Making Sense of Children's Performance on Achievement Tests: The Case of the 5th Grade Science MCAS[1]

Members of the Children and Science Tests Research Team*

Symposium Paper presented at the Annual Meeting of the American Educational Research Association. New York, NY, March 24-28, 2008.

*Members of the Children and Science Tests Research Team are, in alphabetical order: Josiane Hudicourt-Barnes, TERC; Tracy Noble, TERC; Mary Catherine O'Connor, Boston University; Ann S. Rosebery, TERC; Catherine Suarez, TERC; Beth Warren, TERC; Christopher G. Wright, Tufts University. We are grateful to Mike Russell, Boston College and Raquel Magidin de Kramer, Boston College, for conducting the statistical analyses reported herein.

[1] This research is funded by the National Science Foundation, Grant No. REC-0440180. Opinions expressed are those of the authors and not necessarily those of the Foundation.

1. Introduction and overview

Every high-stakes, state level achievement test reports discrepancies—a gap—among the scores of students from various "demographic" groups: students who qualify for free or reduced lunch and those who do not; students who speak English as a first language at home and those who do not; students who identify as White and those who do not. This discrepancy is often called the 'achievement gap', and a number of explanations have been put forward to account for it. Chief among them is the unequal access to resources that exists between low and middle-income school districts and the impact this inequality has on the quality of the teaching and learning that can take place. Within this explanatory framework, it is assumed that if a child in a low-income district answers a science question on the state test incorrectly, it is because the content was not taught, or not taught well.

Ideally, performance on high-stakes tests should be a function of how well students know the material and not a function of other, unrelated factors (Messick, 1989). However, test item creation is notoriously difficult, and approaches to test item development are not as systematic or comprehensive as the field might wish (Frederiksen, 1990, Ruiz-Primo, February, 2002, Shavelson, Carey, & Webb, 1990). Over the years, a variety of factors have been explored that may contribute to differential performance on high-stakes tests. These factors have to do with the nature of the test items themselves (the tasks they embody, their linguistic content and formulation), expectations about how students from particular subgroups will and will not perform, and the high-stakes consequences associated with performance (Abedi et al., 2000/2005; Abedi, Lord, Boscardin, & Miyoshi, 2001; Abedi, Lord, Hofstetter, & Baker, 2001; Butler & Stevens, 1997; Durán, 1989; Freedle, 2003; Hill & Larsen, 2000; O'Connor, 2006; G. Solano-Flores & E. Trumbull, 2003; Solano-Flores, Trumbull, & Kwon, 2003; Steele, 1997; Steele & Aronson, 1995). One important hypothesis is that items can contain sources of difficulty that confound language ability and background knowledge with academic proficiency in a given content area (Abedi et al., 2001; Durán, 1989; O'Connor, 2006; Solano-Flores & Trumbull, 2003; Solano-Flores, 2006).

Unfortunately, the problem has not proved to be as simple as separating factors that are legitimately related to science content knowledge from factors that, while not targeted for measurement, nonetheless affect test scores. There are many factors that may add to item difficulty – perhaps differentially so for low-income students or English language learners – but which, it may be argued, truly tap into science content knowledge. Thus, claims about "unintended item difficulty"—factors that essentially constitute a form of measurement error—must clearly and carefully establish that the factors in question are not related to science knowledge.

Moreover, although students' performance on standardized tests is thought to reveal differences in what they know and can do, it is well recognized that test scores do not tell us *why* students answer as they do (Hamilton, Nussbaum, & Snow, 1997; Messick, 1989; O'Connor, 2006). Little is actually known about the kinds of knowledge, reasoning, and interpretive practices that students use in responding to standardized achievement test items.

In this paper, we report on some factors that we hypothesize to be possible causes of unintended item difficulty for low-income students and English language learners; we share some insights into why students choose the answers they do; and we attempt to characterize these preliminary findings in ways that may lead to some concrete suggestions that test-makers and educators can use in evaluating the meaning of students' test performance.

Our goal in doing this research is to contribute to a science of assessment that is equitable for all students (Pellegrino & Chudowsky, 2003; Pellegrino, Chudowsky, & Glazer, 2001). We do not mean to assert or imply that test score differences never correspond to actual achievement differences. Traditional explanations for differential performance are well documented. Instead, we are looking for systematic sources of unintended and perhaps preventable error in measuring the performance of students from communities placed at risk in school, that is, students living in poverty, students of color, and students who speak a first language other than English at home. One of our working hypotheses, based on findings of Abedi (2002) and others, is that by reducing unintended sources of difficulty for these students, the performance of all students will improve. Thus, our hope is to better understand the specific challenges that such test items present.

The central focus of our research is the multiple choice items on the 5th grade Massachusetts Comprehensive Assessment System (MCAS). Since 2003, the MCAS has included a science test[2] that is administered to all 5th graders. This test assesses the science content specified in the Massachusetts Science and Technology Frameworks (Massachusetts Department of Education, May 2001) for Grades 3 through 5. Like other state tests, the MCAS is intended to ensure accountability for standards-based instruction and Annual Yearly Progress towards meeting the requirements of NCLB. Unlike many state tests, it is regarded nationally as a model for what such tests should be (Gehring, 2001).

Each year, when results are reported, persistent gaps in achievement are found between White students and students of color, between students who qualify for free or reduced lunch (Y-F/RL) and those who do not (N-F/RL), and students who are monolingual speakers of English and those who are limited English proficient (LEP).[3] Table 1.1, for

[2] The 5th grade MCAS science test contains 39 items, 34 of which are multiple choice. It is administered across two days, and is untimed.

[3] For purposes of comparability, we have adopted the subgroup categories used by the MA DOE when working with their dataset. For our interview and modification studies, however, this was not always possible. For example, because we were able to ask participating schools to identify students who did and did not participate in free/reduced lunch programs it was relatively easy to use this designation. However, identifying students' language status was more difficult. We interviewed many students who are not classified as "limited English proficient" (LEP) by the DOE but who the school identifies or who self-identify as speaking a language other than English at home, and many of these students were clearly English language learners. As a result, in our interview and modification studies, we use the category "students who speak a second language at home" (SASL) rather than "LEP."

example, reports student performance in 2005 on the 5th grade Science MCAS by ethnicity and clearly shows a performance "gap" among students of different ethnicities (MA Department of Education, http://profiles.doe.mass.edu/home).

	White	African-American	Hispanic/Latino
Advanced/Proficient	59%	22%	19%
Needs Improvement/Failing	42%	78%	81%

Table 1.1 Results by Race/Ethnicity for 2005 Grade 5 MCAS Science Test

The rest of this paper is organized into four sections. In Section 2, we describe our text analyses of items on the 5th grade science MCAS which resulted in the identification of seven features that we hypothesize to be unintended sources of difficulty. To determine whether these features make statistically significant contributions to item difficulty, we ran correlation and regression analyses on a dataset of student performances of more than 225,000 students in MA made public by the DOE. We report on which factors predict students' performance. In Section 3, we describe the results of clinical interviews with 18 fifth graders; here we report our initial characterization of *why* students performed as they did on two MCAS items. In Section 4, we report on a pilot study – which *preceded by a year* both the text analyses and interview studies – in which we used some preliminary hunches about sources of item difficulty to develop and experiment with a modified set of items. The results of our modifications were both surprising and informative. Finally, in Section 5, we summarize some preliminary implications of this work.

2. Investigating Item-Based Difficulty: Text Analyses and Statistical Studies

2.1 MCAS Content

The 5th Grade Science MCAS is keyed to the MA Science and Technology/Engineering Frameworks (MA DOE, May 2001) and covers learning standards for Grades 3 to 5 in four content areas: (1) Earth and Space Science; (2) Life Science (Biology); (3) Physical Sciences (Chemistry and Physics); and (4) Technology and Engineering. Each of these content areas covers a number of specific topics and standards. Table 2.1 shows the number of standards associated with each of the four content areas, Grades 3 to 5. Each year, a subset of these is selected for testing.

Content Area	# of Standards
Earth & Space Science	15
Life Science	11
Physical Science	12
Technology & Engineering	7
Total # of Standards	45

Table 2.1. Number of standards associated with each of four content areas in MA Frameworks, Grades 3 to 5.

Our analyses revealed that all standards are not equal. First, standards differ in terms of the breadth of the science content they represent. For example, Standard 9 in Physical Science requires students to "recognize that magnets have poles that repel and attract each other," while Standard 3 in Physical Science requires students to "describe how water can be changed from one state to another by adding or taking away heat." The conceptual scope and depth covered by Standard 9 is arguably narrower than that of Standard 3.

Second, standards differ in terms of the likelihood that they will be taught. According to an informal survey of teachers in the Greater Boston area, Standard 9 in Physical Science (magnets) is routinely and uniformly taught by 5th grade while many topics in Technology and Engineering, which comprise 25% of the multiple choice items on the MCAS, are not. Technology and Engineering are often not taught due to lack of curricula and time.

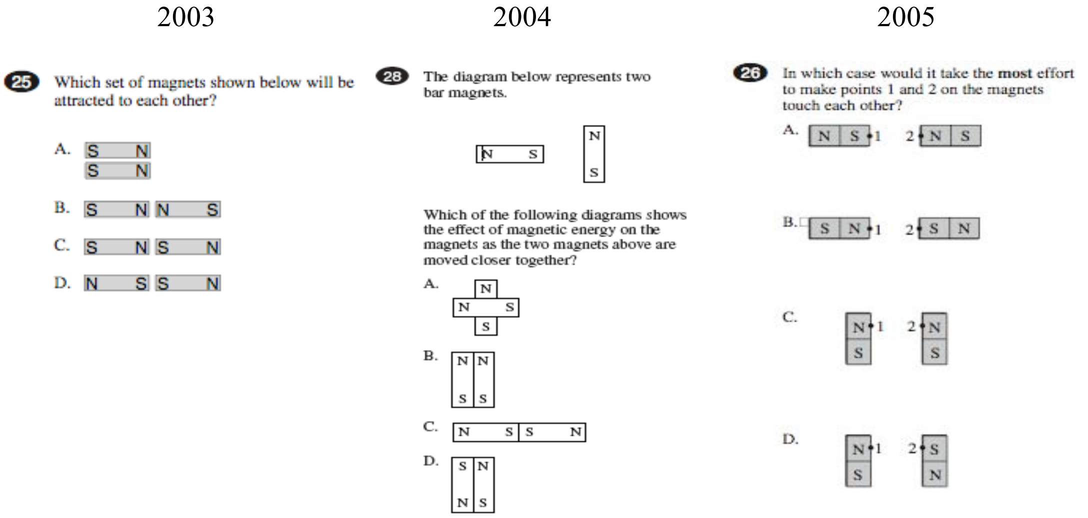
As we became increasingly familiar with the 5th Grade Science MCAS test, we found ourselves asking, do the items measure only students' knowledge of the targeted science content? Perhaps naively, we assumed that items testing the same content standard would be relatively constant in difficulty, modulo access to the science content through classroom instruction. That is, if the main determinant of difficulty is science content, then difficulty--as measured by the percentage of students who answer an item correctly--should vary across years and items tied to the same standard only if the degree to which that content is explicitly taught changes. As we discovered, however, that is not always the case.

2.2 Beyond Content Difficulty: Preliminary Explorations

Our preliminary explorations of items suggested that this idealization – that the percentage of students correctly answering an item is a function only of the science content difficulty and the extent to which that content is taught – may not correspond to reality. To explain this, we will discuss three items that appeared on different years of the MCAS, each of which was written to test Standard 9 in Physical Science (i.e., students will "recognize that magnets have poles that repel and attract each other"). Our attention was initially drawn to these items because, at first blush, they seem to have much in common. However, as we examined them side-by-side, we found they raised many questions. (The items appear in Figure 2.1 below.)

Before sharing our questions, we provide some context for understanding the items. First, it is important to note that each Fall the MA DOE releases both test results from and items administered during the previous Spring. As a result, the items keyed to any given standard must change on a yearly basis so that students do not simply "learn to the test." It is this practice on the part of the MA DOE, which is laudable, that at least in part results in pools of items keyed to the same standard, like the three magnet items below. Second, we believe that by the time they reach 5th grade, most students in MA are familiar with the behavior of magnets. We base our assumption on the observation that magnets are widely available to children experientially, even at early ages, and on the results of an informal survey of teachers that indicated that Standard 9 is taught widely and explicitly by 5th grade.

With this in mind, we turn now to the three magnet items shown in Figure 2.1, each of which is keyed to Standard 9, and appeared on the MCAS between 2003 – 2005.



Which set of magnets shown below will be attracted to each other?

The diagram below represents two bar magnets.

Which of the following diagrams shows the effect of magnetic energy on the magnets as the two magnets above are moved closer together?

In which case would it take the **most** effort to make points 1 and 2 on the magnets touch each other?

Figure 2.1 Three items keyed to Standard 9, Physical Science.

As we looked at these items side-by-side, some differences jumped out at us. We noticed, for example, they differ in length of the stem. The stem of the first item (2003, 25) has 12 words, while the stem of the second (2004, 28) has 30, and the stem of the third (2005, 26) has 21. Research by Abedi et al. (2001) and others has shown that a simple measure of stem length correlates with item difficulty. The items also differ in use of technical terminology. The second item (2004, 28) asks students to identify "which of the following diagrams shows the *effect of magnetic energy* on magnets," while neither of the other items contains technical terms of this nature. Finally, the items also differed in task. The first item (2003, 25) asks students to identify which set of magnets will be attracted to each another, while the third item (2005, 26) asks students to identify "in which case" it would "take the most effort to make points 1 and 2 on the magnets touch each other". Thus, the first item asks students to think about attracting and the third item asks students to think about repelling.

These differences drove us to look at the performance data from the MA DOE for each of these items. The percent of students, for all students and for various subgroups of students, who got each item correct is given in Figure 2.2 below. It is interesting to note that while each of the items poses a question that targets the narrow content associated with Standard 9, the percentage of students getting a given item correct varies dramatically across items.[4]

[4] As a point of comparison, the average proportion of students answering an item correctly on the 54 MCAS items we used in our statistical studies (reported in Sections 2.4.1 and 2.4.2) was .754, with a

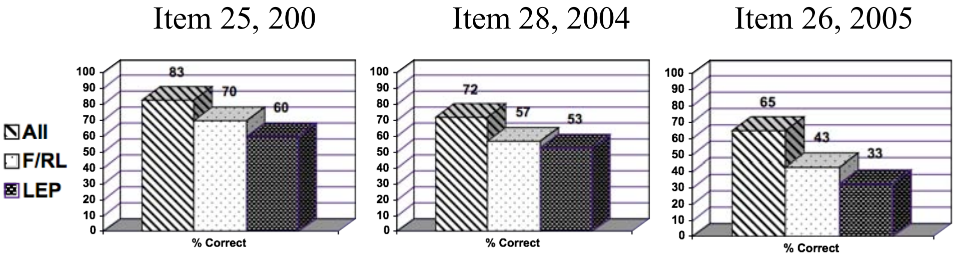

Item 25, 200 Item 28, 2004 Item 26, 2005

Figure 2.2 Percent correct by subgroup for each of the three “magnet” items

As Figure 2.2 shows, 83% of all students got Item 25, 2003, correct; 72% of all students got Item 28, 2004, correct, and 65% of all students got Item 26, 2005, correct. Note that the students on free/reduced lunch (F-R/L) and limited English proficient students (LEP) show a similar, if more dramatic, decrease in performance across items to that of all students. Of F-RL students, 70%, 43%, and 57% got the items correct, while 60%, 53% and 33% of LEP students got the items correct.

This finding, and the results of other exploratory analyses, motivated us to conduct in-depth text analyses of the items in our MCAS corpus. In the next section, we describe seven features that emerged as possible sources of unintended difficulty as a result of our analyses.

2.3 Text and Semiotic Analyses: Seven Item-Based Features

As part of our research, we conducted linguistic and semiotic analyses of the 102 multiple choice items that appeared between 2003-2005. We considered a variety of factors in our analyses. These included linguistic factors (e.g., stem length, length of answer choices, syntactic complexity), conceptual factors (e.g., familiarity of context, type of task students were being asked to complete), and factors related to visual representations (e.g., presence of image, its function in an item). Seven factors emerged as possible sources of unintended difficulty. Note that in our discussion of these seven features, we are remaining open as to whether they might represent *unintended sources of difficulty*, i.e., difficulty not related to the measurement target of science content, *or legitimate difficulty*, i.e., difficulty related to the contexts of school-based teaching and learning of science. We describe each feature below.

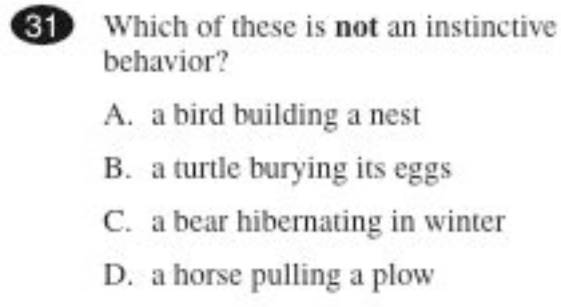
1. Mean number of semantic units per sentence in stem. Some researchers have shown that factors like sentence length, word frequency, and word length correlate with item difficulty, and that the effect is particularly pronounced for English language learners (Abedi, Hofstetter, Baker & Lord,2001; Butler & Stevens,1997). Some researchers see this as an effect of reading difficulty, others as a reflection of students’ familiarity with academic language (Bailey, 2005). Haladyna (2004) advises that item stems be kept as brief as possible. Following this, we hypothesized that the mean number of semantic

standard deviation of .145. Therefore, the difference in the proportion of students passing the first and the third of these three magnet items is approximately 1.25 standard deviations.

units (where semantic units include words, images, and symbols) per sentence in an item's stem would be related to student performance, with bigger means being correlated with lower percent correct scores.

2. Number of semantic units in answer choices. Like length of stem, the length of answer choices is also thought to be correlated with difficulty (Haladyna, 2004). Again, working off previous research, we hypothesized that students would perform less well when there were more semantic units in the answer choices.

3. Negations in stem. Haladyna (2004) advises test-makers to avoid the use of negative words in the stem because it is thought to add to item difficulty. Our text analyses revealed the presence of negations in several items, thus, we investigated this as a possible unintended source of difficulty. For the purposes of analysis, we identified negations by the presence of terms like "not" and "least likely" in the stem. Item 31 from 2003 is an example of an item containing negation. (See Figure 2.3 below.) Instead of asking students to identify the behavior that is instinctive, this item asks students to identify the behavior that is *not* instinctive. Negation in other items in our MCAS corpus function in similar ways.



31 Which of these is **not** an instinctive behavior?

A. a bird building a nest

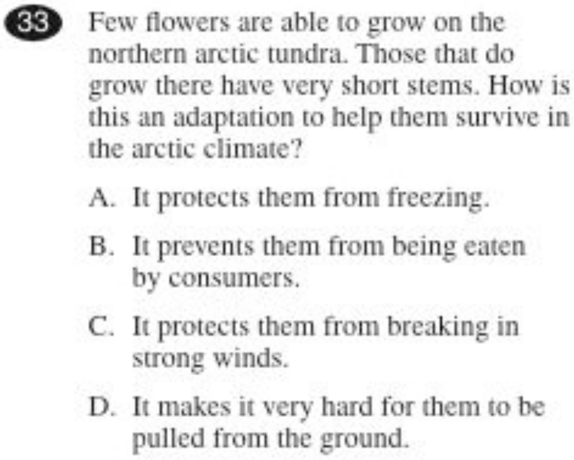
B. a turtle burying its eggs

C. a bear hibernating in winter

D. a horse pulling a plow

Figure 2.3. Item 31, 2003

4. Unfamiliar context. There is a small but growing body of research that suggests that a student's familiarity with the context used in an item may affect performance, and that performance may be differentially affected for students from backgrounds of poverty and for students who are learning English (Solano-Flores & Nelson-Barber, 2001; Solano-Flores & Trumbull, 2003). As we examined items from the 5th grade MCAS, we wondered how familiar the context of each would be to 5th graders. Item 33 from 2004, for example, is a Life Science item that asks students to identify an adaptation that helps flowers in the northern arctic tundra survive in the arctic climate. (See Figure 2.4 below.)



33 Few flowers are able to grow on the northern arctic tundra. Those that do grow there have very short stems. How is this an adaptation to help them survive in the arctic climate?

A. It protects them from freezing.

B. It prevents them from being eaten by consumers.

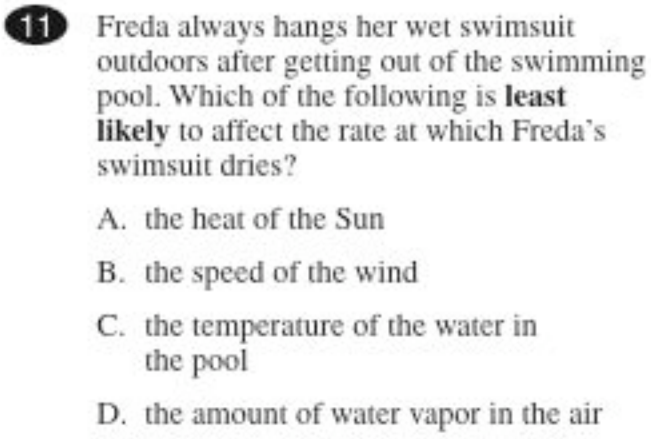
C. It protects them from breaking in strong winds.

D. It makes it very hard for them to be pulled from the ground.

Figure 2.4. Item 33, 2004

Will this item be harder for students who are unfamiliar with its context, that is, who have not learned about adaptations in the context of how plants have adapted to survive in the northern arctic tundra? Our analyses suggested that the context of some items in the MCAS corpus might be unfamiliar to 5th graders; thus, we hypothesized that unfamiliar context might function as an unintended source of difficulty.

5. Atypical perspective-taking. Our initial analyses suggested that some MCAS items require students to think about a scientific process or system from an atypical perspective. For example, the third magnet item (26, 2005) in Figure 2.1, asks students to think about which set of magnets requires the most effort to make points 1 and 2 on the magnets *touch*. It struck us that asking students to think about which set of magnets would take the most effort to push together requires adopting an atypical perspective on magnetic attraction; it is far more typical for students to think about which set of magnets will be attracted to one another. While taking an atypical perspective may not seem problematic in this item, consider Item 11 from 2003, shown in Figure 2.5 below, which we refer to as "Freda's swimsuit."



11 Freda always hangs her wet swimsuit outdoors after getting out of the swimming pool. Which of the following is **least likely** to affect the rate at which Freda's swimsuit dries?

A. the heat of the Sun

B. the speed of the wind

C. the temperature of the water in the pool

D. the amount of water vapor in the air

Figure 2.5. Item 11, 2003, Freda's Swimsuit

In this item, students are asked to identify which of four factors is *least likely* to affect the rate at which Freda's swimsuit will dry. It is typical in everyday life to think about factors that will help a swimsuit or other material dry quickly; it is atypical to think about what will *not* help a swimsuit dry. This item, and others in our corpus like it, led us to wonder whether items that require students to take an atypical perspective are harder than items that do not? Thus, we hypothesized that atypical perspective-taking might be an unintended source of difficulty.

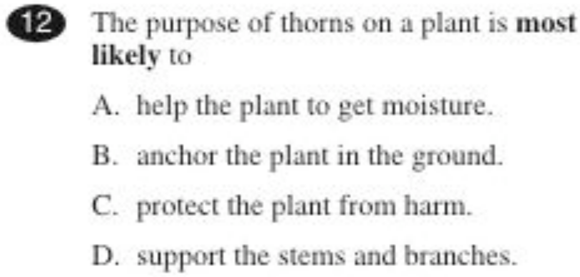
6. Forced comparison. Our text and semiotic analyses also revealed that several items take a form that requires students to go through a series of steps in order to identify the correct answer choice. We call this form "forced comparison" because it forces students to set up a scale among the answer choices and make an exhaustive comparison of the answer choices in order to identify the one that best satisfies an end-of-scale value. These items are characterized by the presence of words like *least, most,* and *best*.

Let's consider an example. "Freda's Swimsuit," shown above in Figure 2.5, is an example of forced comparison. The item asks: "Which of the following is *least likely* to affect the rate at which Freda's swimsuit dries?" To answer this, a student must recognize that the task before her is to construct a scale among the answer choices, in this case the scale is something like, "most likely to least likely to dry Freda's swimsuit." Then she must array the four answer choices along the scale in their correct order. Finally she must identify the answer choice that fulfills the "least likely" end-of-scale value. That is, she must chose the factor from among the four answer choices that will have the *least* chance of affecting the rate at which Freda's swimsuit will dry. Setting up a scale for "Freda" is challenging because all of the variables named in the answer choices (the heat of the sun, the speed of the wind, the temperature of the water in the pool, the amount of water vapor in the air) can in fact influence the rate at which a swimsuit will dry. Thus, the student cannot be sure she has found the answer until she has read and comparatively considered *all* of the answer choices. Our text analyses suggested that other items in our corpus also took this form. Some examples include questions like:

> "What can he most easily change in his design to give him more storage space?"
>
> "Which of these would be most practical to use as a lever?"
>
> "Which of the following is the **least likely** to change from a solid to a liquid state when heat is applied?"

Based on this analysis, we hypothesized that the presence of the forced comparison structure may be an unintended source of difficulty, that is, a source of difficulty unrelated to science teaching and learning.

7. Multiple Frames. As we examined items, we noticed that some items invite students to construct a single frame of meaning while other items invite students to construct multiple frames of meaning. For example, Item 12, 2005 as shown in Figure 2.6 below, invites students to construct a single frame of meaning. Students are asked to identify a scientific relationship (the purpose) that exists between a plant and one of its parts (a thorn). This item is keyed to Standard 2 in Life Science in the MA Frameworks, "identify the structures in plants (leaves, roots, flowers, stem, bark, wood) that are responsible for food production, support, water transport, reproduction, growth and protection."



12 The purpose of thorns on a plant is **most likely** to

A. help the plant to get moisture.

B. anchor the plant in the ground.

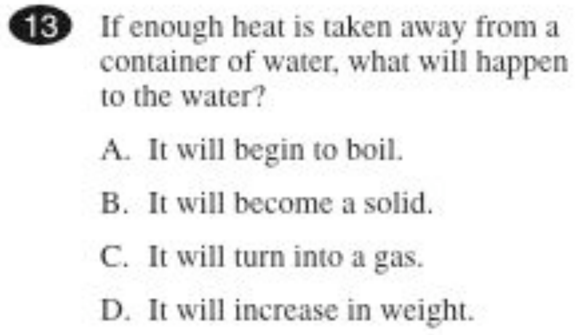
C. protect the plant from harm.

D. support the stems and branches.

Figure 2.6. Item 12, 2005, "Purpose of thorns."

Our analysis suggests that items like this, which explicitly test students' knowledge of a scientific relationship, property, or term are common on the MCAS.

Another type of frame invites students to reason about scientific *processes or models* and predict an outcome. These items often set up hypothetical scenarios about processes or change over time and ask students to run a mental model in order to predict an outcome. Often, they ask students to imagine a scenario by introducing a hypothetical: "if X happens, then what will happen to Y?" An example is given below in Figure 2.7.



13 If enough heat is taken away from a container of water, what will happen to the water?

A. It will begin to boil.

B. It will become a solid.

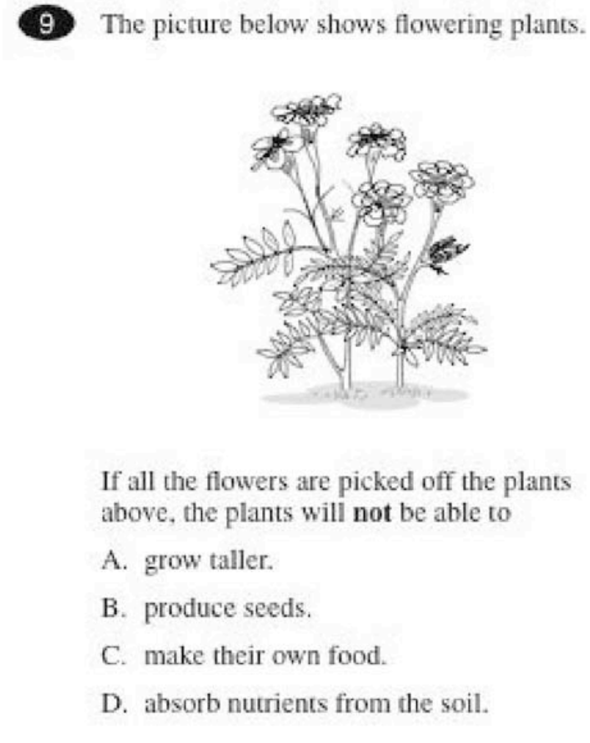
C. It will turn into a gas.

D. It will increase in weight.

Figure 2.7. Item 13, 2004, “Take Away Heat.”

Item 13 invites students to construct a single frame (albeit a different kind of frame than Item 12) which asks about a scientific relationship (the purpose of thorns on a plant). The stem of Item 13 invites students to imagine the process of taking away heat in a container of water in order to predict that the water will become a solid.

It is less important to our current purposes that we identified different types of frames than that our text analyses revealed that some MCAS items invite students to construct single frames while others invite students to construct *multiple* frames. We found, for example, that some items invite students to construct a frame about a scientific relationship *and* to reason about a scientific process, as Item 9, 2003, “Flowering Plants” does (see Figure 2.8 below).



9 The picture below shows flowering plants.

If all the flowers are picked off the plants above, the plants will **not** be able to

A. grow taller.

B. produce seeds.

C. make their own food.

D. absorb nutrients from the soil.

Figure 2.8. Item 7, 2003, “Flowering Plants.”

Like Item 12 above ("Purpose of Thorns," Figure 2.6), "Flowering Plants" is keyed to Life Science, Standard 2. It aims to assess whether students know that flowers produce seeds. The first sentence of the item ("The picture below shows flowering plants.") and the image invite students to think about characteristics of flowering plants, and in particular about the characteristics of the flowering plants shown in the image. Thus, students are initially led to expect that the item will ask them about a relationship, property or term associated with the flowers depicted in the image. The second sentence ("If all the flowers are picked off the plants above, the plants will **not** be able to"), however, invites students to reason about a scientific process, the life cycle of plants; to imagine a hypothetical scenario in which the plants are acted upon ("if all the flowers are picked off the plants"); and to predict an outcome (i.e., what will not happen after the flowers are picked off the plant?). Thus, this item and others like it invite students to construct multiple rather than single frames of understanding.

Based on related work on science tests (Hamilton, et al., 1997), reading comprehension tests (Hill & Larsen, 2000) and work in the field of visual communication (Kress & van Leeuwen, 1996), we began to wonder if such items might present an unintended source of difficulty. Do items like this invite students to engage in what Hill and Larsen (2000) refer to as "frame shift"? That is, students are invited to construct one frame and then a second, and in order to choose the correct answer they must decide which of the two frames the test-maker had in mind. As a result of this analysis, we hypothesized that items that contain multiple frames may be more difficult than those that contain a single frame.

2.4 Predicting Item Difficulty: Some Statistical Results

Once we had identified a set of features that we hypothesized to be possible sources of unintended difficulty, we wanted to know whether they correlated with actual student performance. To investigate this, we grouped all 102 items by standard, and chose the standard-based pools that contained three or more items. We did this because we wanted to look at items keyed to the same standard. We excluded items associated with Engineering and Technology standards, because our survey indicated that they are not consistently taught. This resulted in a subset of 54 items keyed to 14 standards in Earth and Space, Physical, and Life Science.

Each item was coded for the seven features described above. Features 1, 2, 3, 6 and 7 were coded by research staff; Features 4 and 5 (unfamiliar context and atypical perspective taking) were coded by outside coders. Items were coded blind with respect to student performance (percent correct) as calculated from the MA DOE dataset. An item was given a +1 for each feature that was judged to be present[5]. Theoretically, an item could have a score as low as 0 (if it contained none of the features) or as high as 7 (if it

[5] For Feature 1, we calculated the mean number of semantic units per sentence in the stem for each item and for Feature 2, we calculated the mean number of semantic units in the answer choices for each item. An item received a +1 if the mean number of semantic units was greater than the mean for all 54 items + .5 standard deviation. For Features 3-7, an item was given a +1 for each feature that was judged to be present.

contained all of the features). In reality, 32 of 54 items had a score of less than or equal to 1, and 1 item had a score greater than 5. The mean score across 54 items was 1.33.

To explore whether our proposed dimensions of difficulty did in fact predict student performance, we ran a series of correlations and regressions, using the percentage of students in the state who answered the item correctly (sometimes referred to by others as "difficulty level" but here called "percent correct") as our dependent variable. These percentages were calculated on data made publicly available by the MA Department of Education. Approximately 75,000 fifth graders are tested annually in Massachusetts, for a total of 225,000 students between 2003 - 2005.

2.4.1 Correlational Results

Linguistic Load: Features 1, 2 and 3. The first three features are associated with what other researchers have identified as linguistic load (Abedi et al, 2001; Bailey, 2005; Butler & Stevens, 1997; Haladyna, 2004). Previous research suggests that Features 1 (*mean number of semantic units per sentence in stem*), 2 (*mean number of semantic units total in answer choices)*, and 3 (*Negatives in Stem*), should all predict performance on an item, particularly for LEP students.

A correlation analysis for <u>all</u> students who took the MCAS showed that only one of these three linguistic factors correlated significantly with the proportion of students answering the item correctly. This was Feature 2, the mean number of words in the answer choices (r= -.38, p<.01). The relationship was negative, indicating that fewer students answered correctly as the number of words in the answer choices increased. Number of words in the <u>stem</u>, however, did not show any significant correlation with percent correct for all students (r= -.15, ns). Neither did the presence of a negative in the stem (r= -.127, ns).

When we broke out correlations for demographic subgroups, we found that for LEP students only, the presence of a negative in the stem correlated significantly with percent correct: r= -.272, p< .05. That is, fewer LEP students answered an item correctly if it contained a negation.

Extended Scientific Knowledge: Features 4 and 5. Features 4, *Unfamiliar Context*, and 5, *Atypical Perspective-taking,* are arguably related to depth or extensiveness of scientific knowledge. Both of these factors turn out to be strongly correlated with percent correct response. The degree to which outside raters judged the item to contain an unfamiliar context (Feature 4) was negatively correlated with percent correct response: r= -.34, p < .05). That is, if an item contained context that was judged to be unfamiliar to most 5th graders, fewer students answered it correctly.

Similar results held for Feature 5, *Atypical Perspective-taking*. The degree to which outside raters judged the item to contain an atypical perspective (Feature 5) was even more strongly associated with a drop in the percentage of correct responses (r= -.41, p< .01). This means that if an item was judged as demanding an atypical perspective, fewer students answered it correctly. Together, Features 4 and 5 explain over one quarter of the variance in performance for all students in the MA DOE dataset.

Feature 6. Feature 6, *Forced Comparison*, did not display a significant correlation with percent correct responses for all students (r= -.17, ns). However, when considering correlations with subgroups, we found that there was a significant correlation of Forced Comparison with percent correct responses for LEP students (r= -.272, p< .05).

Feature 7. Feature 7, *Multiple Frames*, did not show a significant correlation with percent correct responses in the sample overall, (r= -.19, ns), nor did it display a significant correlation for any subgroup.

2.4.2 Results of Regression Modeling

Based on these correlation results, we constructed several factors that could be studied in a regression analysis. First, we composed a "Linguistic Load" factor, made up of the scores for Features 1, 2, and 3 (*Mean number of semantic units in stem*, and *Mean number of semantic units in answer choices*, and *Negation in stem*).

Next, we combined scores for Features 4 and 5 (*Unfamiliar Context* and *Atypical Perspective-Taking*), each of which asks students to answer questions in ways that go beyond what they might have encountered in a standard classroom lesson on the science topic in the question. We called this composite factor "Extended Science Knowledge."

When both factors were introduced into the regression model for all students using the MA DOE data set, Extended Science Knowledge was a significant contributor to item difficulty (standardized Beta = -.492, p<.001). Linguistic Load was a marginally significant factor (standardized Beta = -.202, p< .093). The R squared for the model with these two factors was .38: over a third of the variance in the model is explained by these two composite factors.

When we considered LEP students as a separate demographic group, however (N=10,942), Linguistic Load is significant in the model, and its contribution is larger (standardized Beta = -.320, p= .007). The Extended Science Knowledge factor was also significant, though not as influential as it was in the model for all students (standardized Beta = -.366, p=.002). The R squared for the LEP student regression model with these two factors is .407: over 40% of the variance in the model was explained by these two composite factors. (Results of these two factors for the complement set, English-Only students, were essentially the same as those listed above for all students.)

Do these two factors play a role in performance for students designated by the MA DOE as qualifying for Free Lunch (Y-F/RL, N=69,907)?[6] A regression model for students qualifying for free lunch showed results midway between those of LEP students and all students (Linguistic Load: Stan.Beta= -.259, p< .03; Extended Scientific Knowledge:

[6] It is important to note that there is overlap between the students in this subgroup and those in the LEP and English Only data sets.

Stan.Beta= -.460, p< .001). The R squared for the FL student regression model with these two factors was .344: over one third of the variance in the model was explained by these two factors.

Although our feature of Forced Comparison did display a significant correlation with percent correct response for LEP students, it did not emerge as significant in any regression model for any subgroup.

2.5 Discussion: Understanding Our Results and Their Limitations

Our correlational and regression studies show that we have identified several features and composite factors that appear to make robust contributions to student performance on 5th Grade MCAS science items. In this section, we will attempt to further explicate the nature of those contributions. We will also consider the larger question, crucial for policy and test-construction, about whether these factors are contributing difficulty that is *unrelated* to science content knowledge. These same studies have also shown that several of the features we strongly suspected of having an impact on performance did *not* reach the level of statistical significance, or did not display a large effect, even when they did reach statistical significance. We will also consider the meaning of these findings.

2.5.1 Extended Science Knowledge: A Factor Predicting Difficulty?

This factor showed the largest effects of any in our investigation. As discussed in Section 2.3, we wondered whether items that contain contexts unfamiliar to most 5th graders or that require atypical perspective-taking would be harder. On the other hand, the argument can be made that students who are able to deal with atypical perspective-taking in thinking about scientific processes, and who are able to extend their knowledge of a concept to an unfamiliar context, should in fact be considered more knowledgeable in the science content that is the main measurement target of MCAS science. From this perspective, these dimensions do not add unintended or spurious difficulty to a science item.

Do they, however, privilege students from affluent backgrounds, and students (whether affluent or not) who have access to the intellectual resources available to highly educated people? Specifically, do students with access to museums, books outside of school, nature preserves, and so on, have access as well to more examples of the science concepts taught in school? Do they have more extensive access to sites and occasions on which extended thinking about a scientific process or phenomenon takes place? If so, we might expect this factor to play a less robust role in predicting difficulty for affluent students. Do our data support this position? Using Free Lunch status as proxy for access to extended science knowledge, we find no difference in the effect size of this factor for Yes-FL and No-FL students: standardized betas are -.46 vs. -.477 respectively.

But does this mean that our intuitions about differential access to extended science knowledge are wrong? We do not think so. It may be that even in the "non-free lunch" population (which includes approximately one third of our sample), the vast majority of students still are not affluent, and do not have access to this extended knowledge. What kind of statistical data would be required to explore this further? We would need to look at item response data in a few very affluent districts with very high levels of parental education, and compare these with item response data in several districts with very low levels of parental education and income. If this comparison does not reveal a substantially greater effect of our Extended Science Knowledge factor for low-income students, then one could reasonably conclude that we have identified a factor that correlates with the quality of science education in schools.

2.5.2 Linguistic Load: How Does It Increase Difficulty?

The data are clear that, particularly for LEP students, there is a noticeable effect of difficulty associated with higher than average numbers of semantic units in the stem and in the answer choices, and with the presence of negatives in the stem. The percent of students who answered an item correctly decreased as the number of semantic units in the item increased and in the presence of a negation. As mentioned above, other researchers have found that word quantity, frequency, and syntactic complexity may make items more difficult (Abedi, Lord, Hofstetter et al., 2001, Haladyna, 2004). Our Features 1 and 2 indicate whether an item has more semantic units than average in the stem or in the answer choices, respectively. This includes both words and images, and because of this, our "linguistic load" includes more than just the processing of words.

Specifically, for some of the easiest items (i.e., highest percent correct response), the answer choices were represented as images rather than words (i.e., there were 0 words in the answer choices). We hypothesize that this in turn is related to the kind of task the item poses for the student. Specifically, items that permit images as answer choices may involve simpler tasks (e.g., those of identification or categorization) than items that require words. Therefore, one aspect of the difficulty level conferred by the factor of Linguistic Load might be argued to be indirectly due to scientific content difficulty. Scientifically easier tasks (e.g. identifying from four images of bird feet "which bird's foot is best for grasping prey?") may themselves be correlated with lower linguistic load.

2.5.3 Findings of Non-Significance

It is important to note that there are limits to our analysis. Although it rests on responses from tens of thousands of students, our item codes are distributed over a subset of 54 items. The frequency of occurrence of some features in this sample is low (e.g., Feature 3, negation and Feature 6, forced comparison) while the frequency of occurrence of other features is higher (e.g., Feature 1: mean number of semantic units per sentence in stem is greater than mean + .5 standard deviation for our sample). Thus, with a sample of only 54 items, it is certainly possible that a feature might actually be a good predictor of difficulty, but not appear frequently enough in our sample to rise to a level of statistical significance in its correlation with pass rate.

Moreover, as discussed earlier, the difficulty of scientific content, the intended measurement focus of the MCAS science test, is not constant across items. Consider the content targeted by the "magnets" items (Figure 2.1, p. 5) as opposed to the content targeted by "Freda's Swimsuit" (Figure 2.5, p. 8) or "Flowering Plants" (Figure 2.8, p. 11). Completely apart from the kinds of dimensions we have discussed so far, items can differ in scientific difficulty merely because the content is more complex or less experientially accessible. Without a parallel theory of difficulty for science content, we cannot fully control for sources of *intended* difficulty in our search for sources of unintended difficulty. The dimension of intended difficulty – ubiquitous in research in science assessment – is a central challenge in the field.

By way of closing this section, let us briefly revisit Features 6 and 7, Forced Comparison and Multiple Frames, respectively. Our intuitions told us that these features should be associated with greater item difficulty and lower percentages of correct responses. Statistically, however, they were not. Does this mean that these features are truly insignificant in the process of item interpretation? We are not yet convinced that this is the case. As we will show in Sections 3 and 4, we found evidence of their effects using alternate methodologies. In Section 3, we present evidence from extended clinical interviews. In Section 4, we present evidence from a pilot study, conducted a year *before* the other analyses, in which we revised 19 items we hypothesized to be problematic and administered them to 100 students who had not seen the original version.

3. Interview Analysis: Evidence for Multiple Frames and Forced Comparison Features

3.1 Introduction

Interview-based studies of students' responses to standardized test items have shown that test scores alone cannot tell us about the complexity of the knowledge and reasoning that students use as they respond to test items (Gee et al., 1992; Hamilton et al., 1997; Haney & Scott, 1987; Hill & Larsen, 2000; Kazemi, 2002; Langer, 1987; Mehan, 1975; MacKay, 1974; O'Connor, 2006; Solano-Flores & Nelson-Barber, 2001; Solano-Flores & Trumbull, 2003). These studies have shown that students use a variety of sources of knowledge and reasoning, many of which are not anticipated by test-makers. For example, Hamilton et al. (1997) found that items that appeared to be testing straightforward factual or declarative knowledge turned out to invite more complicated, model-based reasoning.

This is particularly true of children from communities placed at risk in school. Hill and Larsen's (2000) classic study of children's interpretations of items on the Gates-MacGinitie Reading Test provided detailed examples of the ways in which children's ethnocultural knowledge and experience can conflict with the assumed norms of meaning and language used in the test, including assumptions about word meaning, spatial and temporal framing, real world knowledge, and communicative norms. Solano-Flores and Trumbull (2003) found that students of different cultural background and socioeconomic status tended to interpret seemingly simple words in NAEP items (e.g., "only" as in "His mother has only $1.00 bills.") differently than middle class students, for example, interpreting the word "only" as restricting the number of dollars ("His mother has only one dollar.").

To investigate how students interpret multiple choice items on the 5th grade Science MCAS and why they choose the answers they do, we conducted individual interviews with eighteen 5th graders. We were interested in identifying the sense-making resources (e.g., sources of knowledge, concepts, reasoning practices, interpretive assumptions) these students used in responding to items and discernible patterns in the ways they interpreted particular item features.

Students were recruited for the interview from five urban schools in the Boston area: one charter school and four public elementary schools. Demographic data for the students is provided in Table 3.1.

Total number of students interviewed	Number of students who receive free/reduced lunch	Number of students who speak a language other than English (SASL)
18	9	12

Table 3.1 Demographic Data for Interviewed Students

Students were interviewed on six items. These items represent all the items keyed to two standards, one in Life Science (Standard 2: Identify the structures in plants--leaves, roots, flowers, stem, bark, wood--that are responsible for food production, support, water transport, reproduction, growth and protection.) and one in Physical Science (Standard 3: Describe how water can be changed from one state to another by adding or taking away heat.) All of these items appeared on the MCAS between 2003 - 2005. (See Table 3.2.) These two standard-based item pools were chosen because a) the items within the pools had large ranges for MA DOE difficulty scores (what we refer to as "percent correct" in this paper), and b) items within these pools embodied features of interest based on our text analyses. The percent of students answering each item correctly on the state-wide administration of the MCAS are shown below in Table 3.2.

Item Name	**Year**	**Item #**	**% Correct**
Flowering plants (Life Science)	2003	9	61%
Plant diagram (Life Science)	2003	23	50%
Thorns (Life Science)	2005	12	86%
Freda's swimsuit (Physical Science)	2003	11	54%
Take away heat (Physical Science)	2004	13	48%
Justin's jar (Physical Science)	2004	34	87%

Table 3.2. Six items from two standards chosen for interviews

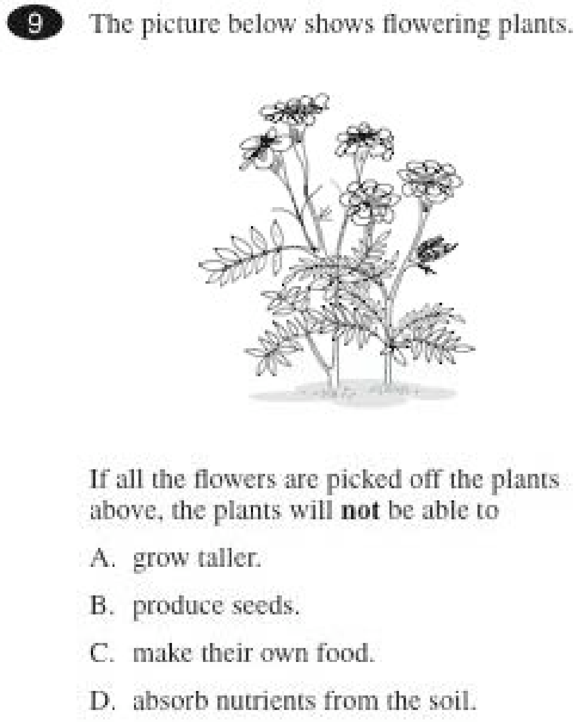
The differences in percent correct values for items designed to test knowledge of the same standard (as shown in Table 3.2) raise questions about the possible causes of these differences. Section 2 laid out several hypotheses about factors other than science content that may be influencing item difficulty. In order to further investigate possible unintended causes of item difficulty, we interviewed students and asked them to think aloud about their reasoning on these six items.

We used variations on methods described by Hill and Larsen (2000) and Solano-Flores and Trumbull (2003). First, students were told that the interviewer's interest was in understanding how (s)he thinks about the questions, and not whether (s)he answered correctly. The student then took a short, untimed test comprised of the six items. Following this, students were interviewed individually following a clinical interview protocol. The protocol asked students to describe their reasoning about each item, and to share their knowledge of particular words and science concepts in the stem and answer choices. We also asked students about the answer choices that were the most popular distractors in state-wide testing. Item protocols were tailored to individual items and their design was based upon protocols pilot-tested in a series of similar interviews in the previous year. Interviews lasted approximately one hour and were videotaped. Students' responses were transcribed and analyzed for evidence of their knowledge about the subject matter and words in each item, as well as the reasoning behind their answer choices.

In this section, we present student responses to one item from each of the standards above (Life Science, Standard 2, Item 9, 2003, "Flowering Plants," and Physical Science, Standard 3, Item 11, 2003, "Freda's Swimsuit") to describe a range of student responses in two different areas of science. Both of these items had low percent correct values on the statewide administration of the MCAS (see Table 3.2), and presented significant challenges to the students we interviewed. In addition, each item exemplifies one or more features of interest described in Section 2.

3.2 Flowering Plants Item: An Example of Multiple Frames

In this section, we analyze students' responses to Item 9, 2003, "Flowering Plants." This item was discussed in Section 2 as an example of Multiple Frames (see Figure 3.1 below.)



9 The picture below shows flowering plants.

If all the flowers are picked off the plants above, the plants will **not** be able to

A. grow taller.

B. produce seeds.

C. make their own food.

D. absorb nutrients from the soil.

Figure 3.1: Item 9, 2003, Flowering Plants

3.2.1 Student Performance Data

When this item was administered state-wide to students in Massachusetts in 2003, 61% of all 5th graders chose the correct answer, B. "produce seeds." (See Table 3.3 below.) Statewide, 44% of students receiving free/reduced lunch assistance (F/RL), and 34% of students labeled as Limited English Proficient (LEP) answered it correctly. In our interview sample of 18 students, 9 (50%) answered it correctly. Four out of 9 (44%) of F/RL students and 4 out of 12 (33%) of students who speak a second language (SASL) answered it correctly.

In the state-wide administration, answer choices A and D were the most commonly chosen distractors. Interview students did not show a preference for any of the distractors. (During the interview, however, when asked if there might be another good answer choice, 33% of students suggested that C, "make their own food," might be another possible option.) What can we learn from the interviews about why students performed so poorly on this item? Did particular features prompt some students, especially F/RL students and students who speak a second language (SASL), to choose distractors rather than the correct answer?

State Administration (approx. 65,000 students)			**Interview Students (18)**		
All Students	F/RL Students	LEP Students	All students	F/RL students	SASL students
61%	44%	34%	50% (9/18)	44% (4/9)	33% (4/12)

Table 3.3 The percent of students in MA and in our interview sample who answered "Flowering Plants" correctly.

3.2.2 Structure and Science of the Item

We asked six teachers if this topic is regularly taught in grades K-5 in their school; four answered yes, one answered no, and one said she did not know. The STC unit, *Plant Growth and Development* (NSRC, 1991) is regularly used in the schools that participated in our interview study. In this unit, a student observes a Wisconsin Fast Plant ^{TM} as it goes through each stage of development. Students explore the cyclical nature of plant development, and the plant's need for nutrients from the soil, as well as water and light. The unit is designed so that students can also explore the interdependence of living things such as bees and plants, including pollination. The unit is typically taught in grade 3 or 4. Thus, it is likely that students who have engaged with this unit have been exposed to the fact that flowers produce seeds.

3.2.3 Students' Responses

In the interviews, 83% (15/18) of the students told us that a plant's flower contains seeds and/or that the flowers are the source of seed production for a plant. Some additionally noted that "produce seeds" can also entail dispersing seeds to form new plants. However, only 50% of interview students answered this item correctly. Of the fifteen students who told us that they knew that flowers make seeds, six (44%) answered the item incorrectly.

Let us take a closer look at the item to better understand what these students might have been doing. As described in Section 2.3, "Flowering Plants" contains multiple frames; that is, the item invites students to identify the function the flower serves for the plant *and* predict what will happen when the flowers are removed from the plant. In Section 2.3, we hypothesized that these multiple frames may lead some students to imagine competing alternative outcomes and thus may represent an unintended but real source of difficulty for students.

Although statistical analyses of the features we coded for (reported in Section 2.4) did not show that the multiple frames feature was significant, data from student interviews paints a different picture. Students' responses to "Flowering Plants" show that while some interpreted the item as asking only about the purpose of flowers, others speculated about various outcomes that could result from picking the flowers off the plants. We share excerpts from students' interviews to illustrate both of these response patterns.

According to interview data, six of fifteen students (40%) who knew the science content interpreted the item as asking about the function of flowers (i.e., a single frame); of these six, five answered the item correctly (see Table 3.4).

	Single Frame	Multiple Frames	Other
Correct	5	3	1
Incorrect	1	5	0
Total	6	8	1

Table 3.4 Number of students who knew the science content answering correctly/incorrectly according to the number of frame(s) they engaged with.

Each of the students who focused on a single frame told us in one way or another that they knew that the flower produces seeds, and that without the flower the plant would no longer be able to produce seeds. George, for example, explained that after the flowers were picked “then the thing won’t be able to produce any seeds” and Kadim told us “it won’t be able to produce seeds without the top… because the top is where the seeds, like that’s where you make the seeds.” Some of the students told us explicitly that they interpreted this question as asking about the function of the flower. Carleen for example explained to us that in science she learned “what the body parts of the flower does [sic].”

Of the six students who interpreted the item as asking about a property of the plant, Rachel was the only one who answered it incorrectly. She told us that the function of a flower might be to breathe in air and nutrients and incorrectly chose answer choice D, “absorb nutrients from the soil.” However, she also said that “produce seeds” might be another good answer choice “because the seeds usually grow in the middle of the flower” and if the flowers are picked off, “it won’t be producing seeds.” None of these six students, including Rachel, speculated about what conditions might have been in place when the flowers were picked or what might have happened to the flowers once they were removed. Excerpts from the interviews with students who focused on a single frame are given in Table 3.5.

Student	**Quote**	**Answered Correctly?**
George	“And I figured out that the flowers produce seeds, so if you take off the flowers, then the thing won't really be able to produce any seeds.”	Yes
Kimberly	“I think the produce seed part, this is like what the bud part does right here, I think.	Yes
Carleen	“Because I thought if they took away that part off, then it wouldn’t produce seeds because um, inside like, inside the flower, there's little seeds in it…[We learned in Science] what the body parts of the flower does”	Yes
Ayana	“A lot of plants have seeds in the flowers…[In third grade] we learned a lot about flowers. We learned that flowers are like male and females.”	Yes
Kadim	“it won’t be able to produce seeds without the top… Because the top is where the seeds-- like that’s where you make the seeds. Because the roots make the water, and the stem makes the food. And then it goes up to the petal which makes the seeds.”	Yes

Rachel	"maybe the top part absorbs air and with the air it might breathe and absorb the soil"	No

Table 3.5 Excerpts from students who focused on a single frame

Our data suggest that eight other students who knew the science content engaged in a different kind of reasoning to answer "Flowering Plants". These students focused on both the property and process frames of the item. Of these eight students, only three (33%) answered the item correctly. (See Table 3.4 above.) These eight students seemed to think the item was asking them to focus on the process and to construct possible outcomes for the plant or the flowers once the flowers were picked off.

For instance, Emily, like many students who used multiple frames, was led to imagine hypothetical agents acting on the plants. She explained that "[they didn't] take the whole plant out and like, step on it, like step on it or something" so the flower might still be able to release seeds. As a result, she did not think that answer choice B, "produce seeds," was a correct option. Another student, Paul, appeared to struggle to bring purpose and process frames together. He thought that if a plant's growth cycle is disconnected by picking the flowers "it probably won't grow taller," but he also thought "that part [the center of the flower] produces seeds so it wouldn't produce seeds either." He carefully explained that the item presented him with many things to think about, such as which part of the plant gets picked off and whether the growth cycle will be interrupted. Paul eventually reasoned his way back to the correct answer, but not before imagining a range of alternate possible outcomes. Excerpts from the responses of students who used multiple frames when answering this item are given in Table 3.6 below.

Student	**Quote**	**Answer Correct?**
Matthew	"Well, like if you have an apple and you cut it in half, there would be seeds inside, so maybe there would be like seeds in the center of it, that when someone picked it or if it, say, was swaying really hard in the wind, then maybe the seeds would scatter, and new plants would grow. Maybe if they got planted…I suppose that, yeah, I mean, you could still grow taller, but I would think that this, that would make it shorter. So maybe that might affect something in the plant that makes it grow a little bit shorter."	No
Emily	"I was gonna pick produce seeds, I was gonna pick produce seeds because it said like take the things off, I didn't think it can grow, like it could grow back…Because um, I always thought that the seeds came from up there so depending on what flower it is, it comes from like the middle, and then the wind comes and like spreads it, when it blows, so that's what I was thinking, because I was thinking that take the whole plant out and like, step on it, like step on it or something, they just said that they would like take, pick it, the little, the flowers, and the middle would stay, so that's what I thought	No

	that the middle would stay and spread it, I thought that it would just not grow.	
Andrew	“And it would be able to produce seeds. That was the other thing I was thinking of because the flower would fall down, really, and the seeds would come out because the seeds are in the flower… The seeds are still there, they’re not destroyed”	No
Paul	“I think there's a cycle, and if the cycle disconnected it probably couldn’t grow taller. And I was thinking, I thought that that part produces seeds…I thought that whole part would be picked off so then it wouldn’t produce seeds either…It would probably die actually.”	Yes
Maria	“there’s a plant, I think it’s the sunflower. When it dies and the flower kind of goes down, the seeds fall into the earth and then it grows again, a new one grows… This might be one of those typical flowers that they die, they produce seeds…Bees can come and they can, like, leave—They can leave pollen and make the food…It might be between the B [produce seeds] and C [make their own food]”	Yes

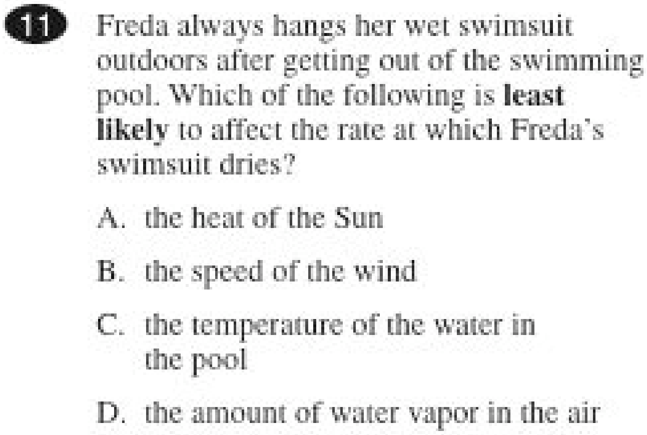
Table 3.6 Excerpts from interviews with students who used multiple frames

Students’ performance on this item (9/18, or 50% correct) stands in contrast to the fact that 15/18, or 83% of students indicated that they knew that flowers contain and/or produce seeds. Their responses show that they often chose one of the three distractors not because they didn’t know that flowers produce seeds, but because they reasoned generatively about the effect that picking off the flowers might have on other processes associated with flowers and plants (i.e., plant growth, development, reproduction). Interestingly, three students chose to change their answers during the course of their interviews (two changed from an incorrect choice to another incorrect choice, and one self-corrected), and all three of these students showed evidence that they were focusing on the multiple frames of the item.

By way of closing, we give one of our students, Andrew, the last word because we feel his response illustrates clearly what we learned about students’ reasoning about multiple frames from their interviews on this item. When asked to think about why other students thought that “produce seeds” might be the correct answer he responded, “Maybe the guy who picked it off didn’t put it down. He took it with him.” And then when describing how the item might be made better Andrew said, “Maybe put a bit more details about the guy who picks the flower” and “what happens to the flower”, because that would change the answer. He said, “if they just left it [the flower] there, the seeds would still grow” but if they took it away “there would be no seeds”, presumably meaning that any flowers left lying on the ground could release their seeds and these seeds could grow into new plants. Here Andrew is explicitly telling us that the item invited him to speculate about possible outcomes in a relatively unconstrained space, and he astutely observes that if the item provided more information this realm of possibilities would be limited and might have implications for which answer choice was correct. We believe that students’ responses to this item provide strong qualitative evidence for our hypothesis that items that contain multiple frames may challenge students in ways unintended by the test-makers and unrelated to students' knowledge of the science content.

3.3 Freda's Swimsuit item: An Example of Forced Comparison

In this section, we analyze students' responses to a second test item, which we call "Freda's Swimsuit". This item appeared in the 2003 MCAS and is keyed to Physical Science Standard 3, "Describe how water can be changed from one state to another by adding or taking away heat". It was discussed in Section 2.3 as an example of Feature 5, Atypical Perspective Taking, and Feature 6, Forced Comparison. The item is reproduced below in Figure 3.2.



11 Freda always hangs her wet swimsuit outdoors after getting out of the swimming pool. Which of the following is **least likely** to affect the rate at which Freda's swimsuit dries?

A. the heat of the Sun

B. the speed of the wind

C. the temperature of the water in the pool

D. the amount of water vapor in the air

Figure 3.2 Item 11, 2003, Freda's Swimsuit

3.3.1 Student Performance Data

When this item was administered state-wide to students in Massachusetts in 2003, 54% of all Table 3.7) Statewide, 35% of F/RL students, and 24% of LEP students answered correctly. In our interview sample of 18 students, 11 (61%) answered it correctly. Of the nine F/RL students in our sample, two (22%) answered correctly, and of the 12 students who speak a second language (SASL), three (25%) answered correctly.

In the state-wide administration, answer choices A. "the heat of the sun" and D. "the amount of water vapor in the air" were the most commonly chosen distracters. Among the students we interviewed, answer choice A, "the heat of the sun", was by far the most commonly chosen distracter. Why did students overall perform so poorly on this item? And why were these distracters so attractive to students?

State Administration (approx 65,000 students)			**Interview Students (18)**		
All Students	Y-F/RL Students	LEP Students	All students	Y-F/RL students	SASL
54%	35%	24%	61% (11/18)	22% (2/9)	25% (3/12)

Table 3.7 The percent of students in MA and in our interview sample who answered Freda's Swimsuit correctly.

3.3.2 Structure and Science of Item

We asked six elementary school teachers from the district from which we drew our interviewees if the topic of this item is regularly taught in grades K-5 in their school. Three out of six said yes, two did not know, and one was torn between yes and no. In this district, the topic of change of state of matter is taught in the context of weather and the water cycle in the 5th grade year. The students we interviewed were in the winter of their 5th grade year, and so may have not yet had this unit. In addition, schools had individual variation in the units they taught and their timing. Only seven of the 18 students we interviewed reported having learned about the topic of this item in school.

To better understand students' experiences with this test item, let's look at the item in more detail. As noted in Section 2.3, this item takes an atypical perspective on the topic of drying, because it asks: "Which of the following is **least likely** to affect the rate at which Freda's swimsuit dries?" To understand what this question is asking, the student must imagine Freda's swimsuit drying, and the various factors that can affect the rate at which it dries, i.e., how fast it dries. Then the student must take an atypical perspective, and look for what will do the *least* to dry the swimsuit. The typical perspective, suggested by the opening statement: "Freda always hangs her wet swimsuit outdoors after getting out of the swimming pool", is that the goal is to dry the swimsuit and thus a student might expect to be asked what will do the *most* to help dry the swimsuit.

Next, the student must interpret the question in more detail. "Which of the following is least likely to affect the rate at which Freda's swimsuit dries?" can have a number of possible interpretations. One interpretation that is consistent with the correct answer is the following: "On an average sunny day in the summer in Massachusetts, when all the factors are present and have typical values for such a day, which factor would do the least to help Freda's swimsuit dry?" Given this interpretation, one must determine which is the correct answer choice.

As discussed in Section 2.3, this item also has the forced comparison feature, and thus students must order the factors listed in the answer choices along a scale of likelihood of helping to dry the swimsuit and choose the one at the *least likely* end. If one interprets the question as in the previous paragraph, then one's ordering of the factors would look like the one shown in Figure 3.3.

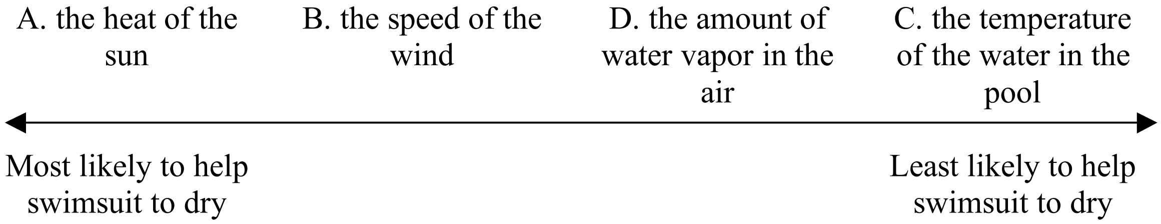

A. the heat of the sun
B. the speed of the wind
D. the amount of water vapor in the air
C. the temperature of the water in the pool
Most likely to help swimsuit to dry
Least likely to help swimsuit to dry

Figure 3.3. Scale of factors ordered from most to least likely

The heat of the sun is the largest contributor to drying on a sunny day, due to the sun's radiation warming the water in a swimsuit left out in the sun, and causing the water to evaporate more quickly. If there is a moderate wind, this can also contribute to drying, largely because the air passing over the swimsuit facilitates the evaporation of water from the swimsuit. The amount of water vapor in the air is the amount of water in its gas state that is mixed with the air we breathe. The more water vapor in the air, the harder it is for additional water to evaporate into the air. Thus the amount of water vapor in the air can help the swimsuit to dry faster by being small. The temperature of the water in the pool is likely to be the smallest contributor to the drying of the swimsuit in this scenario. The temperature of the water in the pool is the starting temperature for the water in the swimsuit, which affects how easily it will evaporate from the swimsuit. However, the water can also leave the swimsuit by dripping, and by having kinetic energy added to the water molecules by the wind rushing over the swimsuit, so the water temperature is likely to have less effect than the other factors.

Our hypothesis is that the atypical perspective and forced comparison features of this item were sources of unintended difficulty for students. Next we look at the students' responses in order to test this hypothesis.

3.3.3 Students' responses

In the interviews, all 18 students reported being familiar with drying wet swimsuits or other clothing. And 15 of 18 (83%) demonstrated that they knew the science. Eleven of 18 students (61%) ordered the factors as described earlier, saying that the heat of the sun is the most likely to affect the rate at which the swimsuit dries, and the temperature of the water is the least likely to affect it[7]. Seven of these eleven students (64%) answered the item correctly.

Table 3.8 shows the number of students who answered the item correctly and incorrectly depending on whether or not they ordered the factors according to the intended order or an alternative order. Note that all seven students who used an alternate ordering of the factors answered the item incorrectly.

[7] Six of the eleven students who arranged the factors in this order did not include "D. the amount of water vapor in the air" because they were not sure what it was.

	Intended Order of Factors	Alternative Order of Factors	Totals
Correct answer	7	0	7
Incorrect answer	4	7	11
Totals	11	7	18

Table 3.8 Number of students answering the item correctly or incorrectly depending on how they ordered the factors.

Atypical Perspective

Interestingly, four of the eleven students who constructed the intended ordering of the factors initially answered the question incorrectly. What would lead these students, who knew the science, to a wrong answer choice? Each of the four students who initially answered this question incorrectly[8] told us that (s)he initially thought the question was asking for the factor that was *most* likely to help the swimsuit to dry, rather than the one that was least likely to help it. (See Table 3.9 for excerpts of students' responses.) All four initially chose the same distracter, "A. the heat of the sun," as their answer, which is consistent with their interpretation of the item as asking for what would help the swimsuit to dry. These students reported interpreting the item from the typical (what will most help dry) rather than the atypical (what will least help dry) perspective.

Student	Reason for choosing "A. heat of the sun"	Changed answer?	Final answer (C is correct answer choice)
Kadim	Thought question asked: "Which [answer choice], like, helped [the swimsuit] dry quicker and faster".	N	A
Emily	"when it says most likely it's basically the best answer, and least likely is like what like can't happen, so the heat of the sun, I picked it, and it was wrong because it said least likely, and not most likely, and I thought it said most likely"	Y	C
Rachel	"Because I thought the question was asking which one do you think might help Freda's swimsuit dry."	Y	C
Teresa	"I thought the question was backwards".	Y	C

Table 3.9 Excerpts from students' responses to "Freda's Swimsuit"

Why did students construct this "backwards", but nonetheless typical interpretation? We need to return to the item (See Figure 3.2) to better understand this. The first sentence of the item, "Freda always hangs her wet swimsuit outdoors after getting out of the swimming pool." suggests that Freda wants to dry her swimsuit. Why else would she have hung it up? A common-sense perspective suggests that one is looking for what will

[8] Three of these four students changed their answer to the correct answer during the course of the interview.

help it to dry. Our interview data clearly show that all eighteen students understood the intended meaning of this sentence and were able to relate it to personal experiences with hanging things out to dry. The second sentence, “Which of the following is **least likely** to affect the rate at which Freda’s swimsuit dries?” contains a complex phrase (“affect the rate at which Freda’s swimsuit dries”) that employs scientific terms (“affect”, “rate”) that have multiple meanings[9]. If a student were having difficulty interpreting this phrase, (s)he might use what (s)he knows from her everyday experience along with the contextual cues contained in the first sentence to reasonably infer that the question is asking for the factor that will *help* Freda’s swimsuit dry quickly.

The typical perspective, that one should look for what will help the swimsuit to dry, is supported not only by a common-sense perspective, but also by conventions of question-asking in science classrooms and science texts. Students are asked what will cause an effect, accelerate the car, make the plant grow faster (or slower), not what will be least likely to do these things. The focus of science curricula on relating causes to effects leads students to expect the typical perspective described in preceding paragraphs, rather than the atypical perspective of this question.

In their hallmark study of the Gates-MacGinitie reading test, Hill and Larsen (2000) analyzed elementary school students’ think aloud responses to reading comprehension items. These students were led to particular distracters when the stem of an item established an initial way of understanding the topic that was later repudiated by additional text in the item. Students who used their initial understanding to interpret the remaining text tended to answer the item incorrectly. We believe a similar phenomenon may be in operation for the students we interviewed.

One student who changed her answer after realizing she had been using a “backwards” interpretation was asked to speculate about why a lot of students choose A, “the heat of the sun” (her initial answer), as their answer. Her reply: “Maybe they read the question wrong.” Our evidence suggests that many students, in fact, may be “reading the question wrong” because of the complex language structure found in this item. Another student, Teresa, ordered the factors influencing the drying of the swimsuit in the way the test-makers intended. However, her initial, and even second and third interpretations of the question led her to incorrect answers. When she first realized that she had read the question wrong, Teresa said: “I actually thought the question said what would help it so I put the heat of the sun.” And after recognizing this, Teresa tried out “D. the amount of water vapor in the air”, and “B. the speed of the wind” as answers, before she settled on the correct answer, “C. the temperature of the water in the pool”. She settled on C as an answer only after restating the item stem as asking for “which one will probably not help

[9] “Affect” and “rate” are used in many different ways in mathematics and science and everyday life, and several students expressed uncertainty about one or both of these terms or gave a non-standard definition of them when asked. Six out of 18 students expressed uncertainty about the meaning of the word “rate” or gave a definition of it that did not include reference to the speed at which the swimsuit dried. Three students expressed uncertainty about the meaning of the word “affect” or gave answers that did not include the idea of acting on or changing the rate.

it [the swimsuit to dry]". For Teresa, who appeared to be a good reader and a self-aware test-taker, this item presented multiple possibilities for interpretation.

The atypical perspective presented by the task of determining "which one will probably not help" the swimsuit to dry, as opposed to which one will help the swimsuit to dry, added unintended difficulty to this item for four of the eighteen (or 22%) students we interviewed. We believe that the incorrect answers of this group are not a reflection of their science knowledge, or even their knowledge of the words in this question, but instead are a result of unintended difficulty due to the combination of this atypical perspective with a forced comparison, making the linguistic structure of this item particularly complex and unexpected.

In the state of Massachusetts as a whole, 17% of students chose A. "the heat of the sun", instead of the correct answer, and A was one of the most popular distracters. It is possible that reasoning like that followed by the students described above may explain why some of these students chose this answer. If that is the case, then they may have been misled by the wording of the item, despite having correct science knowledge.

Alternative ordering of factors.

As noted earlier, we found that seven students did not order the factors contributing to the drying of the swimsuit in the way the testmakers intended. Instead of saying that "A. the heat of the sun" was most likely to help the swimsuit to dry and that "C. the temperature of the water in the pool" was least likely to help it to dry, they ordered the factors in another way. All the students who used an alternative ordering of the factors got this item wrong. Did they simply not know the science? Our analysis of students' responses revealed that three of these students answered incorrectly due to lack of knowledge. One guessed, one did not know what "C. the temperature of the water in the pool" meant, and one understood the item to be about something other than drying the swimsuit.

However, the remaining four students understood all the terms in the item and knew about how things dry. But these students interpreted the question differently from other students, and thus created a different scale and a different ordering of the factors along that scale. We found that all four of these alternative scales are consistent with the language of the item. All of these students told us that the sun would contribute most to the drying of the swimsuit, and knew something about the contribution of the other factors to drying. Yet because the scale they constructed was different from the one the testmakers intended (shown Figure 3.1 above), they answered this item incorrectly.

For example, Nayilah interpreted the question "Which of the following is least likely to affect the rate at which the swimsuit dries" probabilistically, as: Which factor would be least likely *to be present* when one is hanging one's swimsuit outside. Thus, she chose "B. the speed of the wind," and explained that it would be the least likely to be present because "most of the time there's actually no wind outside". Her scale reflects an alternative interpretation of the question in which likelihood is measured as frequency over time as opposed to size of effect for a given period.

The three other students also displayed additional alternative scales that were consistent with the language of the item. Macayla interpreted the question as asking which factor would be least likely to affect the rate at which the swimsuit dries by stopping the drying process. Two other students, Paul and Andrew, interpreted the question as asking something like this: As each variable quantity (heat, speed, amount, temperature) varies from its minimum to its maximum values, which one will have the least effect on the rate at which Freda's swimsuit dries? Rather than imagining a single value of each factor listed in the answer choices, these two students treated the question as a thought experiment in which they could imagine even physically improbable values of the factors, such as when Paul said, "the temperature of the water in the pool, I thought that actually might affect it, because say the water, it was like, it was like freezing, the swimsuit might get frozen." For Andrew as well, the temperature of the water in the pool could vary widely and thus could have a significant effect on the rate at which the swimsuit, soaked with that water, dries. The question as stated does not rule out any of these students' alternative interpretations of the question.

These four students experienced difficulty with "Freda's Swimsuit" because of the multiple possible interpretations of the question, which led to multiple possible scales by which to judge the answer choices. While not due to the forced comparison structure alone, the underspecified nature of the question arose in part from the forced comparison structure, which requires multiple intermediate steps of interpretation. Those students who managed to construct the intended scale tended to get the answer right, while those who constructed an alternate scale did not, regardless of how much science the students knew. In sum, students' responses to "Freda's Swimsuit" provide evidence to support the hypothesis that Feature 6, forced comparison, can add unintended difficulty to items.

3.4 Conclusions

In this section, we have shown that interviewing students provides rich, illuminating, and often unexpected evidence about students' performance on MCAS science items that complement text analyses and statistical analyses. The portraits of student reasoning that emerged from our interviews complement the description of factors affecting item difficulty given in Section 2. While the statistical analyses we ran on Features 6 and 7, forced comparison and multiple frames respectively, did not have a significant effect on student performance on test items across the state, we found these two features to be significant sources of difficulty for many of the students we interviewed. Furthermore, these features were the most significant sources of difficulty for students who demonstrated knowledge of the science but still answered the items incorrectly. In many cases, these sources of difficulty had to do with the existence of multiple, reasonable interpretations of the item. The features of atypical perspective-taking, forced comparison, and multiple frames, in combination, helped to create the conditions in which multiple alternative interpretations could be constructed for the same item.

Finally, we would like to note that while we were able to identify several significant, problematic features of items through text analyses and analysis of state data, we couldn't have predicted ahead of time the range and number of interpretations constructed by students in these interviews. Thus, the interviews provided a perspective that other data could not on the ways that students interacted with each of these features, and on how multiple features interacted in providing unintended sources of difficulty in these items. While large state data sets can provide excellent opportunities to analyze patterns of responses among groups of students, they do not allow us to see *how* students reason as they select an answer. Interviews with students on the other hand can help us to do just that.

4. Item Modification Analysis: Further Evidence

4.1 Introduction

So far, we have described our use of two different approaches to understand possible sources of differential performance on MCAS science items by students who speak a language other than English at home and who qualify for free/reduced lunch programs. Our correlational and regression analyses, reported in Section 2, showed that certain features add to item difficulty for all students, some more so for LEP and Y-F/RL students.

As reported in Section 3, we then used individual student interviews to further investigate possible sources of difficulty for our target students. The interview data strongly suggests that some features that did not reach statistical significance in predicting item difficulty on the state-wide MCAS data nevertheless were substantial obstacles for some students in our interviews. In other cases, interview data showed us that some features do not necessarily constitute obstacles in isolation, but in combination with other features they may invite students to construct alternate interpretations than those intended by testmakers.

In this section we report the results of a third approach to our question: manipulating the contents of an item's stem in an attempt to gain evidence about sources of unintended difficulty. We are not the first to use this approach. A number of researchers have used it to explore the effects of linguistic complexity in test items, with the aim of disentangling the assessment of subject matter knowledge from issues of language comprehension (Abedi et al., 1997, Abedi et al., 1998; Abedi, 2000; Durán, 1989; Durán, O'Connor & Smith, 1987). A major line of work has investigated the effects of modified, or simplified, language on the performance of English language learners (ELLs) (Abedi et al., 1997, Abedi et al., 1998; Abedi, 2000; Durán, 1989; Solano-Flores & Nelson-Barber, 2001).

The results of these studies make up a varied landscape. In some cases, the performance of ELL students improved when the linguistic complexity of items was reduced (e.g., long sentences were shortened, complicated question phrases were simplified) (Abedi, 1997). In other cases, the performance of both ELL students as well as students whose first language is English improved when changes were made to the "non-technical 'ordinary' language" of items but not to "special mathematics vocabulary and structures"

(Abedi et al., 1998, p. 3). Finally, Abedi and his colleagues found that reducing the linguistic complexity of items can reduce the performance-gap between ELL and non-ELL students (Abedi et al, 2000; Abedi et al., 1998).

While item modification does not shed light on test-takers' interpretive processes per se, because it yields percentage of correct responses, it can be an important source of information when students' performance on modified and unmodified versions of items is compared. We developed and employed a variation of these modification studies to triangulate what we were learning from correlational analyses, text analyses of items and analyses of student interviews. Thus our aim in modifying items is not to construct new and better items but rather to evaluate whether, by manipulating specific features of interest, we could affect student performance.

4.2 Modifying and administering items

It is important to note that this pilot study of item modification actually took place *prior* to the work reported in Sections 2 and 3. From the beginning of this project, as we scrutinized MCAS items, we invariably found ourselves proposing ways to improve them. Our intuitions (honed on a variety of sources, including pilot interviews and other researchers' studies of student performance) told us that some items' wordings were opaque or unnecessarily complex, and some items' illustrations were confusing. Eventually, we realized that our intuitions and inchoate hypotheses had to be put to the test, and we decided to construct modified versions for each of 19 items. Items were chosen for modification to represent what we thought were a wide range of potential problems. We modified the stems of these 19 items, leaving all answer choices intact. Our aim was to preserve the science content of an item, making as few changes in the text as possible while still addressing the issues we hypothesized might be adding unintended difficulty for students in our target groups.

Our modifications included changing confusing graphics, clarifying complex language, unpacking or making more explicit connections we thought were crucial to understanding an item, and replacing forced comparison questions with equivalent questions with simpler structures. These are exemplified below. For some of these 19 items, we changed both the item's graphic element and its language. In other cases we changed only the language. In every case we tried to clarify and remove sources of unintended complexity unrelated to science content.

We aggregated the 19 modified versions and the 19 original versions into two test forms, each containing half original and half modified items. (So for example, Form A of the test included the original version of "Freda's Swimsuit" and the modified version of "Desert Fog", while Form B included the modified version of "Freda's Swimsuit" and the original version of "Desert Fog.") We then randomly administered these two test forms to 9 classes of 5th graders in four schools (for a sample of 101 students). We assumed that our random administration of the test within classrooms would allow us to make inferences about how students in our target groups would perform on the modified versus the original version of an item, even though no one student responded to both the original and the modified version.[10]

4.3 Did our modifications succeed?

Our modifications were aimed at detecting unintended difficulty, factors unrelated to science knowledge. What would it mean to succeed in this aim? If we had correctly identified and remedied all sources of unintended difficulty, students in our target groups (and perhaps all students) should show significantly higher pass rates on modified items.

First we looked at total results for all students. A paired samples t-test, comparing percent correct on original and modified versions of each item, showed that there was no significant change in performance overall from original to modified items (t= -.72, df 18, p=.47). We wondered whether there might be an improvement in the performance of some sub-group, perhaps for F/RL students. Again, we found no significant improvement (t= .41, df 18, p= .68).

A closer look at the results for each item—the percentage of our sample that got the item correct—revealed that our null t-test results were in fact obscuring large changes in the difficulty of some items, some in a positive direction and some in a negative direction. While a fair number of items showed little or no change in difficulty, several of our modifications resulted in significant improvements while other modifications appeared to make student performance worse!

As indicated in the graph below, our efforts to remove unintended difficulty had mixed results at best. For some items (bars near zero line), our changes made essentially no difference in student performance. For other items (bars to the right substantially above the line), our modifications seemed to result in much better performance by students overall. But as indicated by the bars extending below the line, some of our modifications actually resulted in a steep drop in performance.

[10] Note that when we compared performance on the original and modified version of an item, we were not comparing performance by the same students. Although we randomized distribution of the two test forms (and thus, the original and modified versions of items) within each of our classrooms, it is still possible that with our small sample there was error from differential levels of ability in each subsample. Accordingly, we found that students who took Form A had a slightly higher average score than students who took Form B. Therefore, we attempted to adjust scores for this difference by using the appropriate scale factor for each item.

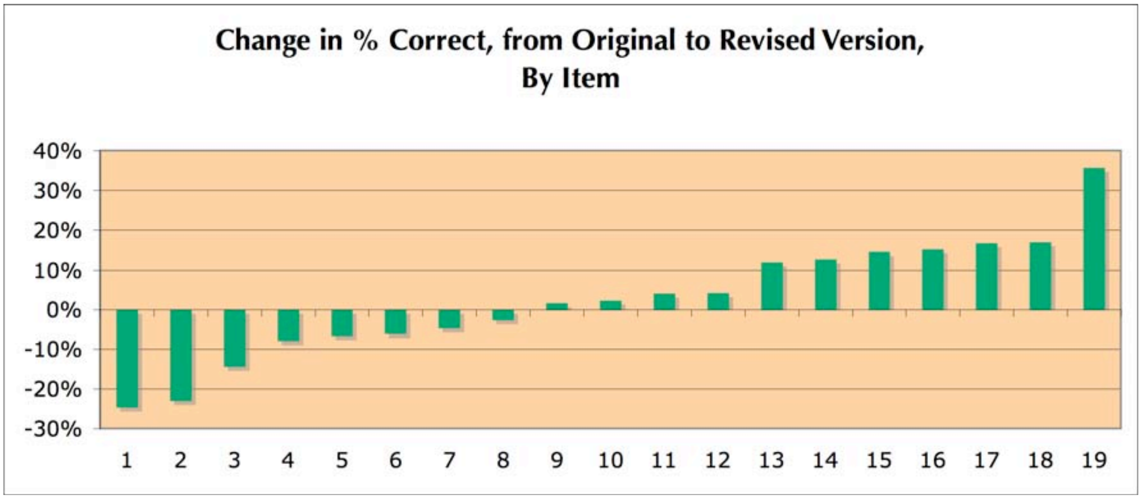


Figure 4.1 Change in percent correct, from original to revised version, by item.

The results in Figure 4.1 are for all 101 students, and are arrayed by item from the one that achieved the largest decrease in percent correct between original and modified version to the one that achieved the largest increase in percent correct between original and modified version.

Is the same range of improvement (or decline) found for a demographic subgroup, such as Y-F/RL students versus students N-F/RL students?[11] In the graph below, the items are arrayed by performance from worst to best for Y-F/RL students, i.e. the item at the far left resulted in the greatest decrease in percent correct responses for Y-F/RL students, and the item at the far right resulted in the greatest increase in percent correct responses for Y-F/RL students. The percent correct responses for N-F/RL students for each item is paired with those for the Y-F/RL students.

[11] We are not able to present results in this section for English Only vs. LEP students. As others have noted, categories and labels for language background are problematic at best. Some include LEP students with low English fluency, bilingual students who are fully fluent in their first language and English, and students who are fully fluent in English but have relatives at home who speak another language. Based on the information provided to us by our cooperating schools, it was difficult to determine the language fluency of the 101 students in our sample. Therefore, we will discuss our modification results only in terms of the groups designated by the free lunch indicator.

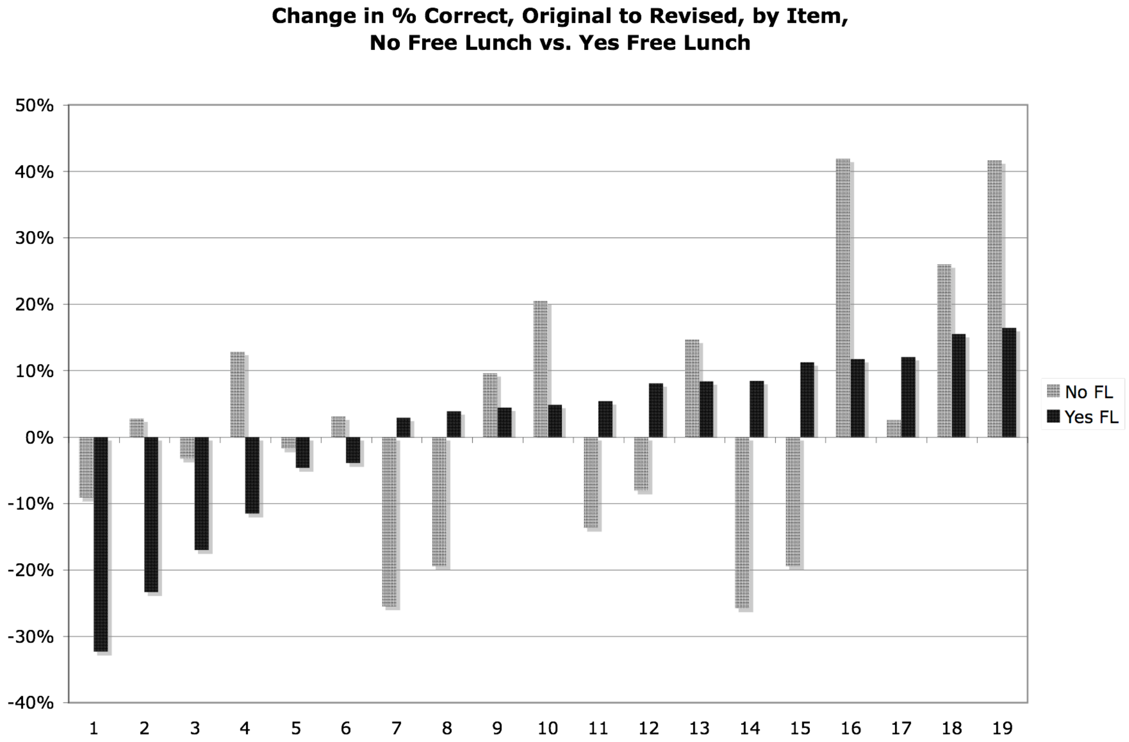


Figure 4.2 Change in percent correct, original to revised, by item for students who do and do not participate in free/reduced lunch programs.

As is apparent by inspection, the performance of N-F/RL students does not precisely track that of Y-F/RL students. The correlation between the two is only 0.250, although the difference between the two groups does not rise to significance in a Student t-test for unpaired samples (t= .78, p= .48). It is evident that for some items, our modifications affected responses for both groups in a similar way. Other items show that some modifications benefited N-F/RL students and worsened performance by Y-F/RL students (e.g. items 2 and 4 above). In still other cases, Y-F/RL students benefited while N-F/RL students did not (e.g. items 7, 8, 11, 12, ad 14 above).

These results do not promise easy answers, yet they provide us with additional albeit preliminary evidence for our hypotheses about unintended item difficulty. In the rest of this section, we briefly present some of the highlights of this complex set of results.

4.4 Two successful modifications

4.4.1 Modifications to "Freda's Swimsuit." The biggest improvement in performance came with our modifications to Freda's Swimsuit. The original version of this item is discussed at length in Sections 2 and 3.

Original stem:
Freda always hangs her wet swimsuit outdoors after getting out of the swimming pool. Which of the following is **least likely** to affect the rate at which Freda's swimsuit dries?

Modified stem:
Freda always hangs her wet swimsuit outside to dry after swimming in a pool. Which of the following would probably **NOT** affect how quickly her swimsuit dries?

Answer choices:

A. the heat of the Sun
B. the speed of the wind
C. the temperature of the water in the pool
D. the amount of water vapor in the air

Our modification removed the phrase "*is **least likely** to*" and replaced it with "*would probably **NOT**.*" This was an attempt to retain the content of the question while removing the feature of forced comparison. We also changed the complex nominal "the rate at which Freda's swimsuit dries" to the less complex "how quickly her swimsuit dries".[12]

Note that we did not modify the feature "Presence of Negative in Stem". It is clear from the interview data that the presence of the negative makes this item more difficult. Several of our interviewees read the original version as asking which factor would *most* likely affect the rate at which the swimsuit dries. By leaving the negative in place in the modified version, we are able to make stronger inferences about the source of improvements.

Results of Modifications. Recall that in the state-wide administration of the original version, 54% of students got this item correct, making it a relatively difficult item. In our sample of 101 students, 28% of students who received the original version of the item got it correct, while 63% of the students who received the modified item got it correct, an improvement of 35 percentage points. A chi square test is significant (X^2= 7.592, df=1, N=58, p= .0058).

When we look at subgroup analyses, we see that our changes resulted in improvements for both Y-F/RL and N-F/RL subgroups students. The original version of "Freda's Swimsuit" elicited a correct answer from 43% of N-F/RL students. The modified version elicited a correct answer from 91% of these students. Among Y-F/RL students, the

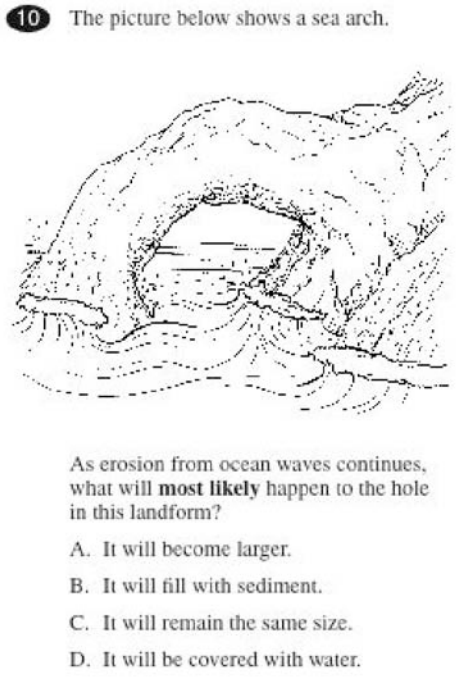
[12] In our clarificational zeal, we also changed "outdoors" to the more common "outside." (A preliminary Google search reveals no instances of the phrase "hang… swimsuit outdoors" but reveals several instances of "hang…swimsuit outside". It seems unlikely that this change would add substantially to student performance. We also made explicit the purpose of hanging the swimsuit outside by adding the purpose adjunct "to dry." As all students were familiar with the scenario of hanging a swimsuit outside to dry, this addition probably did not contribute much either. However, these are empirical questions in need of further investigation.

original elicited correct responses from 23% and the modifications received correct responses from 42%. The effect of the modifications for Y-F/RL students was smaller, but still significant by a chi square test (X^2= 3.88, df=1, N=41, p< .05).

To what can we attribute this positive change? The major modifications in the text involve (a) a removal of the forced comparison, and (b) a simplification of the phrase "the rate at which Freda's swimsuit dries." In light of our interview results, in which all students seemed to be able to understand the "rate" phrase in the original, it seems reasonable to infer that the absence of the forced comparison feature may be the primary reason that student performance improved on the modified item.

Notice that the continued presence of the negative in the modification ensures that this item still displays Feature 5, *Atypical perspective-taking*. It still requires students to think about the processes of change of state, specifically evaporation. However, instead of asking about what factors will facilitate the process of evaporation, readers are asked to think about which factors will not facilitate evaporation. This tripped up some students in our interviews. It may be that when the unintended difficulty of a forced comparison structure was removed, fewer students were challenged by the atypical perspective feature.

Modifications to "Sea Arch." The modification we did for "Sea Arch" resulted in marked improvement for all students and for Y-F/RL students in particular. The original versions appear in Figure 4.3; the modified version follows Figure 4.3.



10 The picture below shows a sea arch.

As erosion from ocean waves continues, what will **most likely** happen to the hole in this landform?

A. It will become larger.

B. It will fill with sediment.

C. It will remain the same size.

D. It will be covered with water.

Figure 4.3 Item 10, 2004, "Sea Arch."

Our modifications left the picture and the introductory sentence in place ("The picture below shows a sea arch.") but changed the stem to read as follows.

> A sea arch is formed by waves crashing against a large rock in the ocean. The waves cause the layers of rock to erode. Gradually, the waves create a hole in the middle of the rock. Over time, what will probably happen to the hole in this sea arch?

What did we change? Notice that the original does not make an explicit connection between erosion and the hole in the rock that constitutes the sea arch. It introduces the sea arch and then asks "As erosion from ocean waves continues, what will happen…". By their linguistic choices, the authors introduce erosion from ocean waves as something that is already shared knowledge. The verb *continues* entails that its argument "erosion from ocean waves" has taken place in the past. By introducing erosion as entailed, the authors add one step to the inferences students must make. Furthermore, by introducing this information in an adverbial, the authors treat the fact that erosion from ocean waves will continue as presupposed, that is, as part of the background to the main clause question about what will happen. Finally, the link between the hole in the sea arch and the process of erosion is implicit in the sequence of sentences in the original. Beyond the label "the picture below shows a sea arch," almost nothing is explicitly asserted.

In the modified version, we made all of these relations more explicit by adding three sentences. (*A sea arch is formed by waves crashing against a large rock in the ocean. The waves cause the layers of rock to erode. Gradually, the waves create a hole in the middle of the rock.)*. In addition, we changed the "forced comparison" format, modifying "what will most likely happen to the hole in this landform" to "what will probably happen to the hole in this sea arch."

Results of modifications. For all students, correct responses to "Sea Arch" improved from 46% to 60%. A chi square test was not significant for the entire sample (X^2= 2.141, df=1, N=101, p= .1433). However, for Y-F/RL students, correct responses improved from the original to the modified version by a total of 16 percentage points, to a 59% correct response rate. This did reach significance as a trend (X^2= 2.77, df=1,N=76, p=.09). To what can we attribute improvement in this case?

Interestingly, our modification made the stem of the item much longer: the original contains 17 words, the modification contains 49. We know from the literature and from the results of our Linguistic Load factor as described in Section 2 that almost tripling the length of the item should add difficulty. Yet performance improved. Thus we can assume that our additional explication was beneficial for student understanding.

"Sea Arch" is one of the items that we coded as exemplifying the feature *Multiple Frames*. It opens with a statement and a picture that represent a static property: this entity is a sea arch. The next sentence points the student to an ongoing process, and invites her to reason about a range of possible outcomes. In addition, "Sea Arch" contains a forced comparison format. As demonstrated with interview data in Section 3, both of these features can add difficulty to the task of constructing understanding.

What did our modification do in terms of these features? From one perspective, each sentence added explicitness about the ongoing process of erosion and its relation to the sea arch. The modified passage explicitly asserts that erosion and the sea arch are related by the feature of the sea arch that is made most salient by the question: that it has a big hole in the middle of it.

Performance in this item improved with modification, but questions remain about how students' step by step interpretation of the text and processing of the answer options was changed by our modifications. To understand this better, and to make specific suggestions for the improvement of items, we need to further explore this through student interviews.

4.5 Examining one of our failures.

Our failures are a bit more puzzling. We will present an analysis of one, "Desert Fog." This modification was particularly egregious in its failure, as N-F/RL students improved slightly (by 3 percentage points) while percent correct responses of Y-F/RL students fell by 23 points; only 32% of Y-F/RL students got this modified item correct.

> **Original stem:**
> Why does a town in the desert rarely experience early morning fog as compared to a town along the coast?
>
> **Modified stem:**
> Towns along the coast often have fog in the early morning, but towns in the desert rarely have fog in the early morning. What is the reason for this difference?
>
> **Answer choices:**
> A. There is less rainfall in the desert.
> B. Temperatures vary more in the desert.
> C. There is less water vapor in the desert air.
> D. There are fewer plants in the desert.

At the time, we considered our modifications to be very well-motivated. In the original version, the fact that towns in the desert rarely experience morning fog is presupposed. (Notice that this is a general property of the content of Wh-questions. If someone asks "Why do they hate us?" they are treating the proposition *"they hate us"* as a fact already given in the discourse, shared by their interlocutors.) In contrast, our modification does not presuppose that the reader knows that towns along the coast have more fog, etc. It first asserts the fact that towns along the coast and towns in the desert differ in the likelihood of early morning fog, and then, in a separate sentence, asks the reader to consider the reason for this difference.

The original version contains another potential source of complexity. The comparative adjunct (*as compared to a town along the coast*) is elliptical: it does not include an explicit mention of what is being compared (namely, the amount of morning fog one would find in "a town along the coast"). It does not explicitly include a predication about towns along the coast. Instead, everything is implied by the comparative. Our modification makes explicit this dimension of comparison.

It is true that our modification increased the number of words in the stem from 20 to 30, but as we saw in "Sea Arch", this is not always a predictor of poorer performance. If the increase in words adds to intelligibility, as we believed it did in this case, it should improve performance in spite of the increase in Linguistic Load.

One hypothesis has emerged as a possible source of unanticipated, unintended difficulty. Notice that in the modified version, the first clause introduces towns along the coast, and states that they often have morning fog. This positions "towns along the coast" as maximally thematic: it is a subject and it is in the first sentence. What follows may be read as a comment on this theme: towns along the coast have lots of fog, and towns in the desert don't. Why is this? Without looking at the answer choices, one would probably begin an answer with a statement about towns along the coast. But notice that all of the answer choices are stated in terms of towns in the desert.

This is a subtle point that concerns non-categorical, gradient, effects of information packaging in sentences and texts. It may or may not be a factor that contributed to the marked degradation in performance by Y-F/RL students. Further exploration is clearly necessary. However, there is a larger importance to our unsuccessful modification.

In general, the members of our research team are experts on language structure and function, and the construction of meaning from texts. We blithely made what we assumed would be improvements to "Desert Fog", and did not notice that we had changed the thematic structure of the question. Our changes resulted in substantially worse performance. This item constitutes a cautionary tale for those who think that it should be easy to improve the fairness and clarity of multiple choice test items. Clearly, much more work is needed before strong recommendations are possible.

4.6 Discussion of Results of Modification Study

The results of our modification study are what one might call a "mixed bag." On the one hand, a modification that eliminated Feature 6, forced comparison structure, improved the performance of all students for "Freda's Swimsuit". The improvement was smaller for Y-F/RL students than it was for N-F/RL students. We also found that although our modifications to "Sea Arch" increased the number of words in the stem, and therefore the Linguistic Load factor, the performance of all students improved. Again the improvement was smaller for Y-F/RL students than it was for N-F/RL students. This suggests that students can benefit from the presence of additional material when that material serves to further explicate the meaning of an item. Finally, the modifications we made to "Desert Fog" resulted in essentially no change for N-F/RL students and a marked decline in the performance of Y-F/RL students. We are somewhat puzzled by this but attribute it at least in part to an unintended change in the thematic structure of the question.

Our modification study provides additional support for the hypothesis that Feature 6, forced comparison, can add unintended difficulty to items, and it suggests that an increase in Linguistic Load, in this case embodied in the number of words in the stem, does not always result in an increase in difficulty. Our clearest finding, however, is perhaps to confirm what test-makers know in spades, that constructing good test items is an extremely complex and difficult process.

Section 5. Conclusions

This paper reports results from a study that was designed to investigate the difference in performance among students from different racial, linguistic, and socioeconomic backgrounds on the 5th grade science MCAS test, i.e., the so-called "achievement gap." We were drawn to this work because the gap in scores on the MCAS was strikingly similar to the gap observed widely for all standardized achievement tests. The second was the puzzlement – and distress – expressed by teachers we know when they compared the daily evidence they had of their students' classroom performance to their MCAS scores. Although their students knew the subject matter being tested, but they still tended to score at lower levels on the MCAS than middle class students. The teachers could not explain this discrepancy to themselves, their students, or parents. Like many others before us, we entered this work hoping to uncover possible sources of unintended difficulty that might be influencing students' performance.

To do this, we used a three-pronged approach that offered multiple, independent and complementary lenses on students' performance. These included: 1) investigations of possible unintended sources of item difficulty through text analyses of MCAS items and statistical investigations of student performance using large datasets made public by the MA DOE; 2) investigations of how and why students choose the answers they do through clinical interviews; and 3) investigations using items we modified to explore preliminary hunches about unintended sources of difficulty. Each of these approaches allowed us to formulate important insights into student performance and also gave rise to further questions for future study.

From our first set of investigations focused on text analyses and statewide performance data, we learned that several features and composite factors affect student performance. These include what we termed: a) linguistic load (i.e., the number of semantic units in the item, in the answer choices, and/or the presence of negation in the stem), and b) extended science knowledge (i.e., items that present unfamiliar contexts and/or that require atypical perspective-taking). When these features are present, student performance tends to decrease.

In our interview study, we learned that features that were not significantly correlated with performance in our first study (e.g., forced comparison, multiple frames) did in fact present students with considerable challenge. Importantly, the interviews revealed that these features were sources of difficulty for students whose scientific knowledge was questionable as well as for those who clearly knew the targeted science. The interviews also revealed that features like forced comparison and atypical perspective may interact in surprising and unpredictable ways to create conditions in which students, sometimes creatively and often quite reasonably, generate multiple alternative interpretations for the same item. In combination, these features can present insurmountable challenges to students.

Finally, our modification study, which it is important to note was conducted *prior* to the other two studies, yielded mixed results. On the one hand, it provided an additional source of support for our hypothesis that the forced comparison structure can add unintended difficulty to items and lead to decreases in student performance. On the other, it suggested that an increase in Linguistic Load does not always result in an increase in difficulty; students' performance can improve if the additional information makes the meaning of an item more explicit.

Test-makers generally worry about two kinds of error patterns. They worry about a) students who do not know the subject matter being tested but manage to get the right answer, and b) students who do know the subject matter being tested and, for whatever reason, get the wrong answer. When these patterns appear, it suggests that items may not be measuring what they purport to measure. Our results suggest that both kinds of error may be in operation for the 5th Grade Science MCAS. Perhaps not surprisingly, we found that students from communities placed at risk in school primarily fell into error pattern (b). Interestingly, however, we also found instances in which middle class students who had advanced subject matter also fell into this error pattern. For example, the son of a university professor who answered "Freda's Swimsuit" incorrectly told us that he would have preferred to do an experiment to figure out the answer to the question!

If we step back and consider the complex phenomena involved in answering a multiple choice item, we should not be surprised by this. Test-taking is, after all, an interactive communicative event in which a student must construct meaning for texts created by a test-maker who is not present. By design, items contain few communicative clues. For their part, test-makers must assume shared meaning for the items and images they employ. But, as we have seen, the range of resources available to students leads to diverse and multiple interpretations. As our own somewhat naïve attempts to modify items illustrate, predicting the meaning that students will construct is exceedingly difficult.

One important implication of this study is the benefit of using multiple methods when investigating a problem as thorny as that posed by the "achievement gap." Our study employed three approaches to investigating, analyzing and understanding student performance, each of which allowed us to ask and explore different kinds of questions. The answers to these questions, in turn, provided us with some important insights into possible sources of unintended difficulty in multiple choice items that may differentially affect students from different backgrounds.

A second implication is to use caution when interpreting the meaning of student performance on tests like the MCAS. When a student does not score well, the tendency is to assume the student has not learned the subject matter material. When a class of students does not do well, the tendency is to assume the teacher has not taught the material, or has not taught it well. When an entire school of students fails, the tendency is to assume that administrators, teachers and students are all at fault. The results reported here, while preliminary, suggest a far more complicated picture that at the very least casts doubt on the use of scores on MCAS and similar high-stakes tests to make consequential decisions about students, teachers, administrators, and schools.

References

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometric issues. *Educational Assessment, 8,(3), 231-257.*

Abedi, J., Bailey, A., Butler, F., Castellon-Wellington, M., Leon, S., & Mirocha, J. (2000/2005). *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives.* (CSE Technical Rep. 663). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education and Information Studies, University of California.

Abedi, J., Lord, C., Boscardin, C. K., & Miyoshi, J. (2001). *The effects of accommodations on the assessment of limited English proficient (LEP) students in the National Assessment of Educational Progress (NAEP)* (CSE Technical Rep. 537). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education and Information Studies, University of California.

Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance*. Los Angeles, CA: Center for the Study of Evaluation.

Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2001). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice, 19*(3), 16-26.

Abedi, J., Lord, C., & Plummer, J.R. (1997). Final report of language background as a variable in NAEP mathematics performance. Los Angeles: Center for the Study of Evaluation, CSE Technical Repot #429.

Bailey, A. (2000/2005). Language analysis of standardized achievement tests: Considerations in the assessment of English language learners. In J. Abedi, A. Bailey, F. Butler, M. Castellon-Wellington, S. Leon, & J. Mirocha, *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives.* (CSE Technical Rep. 663). pp. 79-94. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education and Information Studies, University of California.

Butler, F. A., & Stevens, R. (1997). (CSE Technical Rep. 448). Los Angeles: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Durán, R. (1989). Testing linguistic minorities. In R. Linn (Ed.), *Educational Measurement* (3 ed., pp. 573-587). New York: MacMillan.

Durán, R.P., O'Connor, M.C. & Smith, M. (1987). Methods for assessing reading comprehension skills of language minority students. *Technical Report, CLEAR Project 2.2.* University of California, Los Angeles: Center for Language Education and Research.

Frederiksen, N. (1990). Introduction. In N. Frederiksen & R. Glaser & A. Lesgold & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. ix-xvii). Hillsdale, NJ: Lawrence Erlbaum Associates.

Freedle, R. O. (2003). Correcting the SAT's ethnic and social-class bias: A method for reestimating SAT scores. *Harvard Educational Review, 73*(1), 1-43.

Gee, J.P., Michaels, S., and O'Connor, M.C. (1992). Discourse analysis. In M. Lecompte, W. Milroy and J. Preissle (Eds.), *Handbook of qualitative research in education,* pp. 227–291. New York: Academic Press.

Gehring, J. (2001). Mass. school policies praised as test scores rise, *Education Week*.

Haladyna. (2004). *Developing and validating multiple-choice test items* (Third ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Publisher.

Hamilton, L., Nussbaum, M., & Snow, R. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10*(2), 181-200.

Haney, W. and Scott, L. (1987). Talking with children about tests: An exploratory study of test item ambiguity. In R.O. Freedle and R.P. Durán (Eds.), *Cognitive and linguistic analyses of test performance*, pp. 298-368. Norwood, NJ: Ablex.

Hill, C., & Larsen, E. (2000). *Children and reading tests*. Stamford CT: Ablex.

Massachusetts Department of Education. (May 2001). *Massachusetts Science and Technology/Engineering Frameworks* Malden MA: Massachusetts Department of Education.

Kazemi, E. (2002). Exploring test performance in mathematics: The questions children's answers raise. *Journal of Mathematical Behavior*, *21*, 203-224.

Kress, G. and T. van Leeuwen (1996). Reading images: The grammar of visual design. London, England, Routledge.

Langer, J. (1987). The construction of meaning and the assessment of comprehension: An analysis of reader performance on standardized test items. Cognitive and linguistic analyses of test performance. R. O. Freedle and R. P. Durán. Norwood, NJ, Ablex**:** 225-244.

MacKay, R. (1974). Standardized tests: Objective/objectified measures of "competence". Language use and performance. A. Cicourel, K. Jennings, S. Jennings et al. New York, Academic Press**:** 218-247.

Mehan, H. (1975). The reality of ethnomethodology. New York, John Wiley & Sons.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3 ed., pp. 13-103). New York: MacMillan.

National Science Resources Center. (1991). *Plant growth and development*. Washington, D.C.: Smithsonian Institution-National Academy of the Sciences.

O'Connor, M. C. (2006). The implicit discourse genres of standardized testing: What verbal analogy items require of test takers. In J. Cook-Gumperz (Ed.), *The Social Construction of Literacy* (pp. 264-287). Cambridge: Cambridge University Press.

Pellegrino, J., & Chudowsky, N. (2003). The foundations of assessment. *Measurement, 1*(2), 103-148.

Pellegrino, J., Chudowsky, N., & Glazer, R. (Eds.). (2001). *Knowing what students know*. Washington DC: National Academy Press.

Ruiz-Primo, M. A. (February, 2002). On a seamless assessment system, *In Seamless Science Education. Symposium conducted at the annual meeting of the American Association for the Advancement of Science*. Boston.

Shavelson, R., Carey, N. B., & Webb, N. M. (1990). Indicators of science achievement: Options for a powerful policy instrument. *Phi Delta Kappan, 71*(9), 692-697.

Solano-Flores, G. (2006). Language, dialect, and register: Sociolinguistics and the estimation of measurement error in testing of English language learners. *Teachers College Record, 108*(11), 2354-22379.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching, 38*(5), 553-573.

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

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

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

Steele, C. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52*(6), 613-629.

Steele, C., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*, 797-811.