significant performance differences favoring native English speakers, which ranged from 4.49 percentage points for the Evolution and Natural Selection topic to 13.43 percentage points for the Control of Variables topic (Table 4), DIF analysis identified only 18 items with a moderate Delta score (Zieky, 1993; Zwick, et al., 1999) and three items with a moderate to high Delta score (Table 5), suggesting that test bias toward either EL or non-EL students was a relatively minor issue. (Ten of the DIF items favored ELs and 11 favored non-ELs.)

Table 4
Knowledge, Mental Processing, and Performance of Non-EL and EL Students by Topic

Topic	Overall % Correct	Non-EL% Correct	EL % Correct	Non-EL minus EL
AM	49.72	50.99	38.99	12.00
CV	51.96	53.56	40.20	13.36
EN	45.65	46.09	41.60	4.49
IE	60.86	61.66	51.66	10.00
ME	40.97	41.72	34.83	6.89
EG	46.07	46.99	39.26	7.74
NG	41.55	42.14	35.66	6.47
PT	43.63	44.03	38.48	5.55
WC	46.53	47.08	41.78	5.30
Total	46.64	39.78	47.30	7.52

Table 5
*Summary of DIF Findings by Topic: Number of Items Having Moderate and Moderate to High Delta Scores**

Topic	N Items	Items with moderate Delta	Items with moderate to high Delta
AM	44	2	0
CV	21	0	0
EN	39	1	0
IE	44	2	1
ME	44	3	0
EG	91	0	1
NG	95	3	0
PT	35	2	0
WC	62	5	1
Total	475	18	3

A high Delta score suggests the item is biased either toward or against EL students (direction of bias is not specified)

A one way analysis of variance showed no significant differences among the mean knowledge complexity scores and mean mental processing complexity scores for the DIF items that favored ELs, the DIF items that favored non-ELs, or the no-DIF items. When comparing overall percent correct for items in these categories, the ANOVA showed that items that favored ELs were more difficult than items that favored native English speakers and non-DIF items. Mean difference scores (Non-ELs minus ELs) showed EL students doing better than expected on the DIF items that favored them and worse than expected on the DIF items that favored non-EL students, as anticipated (see Table 6.)

Table 6

One way ANOVA of Cognitive Complexity Scores and % Correct for Year 1 DIF Items Favoring non-ELs, DIF Items Favoring ELs, and Non-DIF Items

	DIF items favoring non-EL	DIF items favoring EL	Non-DIF items	F	p
Mean Knowledge	2.96	3.10	3.05	.110	n.s.
Mean Mental Processing	3.09	3.10	3.28	.277	n.s.
Mean % correct	57.16%	40.69%	46.43%	4.009	<.05
Non-ELs minus ELs	17.41%	-3.16%	7.53%	40.170	<.001
Number of items	11	10	455		

Combining Cognitive and Linguistic Complexity Measures in Regression Analyses

Item Syntactic Complexity and Item Length. The item-level measure of syntactic complexity that we calculated on 472 items by counting the # complex sentences (sentences with a syntactic complexity of .200 or less) was significantly related to percent correct at the .001 level for both EL and non-EL students (R^2 = .029 for ELs, .053 for non-ELs). The number of words in an item was also significantly related to percent correct at the .01 level for EL students and the .001 level for non-EL students (R^2 = .018 for ELs, .032 for non-ELs). These results can be seen in Table 7 along with the regression coefficients for the two cognitive complexity measures. The direction of the sign in each case tells us that the more complex an item is, either linguistically or cognitively, the more difficult it is for both EL and non-EL students.

Table 7

Using Simple Regression to Predict Item Difficulty from Cognitive and Linguistic Complexity Measures for EL and non-EL Students

EL Students	R^2	β	p
Knowledge	0.053	-0.230	<.001
Mental Processing	0.048	-0.218	<.001
# Complex sentences	0.029	-0.170	<.001
# Words (leaves)	0.018	-0.133	<.01
Non-EL Students	R^2	β	p
Knowledge	0.074	-0.271	<.001
Mental Processing	0.049	-0.222	<.001
# Complex sentences	0.053	-0.231	<.001
# Words (leaves)	0.032	-0.179	<.001

Results of Multiple Regression Analyses. When our four cognitive and linguistic complexity variables (knowledge complexity, mental processing complexity, item-level syntactic complexity, and item length) were used to predict item difficulty in a multiple regression analysis, we found that these four complexity variables explain 8.3% of the variance in item difficulty for EL students and 10.9% of the variance for non-EL students. All variables in the multiple regression equation are significant at the .05 level (Table 8).

Table 8
Using Multiple Regression to Predict Item Difficulty from Cognitive and Linguistic Complexity Measures for EL and non-EL Students

EL Students (R^2 = .083)	β	p
Knowledge	-0.162	<.01
Mental Processing	-0.179	<.01
# Complex sentences	-0.173	<.01
# Words (leaves)	0.165	<.05
Non-EL Students (R^2 = .109)	**β**	**p**
Knowledge	-0.203	<.001
Mental Processing	-0.133	<.05
# Complex sentences	-0.218	<.001
# Words (leaves)	0.141	<.05

Item Length. It should be noted that when the number of words (leaves) is used in the multiple regression equation, the sign of the regression coefficient is now positive. That is, longer sentences actually have a positive effect on student performance after cognitive complexity and linguistic complexity are controlled. This can be explained by the fact that the correlation among these variables is moderately high: For mental processing and number of words, r = .599; for number of complex sentences and number of words, r = .597, and for knowledge and number of words, r = .541. It is not surprising that the more words there are in a test item, the more likely it is that an item will be cognitively and linguistically complex. But after controlling for those factors, it appears that additional words may actually help students understand what the item is asking.

ETS Language Muse Analysis. At this time, there is nothing to report from the ETS Language Muse analysis of 35 linguistic features. Preliminary analysis shows that a relatively small number of features are statistically related to item difficulty for both EL and non-EL students. The presence of some of the features seems to be related to higher student performance and some to lower student performance. Additional work is needed to sort out these differences.

Summary. The results of this study demonstrate that it is possible to design measures of cognitive (and linguistic) complexity that help explain item difficulty for both EL and non-EL students. The variance in item difficulty that is explained is small, but it is statistically significant in each case. It is understandable that the variance explained would be small, given that the strongest predictor of student performance is likely to be the knowledge they have, not structural features of test items. It also appears that these predictors operate similarly for EL and native English speakers, suggesting that cognitive and linguistic complexity affect the performance of both groups. This, too, is not surprising given that there is a wide range of language proficiency in native English speakers as well as in students whose primary language is not English. Finally, there are differences in how these item features operate, topic by topic, and this will be part of our analysis going forward.

References

Burstein, J., Shore, J., Sabatini, J., Moulder, B., & Holtzman, S. (2012, April). *The Language Muse^{SM} System: Linguistically-focused instructional authoring*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Vancouver, BC.

DeBoer, G. E., Herrmann-Abell, C. F., Gogos, A., Michiels, A., Regan, T., & Wilson, P. (2008). Assessment linked to science learning goals: Probing student thinking through assessment. In J. Coffey, R. Douglas, & C. Stearns (Eds.), *Assessing student learning: Perspectives from research and practice* (pp. 231-252). Arlington, VA: NSTA Press.

Ferrara, S., Huff, K., & Lopez, E. (2010). Targeting cognition in item design to enhance valid interpretations of test performances: A case study and some speculations. In *Cognition and Valid Inferences about Student Achievement: Aligning Items with Cognitive and Proficiency Targets*. Denver, CO.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334.

Haladyna, T. M., & Downing, S. M. (1989). Validity of a taxonomy of multiple-choice-item-writing rules. *Applied Measurement in Educa*tion, *2*(1), 51-78.

Hess, K. K., Jones, B. S., Carlock, D., & Walkup, J. R. (2009). Cognitive rigor: Blending the strengths of Bloom's taxonomy and Webb's depth of knowledge to enhance classroom-level processes. Education Resources Information Center.

Holland, P. W. and Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Ed.), *Test validity*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kehoe, J. (1995). Writing multiple-choice test items. *Practical Assessment, Research & Evaluation, 4*(9), Available: http://pareonline.net/getvn.asp?v=4&n=9. Retrieved 3/8/08.

Klammer, T. P., Schulz, M. R., & Della Volpe, A. (2000). Analyzing English grammar (3rd ed.). Needham Heights, MA: Allyn & Bacon.

Klein, D. & Manning, C. (2014). The Stanford Parser: A statistical parser. Version 3.5.0. http://nlp.stanford.edu/software/lex-parser.shtml.

Knaus, K., Murphy, K., Blecking, A., & Holme, T. (2011). A valid and reliable instrument for cognitive complexity rating assignment of chemistry exam items. *Journal of Chemical Education, 88*(5), 554–560.

Kopriva, R. J., & Winter, P. C. (2012). *Designing a cognitive skills framework for item development*.

Krippendorff K. (2004). Reliability in Content Analysis: Some Common Misconceptions and recommendations. *Human Communication Research, 30*(3), 411-433.

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

Li, M. (2001). A framework for science achievement and its link to test items. (Unpublished doctoral dissertation). Stanford University, Palo Alto, CA.

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review*, *78*(2), 333-368.

Trumbull, E., Nelson-Barber, S., & Huang, C.W. (2016, April). Exploring the Role of Linguistic Factors in the Performance of English Learners on Science Assessments. Paper presented at the annual meeting of the American Evaluation Research Association, Washington, DC.

Oosterhof, A., Rohani, F., Sanfilippo, C., Stillwell, P., & Hawkins, K. (2008). *The capabilities-complexity model* (No. 108). Center for Advancement of Learning and Assessment: Florida State University.

Shavelson, R., Ruiz-Primo, M., Li, M., & Ayala, C. C. (2003). Evaluating new approaches to assessing learning. (CSE Report 604). National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles. Retrieved from http://www.cse.ucla.edu/products/Reports/R604.pdf

Solano-Flores, G., Trumbull, E., & Kwon, M. (2003, April). The metrics of linguistic complexity and the metrics of student performance in the testing of English language learners. Paper presented at the annual meeting of the American Evaluation Research Association, Chicago, IL.

Wang, T., Liaw, Y.-L., Li, M., & Taylor, C. (2014). Identifying science item context characteristics for English Language Learners (ELLs) and non-ELLs by Differential Item Functioning (DIF). Presented at the Annual Conference of the American Education Research Association, Philadelphia, PA.

Webb, N. L. (2002). *Depth-of-knowledge levels for four content areas.* Unpublished manuscript.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P.W. Holland & H. Wainer (Eds.) *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Erlbaum.

Zwick, R., Thayer, D.T., and Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement*, *36*(1), 1-28.