

UC Berkeley

UC Berkeley Previously Published Works

Title

Exploring the item features of a science assessment with complex tasks

Permalink

<https://escholarship.org/uc/item/1bz699tf>

Authors

Collier, Tina
Morell, Linda
Wilson, Mark

Publication Date

2018

DOI

10.1016/j.measurement.2017.08.039

Peer reviewed

Exploring the Item Features of a Science Assessment with Complex Tasks

Tina Collier^{a1}, Linda Morell^a, & Mark Wilson^a
Berkeley Evaluation and Assessment Research (BEAR) Center
University of California, Berkeley

^aBerkeley Evaluation and Assessment Research (BEAR) Center
University of California, Berkeley
2000 Center Street, Suite 301
Berkeley, CA 94704
USA

Corresponding Author: Tina Collier (colliert@ada.org)

Item Features of a Science Assessment

¹ Present Address: 211 East Chicago Avenue, Chicago, IL 60611

Abstract

Item explanatory models have the potential to provide insight into why certain items are easier or more difficult than others. Through the selection of pertinent item features, one can gather validity evidence for the assessment if construct-related item characteristics are chosen. This is especially important when designing assessment tasks that address new standards. Using data from the Learning Progressions in Middle School Science (LPS) project, this paper adopts an “item explanatory” approach and investigates whether certain item features can explain differences in item difficulties by applying an extension of the linear logistic test model. Specifically, this paper explores the effects of five features on item difficulty: type (argumentation, content, embedded content), scenario-based context, format (multiple-choice or open-ended), graphics, and academic vocabulary. Interactions between some of these features were also investigated. With the exception of context, all features had a statistically significant effect on difficulty.

Keywords: linear logistic test model, item explanatory models, item features

1 Introduction

This paper adopts an “item explanatory” approach, where the focus is on investigating whether certain item features can explain differences in item difficulties by applying an extension of the linear logistic test model (LLTM; Fischer, 1973) to a middle school science assessment that was designed to follow the Next Generation Science Standards (NGSS; NGSS Lead States, 2013). Explanatory item response models (EIRMs; De Boeck & Wilson, 2004) have the potential to provide explanations for the item responses, unlike descriptive models where item responses are merely described by the estimated parameters. While more traditional models output a list of estimated item difficulties, an “item explanatory” approach results in a list of estimated difficulties for item features. These item features must be selected a priori and, if content-related features are chosen, have the potential to provide strong content validity support for an assessment. Specifically, this paper explores the effects of five features on item difficulty: type, context, format, graphics, and academic vocabulary.

1.1 The Next Generation Science Standards (NGSS)

The NGSS is a U.S. initiative designed to increase understanding of science, create common standards for teaching across the U.S., and develop more interest in science in school-age students in the hopes that more of them will major in a science-related area of study in college. The NGSS provides performance expectations to reflect a reform in science education that includes three dimensions: (1) developing disciplinary core ideas (DCI), (2) linking these core ideas across disciplines or crosscutting concepts, and (3) engaging students in scientific and engineering practices—based on contemporary ideas about what scientists and engineers do. The emphasis, in particular, is on integrating these three dimensions so that core ideas are not taught in isolation, but connect to larger ideas that also involve real-world applications. Rather than learn a wide breadth of disconnected content topics, the goal is to develop a deeper understanding of a few core ideas that set a strong foundation for all students after high school. The Learning Progressions in Middle School Science (LPS) project, described in the next section, examined two of these three dimensions and designed an assessment to reflect their integration.

1.2 Item Features for the Learning Progressions in Middle School Science (LPS) Project

One of the main research goals for the Learning Progressions in Middle School Science (LPS) project², was to explore the relationship between science content knowledge, a DCI, and scientific argumentation, a scientific practice. To further explore this relationship, the assessment was divided into three “complex tasks”, which consist of three item *types*: (1) argumentation items assessing argumentation competency in a specific scientific context (e.g. two students arguing over what happens to gas particles placed in a container), (2) content science items embedded within the same scientific context (e.g. what happens when you insert gas particles into a sealed container), and (3) content science items assessing knowledge of other concepts in the same science domain but not so closely associated with the context (e.g. compare the

² Details of the project are not discussed here, though several resources are available for interested readers (Osborne et al., 2013a; Osborne et al., 2013b; Wilson, Black, & Morell, 2013; Yao, 2013; Yao, Wilson, Henderson, & Osborne, 2015).

movement of liquid water molecules with the movement of ice molecules). In this paper, these three item types are referred to as ‘argumentation’, ‘embedded content’, and ‘content.’ Examples are provided in Figures 1 through 3.

The ‘complex tasks’ are each set within a *context*—that is all the items within a complex task share a common setting. These contexts are what happens when someone (a) chops onions, (b) inserts gas particles into a container, and (c) mixes sugar into a glass of water. These contexts will be referred to as ‘Onions’, ‘Gases’, and ‘Sugar’ for easier reference. The embedded content and argumentation items were presented in these contexts, while the content items were related (i.e., they were more generalized but about the same concepts). Note that while the context of the content items are more generalized, they were still designated into a context by the test developers.

Mark and Kian are discussing what happens when they chop onions. They have two different ideas.

Mark says:
Chopping onions makes me cry because when I cut the onion, some gas is released. The gas goes into the air and gets into my eyes.

Kian says:
I disagree. Chopping onions makes you cry because when the knife slices the onion, some liquid squirts out of the onion and into your eyes.

What is Mark's idea about why people cry when they cut onions?

Mark's idea is that...

Figure 1. An argumentation item from the *Onions* complex task.

Have you ever noticed that when people chop onions they look like they are crying?

In the space below, explain how you think a chemical from the onion could get into a person's eye.

A chemical could get into a person's eye by...



Figure 2. An embedded content item from the *Onions* complex task.

Describe the arrangement of molecules in ice, liquid water, and water vapor.

The arrangement of molecules in ice is...

- Packed closer together than liquid water and in a repeating pattern
- Spread further apart than liquid water and in a repeating pattern.
- Packed closer together than liquid water and in a random pattern.
- Spread further apart than liquid water and in a random pattern.

Figure 3. A content item from the *Onions* complex task.

The remaining three item features explored in this paper are often tested in psychometric studies to examine whether they have an unintended effect on the item difficulties. For instance, the *format* refers to whether an item is open-ended or forced-choice (e.g., multiple-choice). Previous studies have suggested that multiple-choice items are easier for students than open-ended ones (Hohensinn & Kubinger, 2011; Kubinger, 2008). Because the assessment includes a combination of both, this feature is investigated to see this finding holds true for the LPS data.

The *graphics* feature includes three categories: schematic representations, pictorial representations, and no graphics. Schematic representations are defined as abstract pictures whose “schematic meaning is provided by the symbolic/visual representation in the item” (Martinello, 2009, pp. 166). An example would include an image of the movement of gas particles. This contrasts with pictorial representations, which are concrete images that simply illustrate the details of objects described in an item.

Lastly, whether an item contains *academic words* is explored. Academic vocabulary words are those that are not among the 2,000 most common words and occur most often in academic texts (Coxhead, 2000). Unlike technical vocabulary—which are the specialized words

specific to a discipline, academic vocabulary words are more generalized and span across many content areas (Stevens, Butler, & Castellon-Wellington, 2001). This distinction is important for many studies investigating the language effects of content assessments because while technical vocabulary is deemed to be construct-relevant, academic vocabulary is often seen as construct-independent and, subsequently, may interfere with the interpretations of student scores on assessments (Avenia-Tapper & Llosa, 2015; Haag, Roppelt, & Heppt, 2015; Wolf and Leon, 2009). Coxhead’s (2000) academic word list (AWL) is used here to identify academic words on the assessment³. Note that the word “evidence” is on the AWL, but will not be counted as an academic word in this paper because “evidence” is central to the argumentation construct. Thus, “evidence” is deemed to be construct-relevant, whereas other words on the AWL may be considered construct-independent.

1.3 Research Questions

This paper explores the effect of each item feature on the overall item difficulty. Specifically, the research questions are:

- RQ1. Which of the following, if any, item features—type, context, format, inclusion of graphics, and inclusion of vocabulary from the Academic Word List (AWL)—contribute to the explanation of item difficulty?
- RQ2. Does the item type feature interact with any of the other features to have a statistically significant effect on the item difficulties?

Because of the added complexities in interpreting interactions, only one feature was explored in detail for RQ2. Item type was chosen because it directly relates to the content of the items and is also related to the main research goal of LPS—exploring the relationship between content knowledge and argumentation.

2 Methods

2.1. Sample

In the spring of 2014, a total of 282 eighth and tenth grade students from a large urban U.S. school district took the assessment during regular school hours on a computer for one class period. Four students have provided no information for the test (i.e. all missing responses), leaving a final sample of 278 students.

The sample consisted of 119 grade 8 and 159 grade 10 students from 3 schools. Demographic information was missing for one student. There were more girls (n=172) than boys (n=105). A high percentage of this group of students were classified as gifted students (n=169, 60.79%), as this was a convenience sample. Eleven (3.96%) were classified as special education students.

³ Cobb’s website *Web Vocabprofile Classic* at <http://www.lex tutor.ca/vp/eng/> automatically sorts texts and provides counts for four types of words: the 1000 most frequent words, 1001-2000 most frequent words, words on Coxhead’s (2000) Academic Word List, and off-list words.

2.2. Instrument

A subset⁴ of the original 2013-2014 LPS items were used, for a final total of 39 items across the three complex tasks. Table 1 illustrates the distribution of these final item types across the three different contexts – Onions, Gases, and Sugar. There were 20 argumentation items, 7 embedded content, and 12 content items. All embedded content and content items were scored following the content learning progression, while all argumentation items were scored following the argumentation learning progression. A brief description of the two learning progressions is provided in the Appendix. Readers interested in learning more about the progressions can also refer to additional resources (Osborne et al., 2013a; Osborne et al., 2013b; Wilson, Black, & Morell, 2013; Yao, 2013; Yao, Wilson, Henderson, & Osborne, 2015).

Table 1

Distribution of types of items across the three complex tasks

Task	Argumentation	Embedded Content	Content	TOTAL
Onions	9	1	4	14
Gases	4	3	6	13
Sugar	7	3	2	12
TOTAL	20	7	12	39

2.3. Model

The linear logistic test model (LLTM; Fischer, 1973) decomposes the item difficulty into a linear combination of item features. Because the items in the data are polytomously scored, the extension for the LLTM is described here. This extension is sometimes referred to as the linear partial credit model (Fischer & Ponocny, 1994) which is also similar to the multifacet Rasch model (Linacre, 1989). This model builds from the partial credit model (Masters, 1982), which models the log odds of the probability that student p with ability θ_p will respond in category j instead of category $j-1$ on item i :

$$\frac{P(X_i=j|\theta_p)}{P(X_i=j-1|\theta_p)} = \exp(\theta_p - \delta_{ij}) \quad (1)$$

where δ_{ij} is a parameter for the difficulty for step j of item i and $\theta_p \sim N(0, \sigma_{\theta_p}^2)$.

⁴ The original data included two content constructs, Macroscopic Properties and Particulate Explanations of Physical Changes. There were six content and no embedded content items related to Macroscopic Properties. Thus, to avoid any misleading conclusions when exploring the difficulties of content and embedded content items, the Macroscopic Properties items were removed.

The LLTM, on the other hand, models δ_{ij} differently, while all other terms stay the same. Specifically, δ_{ij} is defined as:

$$\delta_{ij} = \delta_i + \tau_{ij} \quad (2)$$

and

$$\delta_i = \sum_{m=1}^M \beta_m X_{\zeta} \quad (3)$$

where δ_i is the item difficulty, τ_{ij} is the step deviate parameter, X_{ζ} is the value of item i on feature m , and β_m is the regression weight for item feature m . Notice that the item step parameter, δ_{ij} from (1) is replaced with the linear combination of the difficulties for the item features. To answer RQ1, the overall item difficulty across the steps can be written as:

$$\delta_i = \beta_1 \text{type}_i + \beta_2 \text{task}_i + \beta_3 \text{format}_i + \beta_4 \text{graphics}_i + \beta_5 \text{AWL}_i \quad (4)$$

where the coefficients for each feature determines if the item becomes easier or more difficult.

For RQ2, interaction effects are added to (4). As an example, (5) shows the overall item difficulty for a model that includes interaction terms for type and all other features.

$$\delta_i = \beta_1 \text{type}_i + \beta_2 \text{task}_i + \beta_3 \text{format}_i + \beta_4 \text{graphics}_i + \beta_5 \text{AWL}_i + \beta_6 \text{type}_i * \text{task}_i + \beta_7 \text{type}_i * \text{format}_i + \beta_8 \text{type}_i * \text{graphics}_i + \beta_9 \text{type}_i * \text{AWL}_i \quad (5)$$

Because (4) can be found by constraining some of the parameters in (5) (i.e. $\beta_k = 0$, for $k > 5$), these two models can be directly compared using a likelihood ratio test. ConQuest 3.0 (Adams, Wu, & Wilson, 2012) was used for all analyses.

3 Results

3.1 Item Analysis

Table 2 provides the frequencies for each of the five features on the assessment. For type, the items are not distributed evenly across the three categories. Argumentation items make up approximately half of the items, while only 17.95% of the items are categorized as embedded content. On the other hand, the items are about evenly distributed for the item contexts, format, and graphics. Only 13 items contained academic words.

Table 2

Frequencies for Each Item Feature on the LPS Assessment and the LLTM Result (RQ1)

Item Feature	Count	Percentage	Estimate (SE)	p
<i>Type</i>				<0.001
Argumentation	20	51.28	0.07 (0.03)	0.02
Embedded Content	7	17.95	-0.49 (0.04)	<0.001
Content	12	30.77	0.43* (0.04)	<0.001
<i>Context</i>				0.11
Sugar	12	30.77	-0.07 (0.04)	0.08
Onions	14	35.90	0.01 (0.03)	0.74
Gases	13	33.33	0.06*(0.03)	0.05
<i>Format</i>				<0.001
Multiple-Choice	19	48.72	0.75 (0.02)	<0.001
Open-Ended	20	51.28	-0.75*(0.02)	<0.001
<i>Graphics</i>				<0.001
Schematic	11	28.21	0.06 (0.03)	0.05
Pictorial	14	35.90	0.24 (0.04)	<0.001
None	14	35.90	-0.31* (0.03)	<0.001
<i>Academic Words List (AWL)</i>				<0.001
Yes[†]	13	33.33	-0.15 (0.02)	<0.001
No	26	66.67	0.15* (0.02)	<0.001

Note[†]: “Yes” means that an item contains at least one AWL word. It does not account for the number of AWL words in an item.

Note*: Indicates the result is constrained for the model to be identified. In ConQuest, this is done by setting the last category for each feature to be equal to the negative sum of all other categories for that feature.

To get a better sense of the items, a partial credit analysis was conducted. The coefficient alpha was 0.84 and the EAP reliability was 0.83. Overall, the items appeared to fit well, as the weighted mean square fit statistics ranged from 0.85 to 1.20 (Wu, Adams, Wilson, & Haldane, 2007). Figure 4 is the Wright Map for this analysis.

The Wright map has two distinct sections: the student ability and the item distribution. Both of these distributions use the same scale, the logit scale. The student ability distribution is shown on the left column, while the item thresholds are shown on the right. The logit scale is located on the far right of the map. The numbers to the left of the icons for each item indicates the x^{th} threshold for that item. From Figure 4, it appears that there were few items that matched the students at the top end of the distribution. For these students, the items were most likely easy for them. However, for the rest of the students, there seemed to be good balance of items that matched the student abilities.

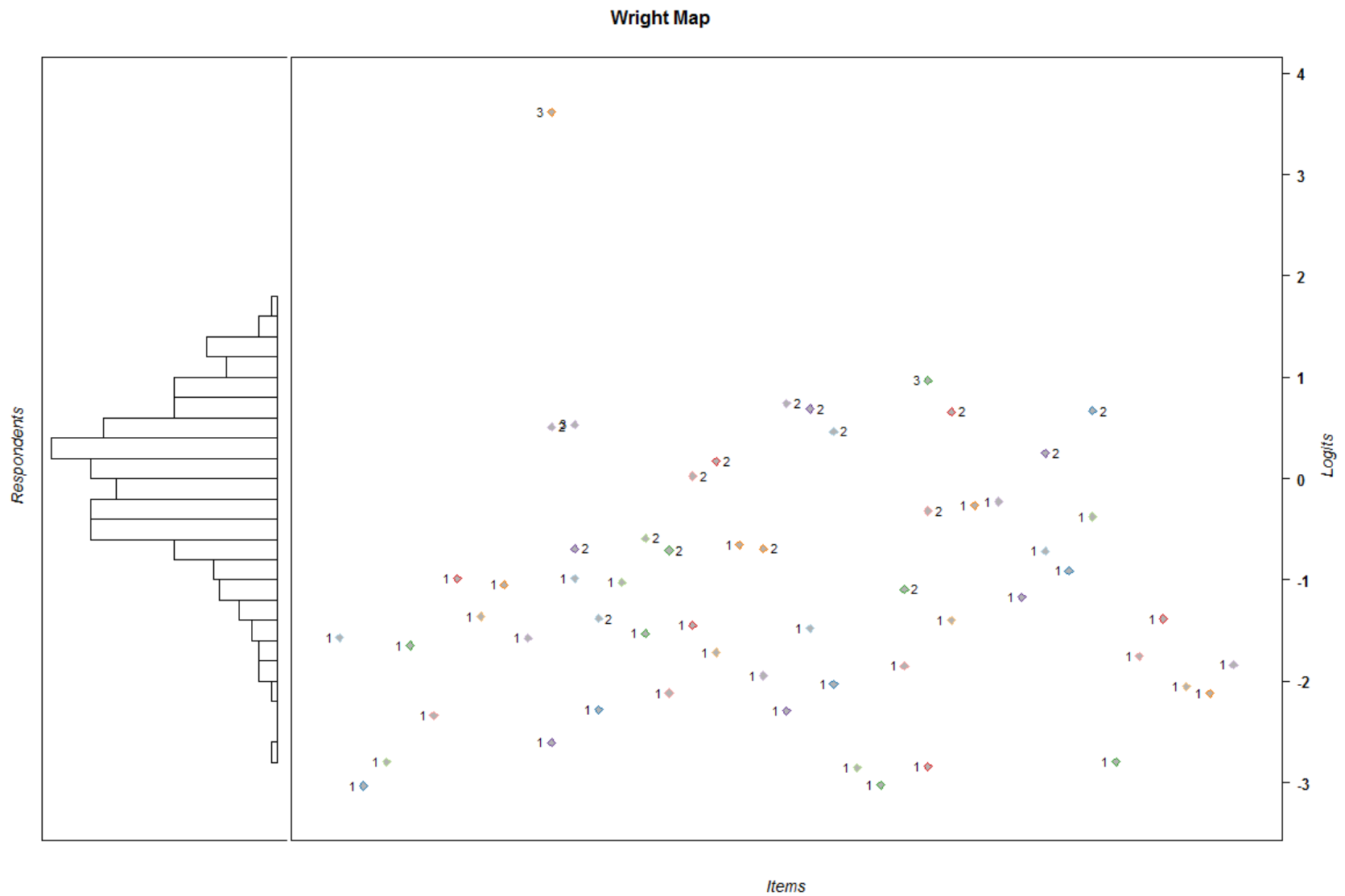


Figure 4. Wright Map from the partial credit analysis. Generated from the WrightMap package (Torres Iribarra & Freund, 2016).

3.2 Results for RQ1: LLTM

Results from the LLTM analysis are shown in Table 2. For the type feature, the embedded content items were estimated to be the easiest, while the content items were identified as the most difficult. Argumentation items lie somewhere in between. The results suggest that the variation across types is statistically significant ($\chi^2=175.30, df=2, p<0.001$).

The variation between context features, on the other hand, is much smaller—maybe even non-existent. These results suggest that the item contexts do not vary in their difficulty ($\chi^2=4.37, df=2, p=0.11$). This finding is reassuring since the interest was never on how well students perform on items about the specific topics of dissolving sugar or chopping onions.

Surprisingly, the multiple-choice feature was estimated to increase the difficulty of an item whereas the open-ended feature was estimated to decrease the difficulty. This result may be due to the fact that some open-ended items required students to carry out a simple operation such as “identify a claim,” while other items required students to carry out more complex operations like “explain how the evidence supports your answer.” These two open-ended items should differ in their difficulty—the first is simply writing down a claim while the latter requires a more thoughtful explanation. These differences are not accounted for in the LLTM and this shows one of the limitations of decomposing item difficulties into only a handful of features.

For the graphics feature, pictorial representations were found to be the most difficult, followed by the schematic representations, and lastly by no graphics in the items. However, note that the value for schematic representation is low, suggesting that it has little to no effect on item difficulty. There is significant variation for this feature ($\chi^2=49.29, df=2, p<0.001$).

Items with AWL words were estimated to decrease the item difficulty by about 0.15 logit, while items with no words from the AWL were estimated to increase the item difficulty by about 0.15 logit. This difference was statistically significant ($\chi^2=51.19, df=1, p<0.001$). It is unusual that items with words from the AWL would be easier than items without words from the list and this finding requires more investigation.

3.3 Results for RQ2: LLTM with Interactions

To answer RQ2, a LLTM with interactions model was applied. Because the context item feature was not statistically significant from the simple LLTM, it was excluded from this model. Two interaction terms were added: *type*format* and *type*AWL*. Interactions between types with graphics were not modeled because all embedded content items had some sort of graphic representation and none of the content items had pictorial representations. The results are shown in Table 3.

Table 3

Results for RQ2: LLTM with Interactions

Item Feature	Estimate (SE)	p	Feature Interaction	Estimate (SE)	p
<i>Type</i>		<0.001	<i>Type*AWL</i>		<0.001
Argumentation (ARG)	-0.18 (0.03)	<0.001	ARG*Yes	-0.30 (0.03)	<0.001
Embedded (EMB)	-0.29 (0.05)	<0.001	EMB*Yes	0.23 (0.04)	<0.001
Content (CON)	0.47* (0.04)	<0.001	CON*Yes	0.06* (0.03)	0.05
			ARG*No	0.30* (0.03)	<0.001
<i>Format</i>		<0.001	EMB*No	-0.23* (0.04)	<0.001
Multiple-Choice (MC)	0.75 (0.03)	<0.001	CON*No	-0.06* (0.03)	0.05
Open-Ended (OE)	-0.75* (0.03)	<0.001			
<i>Graphics</i>		<0.001	<i>Type*Format</i>		<0.001
Schematic	0.06 (0.03)	0.05	ARG*MC	-0.02 (0.03)	0.50
Pictorial	0.25 (0.03)	<0.001	EMB*MC	0.26 (0.04)	<0.001
None	-0.31* (0.03)	<0.001	CON*MC	-0.25* (0.04)	<0.001
			ARG*OE	0.02* (0.03)	0.50
<i>AWL</i>		0.58	EMB*OE	-0.26* (0.04)	<0.001
Yes	-0.01 (0.02)	0.62	CON*OE	0.25* (0.04)	<0.001
No	0.01* (0.02)	0.62			

Note*: Indicates the result is constrained for model to be identified.

The interaction term *type*format* was statistically significant ($\chi^2=51.03, df=2, p<0.001$). Multiple-choice embedded content items contribute an additional 0.26 logit to the item difficulty, while multiple-choice content items are estimated to decrease the item difficulty by 0.25 logit. For argumentation items, it appears that format does not contribute much, if anything, to the item difficulty (i.e. the estimated parameter is small).

The interaction term *type*AWL* was also statistically significant ($\chi^2=130.31, df=2, p<0.001$). Argumentation items with AWL words are estimated to decrease item difficulty by 0.30 logit, whereas embedded content items are estimated to contribute an additional 0.23 logit to the item difficulty. Lastly, content items with AWL words are estimated to contribute an additional 0.06 logit to the item difficulty. The interaction term shows some of the nuances of the effect of AWL—that is, with embedded content items, items containing words from the AWL are estimated to be slightly more difficult. However, for argumentation items, the opposite is true. Perhaps the conversational format of the argumentation items is easier for comprehension even with the inclusion of AWL words. Whether or not a content item contains AWL words does not seem to contribute much to the item difficulty, as the estimated parameter is small. This result could be due to the familiarity of content items for students, as they may be accustomed to encountering academic words in these types of items.

3.4 Post-Hoc Analysis on Academic Word List Feature

A post-hoc analysis on the academic words appearing in the three types of items was conducted to learn more about its effect. Figure 5 lists the words from the AWL found on the assessment. From the list, most of the AWL words are on content items and include words like: energy, structure, spheres, volume, predict, and contact. Only one word appears on the embedded content items: chemical. Lastly, four words appear on the argumentation items: released, chemicals, selected, and created. The word family “chemical” is the only AWL word that is present on both embedded content and argumentation items. “Chemical” is also the only academic word listed in the embedded content items—and it is on one of the more difficult ones too. This word is interesting because it is on the content progress map, so its difficulty may be considered partially construct-relevant for the embedded content item—though maybe appropriately not so for argumentation.

This leads to a possibility of two academic word lists: (a) one where the academic jargon is *specific* to the content of the items, and (b) one where the AWL is *general* across contexts. Then, the following words can be categorized into group (a): energy, structure, spheres, chemical, released, and volume, and group (b): selected, predict, contact, and created. Another LLTM analysis was run with no interactions and using this categorization for AWL. The results are shown in Table 4. Because context was not statistically significant in the previous LLTM analysis, this variable was removed from the model.

Academic Word List	Item Type
Energy	Content
Structure	Content
Spheres	Content
Chemical	Embedded Content, Argumentation
Released	Argumentation
Volume	Content
Selected	Argumentation
Predict	Content
Contact	Content
Created	Argumentation

Figure 5. The academic words on the LPS test and the corresponding item type. Coxhead's (2000) Academic Word List (AWL) was used.

This new division shows that items with AWL words specific to the science context of the items is associated with a difficulty increase of approximately 0.21 logit, whereas the presence of those that are more general across contexts is associated with a difficulty decrease of approximately 0.45 logit. In contrast, the absence of AWL words is associated with a higher difficulty by about 0.24 logit. These results show some of the nuances of AWL words, especially those used in a science context which have specific scientific meanings. More general AWL words were found to be associated with lower difficulty of an item by almost half a logit. However, it is still unclear why having no AWL words would be predicted to increase the difficulty of an item compared to the general AWL words.

Table 4

Post-hoc LLTM analysis with updated AWL category

Item Feature	Estimate (SE)	p
<i>Type</i>		<0.001
Argumentation	0.14 (0.03)	<0.001
Embedded Content	-0.57 (0.03)	<0.001
Content	0.43* (0.03)	<0.001
<i>Format</i>		<0.001
Multiple-Choice	0.74 (0.02)	<0.001
Open-Ended	-0.74* (0.02)	<0.001
<i>Graphics</i>		<0.001
Schematic	0.14 (0.03)	<0.001
Pictorial	0.22 (0.03)	<0.001
None	-0.36* (0.03)	<0.001
<i>Academic Words List (AWL)</i>		<0.001
Specific	0.21 (0.04)	<0.001
General	-0.45 (0.04)	<0.001
None	0.24* (0.03)	<0.001

Note*: Indicates the result is constrained for the model to be identified.

4 Discussion

The purpose of this paper was to investigate how well certain item features can explain the item difficulties on the LPS assessment. There are several findings worth mentioning here. First, related to RQ1, item context was not a statistically significant predictor for item difficulties. This is reassuring, since the context was not part of the intended construct. In fact, these contexts were chosen in the hope that they would be familiar enough for students to limit construct-irrelevant variance. The other features (type, format, graphics, and AWL words) were statistically significant.

In addition, some features were flagged for having large fit statistics. For instance, open-ended items (part of the format feature) were identified as having more variation than predicted.

This could be due to the fact that some open-ended items required much shorter responses, while others required a more detailed explanation. Perhaps a finer-grained distinction between different types of open-ended items may be useful for future analyses. This could even be done with the multiple-choice items, where the number of response options varied from two to five. For this assessment, two items had two response options, two had five response options, and 15 had four response options.

While this study had unexpected results (e.g. the effect of AWL words), the LLTM with interactions showed the nuances of some of the effects. Some effects should be interpreted with caution, since the LLTM can greatly reduce the number of estimated parameters (e.g. an original 39 items reduced to 5 item features). Incorporating interactions among the features can help with identifying the feature effects in more detail.

The interaction effects of *type*AWL* are noteworthy. At first, it seemed strange that the inclusion of academic words in items would decrease the difficulty. However, by examining the interaction effects, we found that this was only true for argumentation items. For the embedded content items, inclusion of AWL words increased the difficulty as we might have expected. It is unclear why the trend is different for argumentation items. One possibility may be due to subject-matter—argumentation items may appear more conversational (in general) than the other two item types, hence the inclusion of academic words may not actually interfere with understanding the item. Of course, further investigations into this finding is needed.

The post-hoc LLTM analysis, with the division of AWL words into specific and general categories, shows that the presence of AWL specific words is expected to contribute to the difficulty of an item, as does the absence of AWL words. The presence of AWL general words was found to actually be associated with easier items. Future analyses examining AWL words in specific science contexts will be needed to provide more information into how these words may affect item difficulty.

There are many possibilities for future explorations, especially if there is another round of data collection with this particular assessment. One easy extension is to add other item features that may have predictive ability for estimating difficulty. Differential facet functioning (DFF; Xie & Wilson, 2008) is another possible extension. For future studies, if there were a more diverse sample with distinct groups to explore, then DIF and DFF can provide powerful explanatory information to the LPS items and constructs.

Acknowledgements

The work was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A100692 to University of California, Berkeley. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Appendix

A.1 Content Learning Progressions from the LPS Project

The structure of matter learning progression is hypothesized to include six related, but distinct constructs. Shown in Figure A.1, the constructs for this progression are represented by boxes with the arrows pointing towards more sophisticated constructs. Thus, the progression starts at the bottom, with Macro Properties (MAC) as the easiest, followed by the Changes of State and other Physical Changes (PHS), and ends with Particulate Explanations of Physical (EPC) and Chemical (ECC) Changes as the most difficult. Two additional constructs, Measurement and Data Handling (MDH) and Density (DMV), were identified as auxiliary constructs—constructs that aid in the understanding of the four core ones but not necessarily central. This classification was helpful because, due to time and resource constraints, not all constructs could be investigated in great detail. This allowed the research team to prioritize and gather high quality empirical evidence for the constructs of most interest. Although not illustrated in Figure A.1, each construct contains more detailed descriptions, called construct maps. Each construct map covers increasingly sophisticated descriptions of student thinking in these areas.

This study used items from only one content construct: Particulate Explanation of Physical Changes (EPC). This was chosen because concepts from this construct fit well with the argumentation items on the assessment, as they covered similar ideas. The construct map for EPC is shown in Figure A.2. EPC contains two strands, *A: molecular models of physical changes* and *B: molecular representations of different states of matter*. Strand A consists of three sub-strands, describing phenomena for mixing and dissolving, compression and gases, and phase change and heating. Strand B consists of two sub-strands, density and arrangements and movements. Both strands contain three levels; Level 1 describes the simpler levels of understanding within each sub-strand, whereas Level 3 describes the more complex and sophisticated understandings within each sub-strand.

A.2 Scientific Argumentation Learning Progression from the LPS Project

In addition to the content learning progression, a separate progression was developed for scientific argumentation, which is shown in Figure A.3. Unlike for content, this learning progression reads from top to bottom, with the less sophisticated argumentation practices listed at the top and the most sophisticated at the bottom. It is based on Toulmin's (1958) model of argumentation and contains three main elements: claims, evidence, and warrants. *Claims* are statements that an arguer states is true. *Evidence* are the data used to support these claims and these depend on the *warrants*, or the explanations of how the evidence supports the claims.

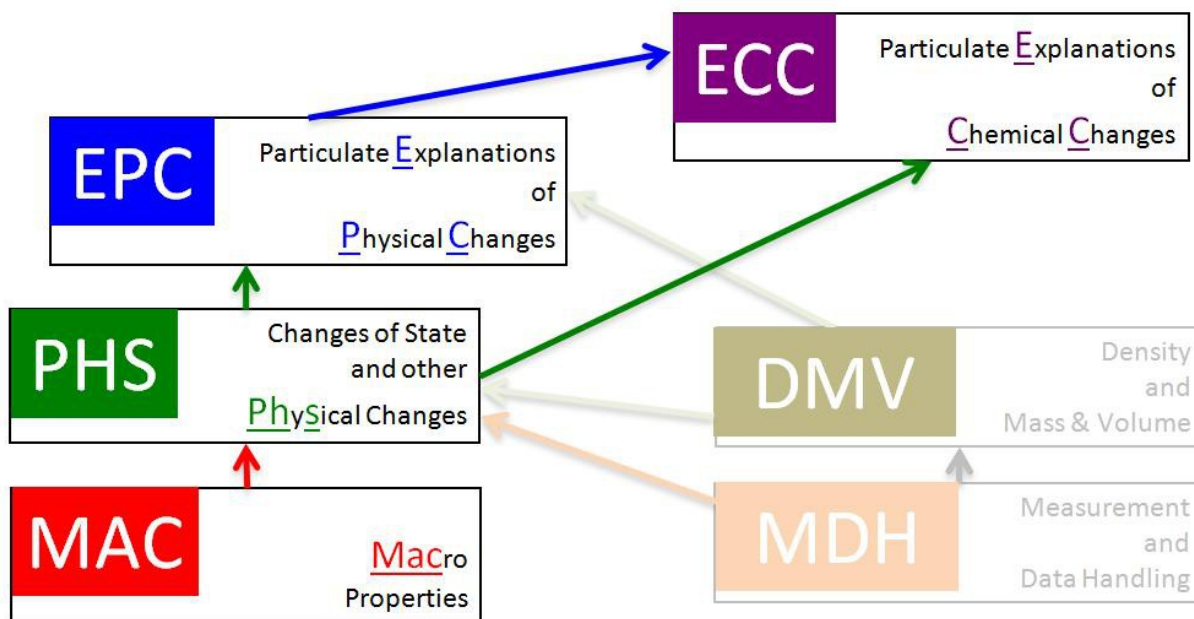





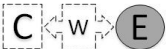
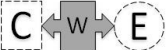
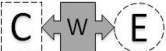

Figure A.1. Structure of matter learning progression from the Learning Progressions in Science (LPS) Project.

The first column in the progression represents the three distinct levels (Levels 0, 1, and 2), each with its own sublevels (e.g. Levels 1a, 2a). Like for content, higher numbers represent more difficult practices and a deeper understanding of the area. The second and third columns represent whether an argument requires students to construct ones' own element or critique someone else's, while the fourth column includes a description of the level. These columns are based on the notion that argumentation is a dialectic between construction and critique (Ford, 2008). The construction of scientific claims, for instance, are subject to the critique and scrutiny by the community. Scientists often engage in both practices. In some more difficult levels (e.g. Level 2A), both of these skills—constructing and critiquing—are required.

After some earlier analyses were completed, the research team decided to incorporate cognitive load theory into this progression as well. The idea is that the more elements that are required in an argument, the more sophisticated argumentation skills are required. The last column in the learning progression provides a visual representation of this addition. The grayed figures indicates which element is needed to successfully argue at a certain level and one can observe that the highest level in this progression also contains the most required elements.

Levels	EPC Strand A : Molecular models of physical change			EPC Strand B: Molecular representations of different states of matter	
	Mixing and Dissolving	Compression of gases	Phase change and heating	Density	Arrangements and Movements
3			Can explain that when a solid or liquid is heated, it occupies more volume because of the faster movements of molecules	Can explain why with the same number of different molecules in the same volume, the densities of the two materials cannot be the same. Can explain why with different numbers of the same molecules in a given volume, the two materials cannot have the same density. Can explain why with different numbers of different molecules in a given volume, the two materials can have the same density.	Knows that in ice the spaces between the molecules are empty.
2	Knows that, when two different substances are mixed, the molecules of the substances mix together at random Knows that when sugar is dissolved in water, the sugar can't be seen because it has split up and the pieces are mixed in the water.	Knows that when a volume of gas is compressed (or expanded), the molecules move closer together (or further apart) and are still distributed at random, but the molecules do not change their size or their mass	Knows that in phase changes, the molecules speed up - from solid to liquid and from liquid to gas	Can give a partial explanation why with the same number of different molecules in the same volume, the densities of the two materials cannot be the same. Can give a partial explanation why with different numbers of the same molecules in a given volume, the two materials cannot have the same density. Can give a partial explanation why with different numbers of different molecules in a given volume, the two materials can have the same density.	Can explain effects of the free movement of gas particles. Can describe the movements of molecules in ice, water and water vapor. Knows that the particles of a gas move freely to fill any space.
1	Knows that when a substance is dissolved, the substance's mass is conserved.	Knows that when a gas is compressed (or expanded), the number of molecules in that gas does not change		Can recognize that with the same number of different molecules in the same volume, the densities of the two materials cannot be the same.	Can describe the arrangements of molecules in ice, water and water vapor.

Figure A.2. Construct map for Particulate Explanations of Physical Changes (EPC).

Lev.	Constructing	Critiquing	Description	Representation of elements
0			No evidence of facility with argumentation.	
0a	Constructing a claim		Student states a relevant claim.	
0b		Identifying a claim	Student identifies another person's claim.	
0c	Providing evidence		Student supports a claim with a piece of evidence.	
0d		Identifying evidence		
1a	Constructing a warrant		Student constructs an explicit warrant that links their claim to evidence.	
1b		Identifying a warrant	Student identifies the warrant provided by another person.	
1c	Constructing a complete argument		Student makes a claim, selects evidence that supports that claim, and constructs a synthesis between the claim and the warrant.	

1d	Providing an alternative counter argument	Student offers a counterargument as a way of rebutting another person's claim.	
2a	Providing a counter-critique	Student critiques another's argument. Fully explicates the claim that the argument is flawed and <i>justification</i> for why that argument is flawed.	
2b	Constructing a one-sided comparative argument	Student makes an evaluative judgment about the merits of two competing arguments and makes an explicit argument for the value of <i>one</i> argument. No warrant for why the other argument is weaker.	
2c	Providing a two-sided comparative argument	Student makes an evaluative judgement about two competing arguments and makes an explicit argument (claim + justification) for why one argument is stronger and why the other is weaker (claim + justification).	
2d	Constructing a counter claim with justification	This progress level marks the top anchor of our progress map. Student explicitly compares and contrasts two competing arguments, and also constructs a new argument in which they can explicitly justify why it is superior to each of the previous arguments.	

Figure A.3. Argumentation learning progression from the Learning Progressions in Science Project (LPS).

References

- Adams, R. J., Wu, M., & Wilson, M. (2012). *ConQuest 3.0* [computer program]. Hawthorn, Australia: ACER.
- Avenia-Tapper, B. & Llosa, L. (2015). Construct relevant or irrelevant? The role of linguistic complexity in the assessment of English language learners' science knowledge. *Educational Assessment, 20*(2), 95–111. <http://doi.org/10.1080/10627197.2015.1028622>
- Cobb, T. (n.d.) Web Vocabprofile. Retrieved from <http://www.lex tutor.ca/vp/>
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238.
- De Boeck, P. & Wilson, M. (2004). A framework for item response models. In P. De Boeck & Wilson, M. (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 3-41). New York: Springer.
- Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*(6), 359–374.
- Fischer, G. H. & Ponocny, I. (1994). An extension of the partial credit model with an application to the measurement of change. *Psychometrika, 59*(2), 177-192.
- Ford, M. (2008). Disciplinary authority and accountability in scientific practice and learning. *Science Education, 92*(3), 404–423. <http://doi.org/10.1002/sce.20263>
- Haag, N., Roppelt, A. & Heppt, B. (2015). Effects of mathematics items' language demands for language minority students: Do they differ between grades? *Learning and Individual Differences, 42*, 70–76. <http://doi.org/10.1016/j.lindif.2015.08.010>
- Hohensinn, C. & Kubinger, K. D. (2011). Applying item response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement, 71*(4), 732–746. <http://doi.org/10.1177/0013164410390032>
- Kubinger, K. D. (2008). On the revival of the Rasch model-based LLTM: From constructing tests using item generating rules to measuring item administration effects. *Psychology Science, 50*(3), 311–327.
- Linacre, J. M. (1989). *Multi-facet Rasch measurement*. Chicago: MESA Press.
- Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational Assessment, 14*(3-4), 160–179. [doi:10.1080/10627190903422906](http://doi.org/10.1080/10627190903422906)
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.
- NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States*. Washington, D.C: The National Academies Press.

- Osborne, J., Henderson, J. B., MacPherson, A., & Szu, E. (2013a, May). *Building a learning progression for argumentation in science education*. Presentation at the American Educational Research Association (AERA) conference, San Francisco.
- Osborne, J., Henderson, J. B., Szu, E., MacPherson, A., & Yao, S.-Y. (2013b, May). *Validating and assessing a new progress map for student argumentation in science*. Presentation at the American Educational Research Association (AERA) conference, San Francisco.
- Stevens, R. A., Butler, F. A., & Castellon-Wellington, M. (2001). *Academic language and content assessment: Measuring the progress of English language learners (ELLs)* (Technical No. 552). Los Angeles, CA: CRESST/ University of California, Los Angeles.
- Torres Iribarra, D. & Freund, R. (2016). WrightMap: IRT item-person map with ConQuest integration. Available at <http://github.com/david-ti/wrightmap>
- Toulmin, S. (1958). *The uses of argument*. Cambridge: Cambridge University Press.
- Wilson, M., Black, P., & Morell, L. (2013, May). *A learning progression approach to understanding students? Conceptions of the structure of matter*. Presentation at the American Educational Research Association (AERA) conference, San Francisco.
- Wolf, M. K. & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educational Assessment*, 14(3), 139–159. doi:10.1080/10627190903425883
- Wu, M., Adams, R., Wilson, M., & Haldane, S. A. (2007). *ConQuest (Version 2.0) Manual*. Hawthorn, Australia: ACER.
- Xie, Y., & Wilson, M. (2008). Investigating DIF and extensions using an LLTM approach and also an individual differences approach: An international testing context. *Psychology Science Quarterly*, 50(3), 403–416.
- Yao, S.-Y. (2013). *Investigating the validity of a scientific argumentation assessment using psychometric methods* (Unpublished doctoral dissertation). University of California, Berkeley.
- Yao, S.-Y., Wilson, M., Henderson, J. B., & Osborne, J. (2015). Investigating the function of content and argumentation items in a science test: A multidimensional approach. *Journal of Applied Measurement*, 16(2), 171-192.