Assessment Linked to Middle School Science Learning Goals: A Report on Field Test Results for Four Middle School Science Topics

George E. DeBoer, Cari F. Herrmann-Abell,
Jill Wertheim, and Jo Ellen Roseman
AAAS Project 2061

NARST Annual Conference
Garden Grove, CA

April 19, 2009

This paper is a summary of certain aspects of an assessment development project being conducted at AAAS Project 2061 (DeBoer, et al., 2007, 2008a, 2008b). The project involves developing multiple choice assessment items for 16 middle school science topics closely aligned to the ideas in *Benchmarks for Science Literacy* (American Association for the Advancement of Science [AAAS], 1993) and the *National Science Education Standards* (National Research Council [NRC], 1996). The items can be used to gather data on what students currently know, the growth in their knowledge over time, or gains in their knowledge made following instruction. In this paper, we discuss the results of field testing of four topics: Matter and Energy in Living Systems; Chemical Reactions and Conservation of Matter; Plate Tectonics; and Control of Variables.

The item development procedure involves rigorous qualitative alignments and the use of quantitative psychometric methods. It is this balance of the two approaches that distinguishes our work from many other approaches to item development. The items discussed in this paper have gone through this rigorous development process, including pilot testing, expert review, and various statistical analyses.

Qualitative Aspects of Item Development

- *Elaboration of the target learning goals into clarification statements, which serve as precise item writing specifications*. Each content standard is unpacked or elaborated into more detailed statements that describe precisely what our expectations are for students. Our goal is to describe the boundaries for testing so that only those things are tested that are specified in these clarification statements. If we find that an item is outside the bounds of what is indicated in the clarification statement, the item is either revised to fit with the clarification statement or the clarification statement is expanded to include the additional knowledge. This forces us to look closely at the items in relation to the learning goals so that they are in agreement with each other.
- *Identification of misconceptions, naïve conceptions, and alternative ideas that students have.* In addition to learning what students do not know, we are also interested in finding out what alternative ideas they have. For some topics there is a rich literature that summarizes the various misconceptions and naïve conceptions that students have. When the research literature is inadequate for identifying the ideas that students have, we use open-ended interviewing to supplement the research literature, and we create answer choices to probe for suspected misconceptions during pilot testing.
- *Use of misconceptions as distractors in test items*. After we identify the misconceptions that students are likely to have, we embed them in the test questions (see Sadler, 1998 for a discussion of distractor-based multiple choice testing). Almost all of our field test questions have one or

more misconceptions that they are testing at the same time they are testing the ideas in *Benchmarks* and *Standards*. Because we are field testing each item with at least 1000 students per item, and using a standard procedure across all topics, we can provide an up-to-date assessment of the most common misconceptions that students have about these ideas and the relative strength of these misconceptions.

- *Getting feedback from students about the appropriateness of the items from pilot testing and interviewing*. We also draw heavily on student responses to our test items during pilot testing. Students tell us whether anything is confusing about an item, whether there are words they don't understand, and they tell us why each answer choice is correct or incorrect. We use the match (or mismatch) between their answer selections and their explanations to identify items that are likely to yield either false negative or false positive responses. Our goal is to remove as much of the construct irrelevant variance as possible through this means. We have also found it informative to give the test questions to older students. If students who should know the correct answer give us incorrect answers, we then look more closely at the item to see if there is something wrong with the way it is written.
- *Analysis of Alignment to Target Learning Goals*. The items are both developed and later analyzed according to our set of Assessment Alignment criteria and indicators (DeBoer, et al., 2007, 2008a, 2008b). Panels that include experts in both science and science education are convened to evaluate each item's alignment, accuracy, and appropriateness. In short, the panels are asked to consider whether or not the knowledge specified in the targeted learning goals are both necessary and sufficient to correctly answer the item.

Quantitative Aspects of Item Development

- *Summary of what the students who were tested know for each item and for each cluster of items*. We determine the percentage of students who answered each question correctly and which answer choices they selected if they did not answer correctly. Clusters of items are used to measure students' understanding of key ideas under each topic heading. We also look at how well students did on each of the items under the various key ideas. Student performance depends on what they know, but we find that it also depends on the context of the item and on the kind of thinking that is expected of students.
- *Summary of misconceptions the students who were tested have*. We obtain a measure of the misconceptions that students have by counting the number of times that the students select answer choices that reflect those misconceptions and dividing by the total number of times they could have selected the misconception. If the misconception appears as an answer choice six times, and the average number of times it is chosen by the students responding is two, then the measure of that misconception in that group of students is 2/6 or .33.
- *Rasch difficulty measure (item maps).* We use Rasch modeling for a number of reasons. For one, we use the logit scores that Rasch provides to give us a difficulty score that is normalized from sample to sample. We typically field test about 50 items on 4-6 different test forms. We use several linking items common to all forms so that we can compare the data across forms. The Rasch item maps show us the spread of difficulty of the test items and the match between the ability of the student sample that we tested and the difficulty of the items for that sample. We learn from this if the set of items is a reasonable measure of middle school knowledge of each topic. We can also compare results across topics to see how well the ideas under each topic are currently understood by students in this country. In the data that are presented later, we will show, for example, that the match is reasonably good between the difficulty of the items and the ability of the students for the physical science topic dealing with chemical reactions and conservation of mass, but for the life science topic dealing with matter and energy in living

systems, the match is not as good. Fewer students understand the basic ideas described in *Benchmarks* and *Standards* for that topic than they do for the chemistry topic.

- *Rasch item and person reliability*. Rasch provides a reliability measure for the entire set of items for a topic. The result is in part an indicator of whether there were enough students to provide a reliable measure of the difficulty of each test item. Rasch also provides a reliability measure for the sample of students. The result is in part an indicator of whether there were enough items to provide a reliable measure of the ability of students at each ability level.
- *The point-measure correlation coefficients of the items are used to determine the contribution of each item to the set of items on a topic (related to dimensionality).* When a point-measure correlation coefficient is very low, we examine the item to determine if it is not well aligned to the topic or if it is measuring a particular aspect of the topic that is not well represented in the other items.
- *Differential item functioning*. We examine results by gender and whether English is the student's primary language. In the paper we will provide examples of items that show significant under or over performance on the part of various groups of students.

General Methods

We tested middle school students in grades six through eight from school districts across the country. Each test form was distributed in such as way as to ensure that all questions were answered by high, middle, and low performing students in different parts of the country and in different urban, suburban, and rural settings. The sample was generated through the help of the Building a Presence Program of the National Science Teachers Association.

Because we were testing more items than students could complete in a typical class period, multiple test forms were created that contained subsets of the available questions. Also, because we were interested in describing both the students' understanding of the topic as a whole and their understanding of the ideas within each topic, some of the test forms included a random selection of all of the items and some of the test forms included items from selected clusters of the ideas being tested. In addition, for each form of the test, half of the students took the items in reverse order so that the last items on the test would not be disproportionately omitted if students ran out of time.

Topic 1: Substances, Chemical Reactions, and Conservation of Mass

Description of the Sample Tested. In chemistry, the field testing included 3337 students in grades six through eight. A total of 133 teachers from 122 schools in 30 states participated in the field testing. Approximately 25% of the students were in sixth grade, 43% in seventh grade, and 32% in eighth grade. Students from a wide range of urban, suburban, and rural school districts across the country responded to the items. Approximately 46% of the students were students of color, and 10% of the students indicated that English was not their primary language. About half of the students were female and half were male. The field test was administered in the spring of 2008. Approximately 1800 students responded to each item.

We also administered the items to populations of students who were likely to have the knowledge being targeted by the items. We gave the test to 474 students who had taken high school chemistry and were enrolled in a college level introductory chemistry course at a public university in a northeastern state, but who had not yet had any instruction at the college level. We refer to this group as High School graduates. We also tested 1218 students were currently enrolled in a chemistry course at two universities—a public

university in a southern state and a public university in a northeastern state. Approximately 46% of the college students were freshman, 36% were sophomores, 13% were juniors, and 4% were seniors. Approximately half of the students were female and half were male, and 45% of the students were students of color.

Description of the Target Learning Goals. The topic Substances, Chemical Reactions, and Conservation of Mass focuses on five key ideas. The ideas are based on Chapter 4, Section D of *Benchmarks for Science Literacy* (AAAS, 1993) and Physical Science Content Standard B of *National Science Education Standards* (NRC, 1996). The five key ideas are:

Idea A: A pure substance has characteristic properties, such as density, a boiling point, and solubility, all of which are independent of the amount of the substance and can be used to identify it.
Idea C: Many substances react chemically in predictable ways with other substances to form new substances with different characteristic properties.
Idea D: When substances interact to form new substances, the atoms that make up the molecules of the original substances rearrange into new molecules.
Idea G: Whenever substances interact with one another, regardless of how they combine or break apart, the total mass remains the same.
Idea H: Whenever atoms interact with each other, regardless of how they are arranged or rearranged, the number of each kind of atom stays the same and, therefore, the total mass stays the same.

Each key idea was further clarified in order to state precisely what students would be expected to know. These clarification statements act as item writing specifications that ensure a close alignment between the items and the learning goals. The clarification statement for Idea C says:

> Students should know that when substances react chemically, one or more new substances are formed. They should know that if a new substance does not appear, a chemical reaction did not occur. They should know that the products of a chemical reaction can be identified as new substances because each product has different characteristic properties from the original substances under the same conditions. They should know that liquids, solids, or gases can be reactants or products in chemical reactions. Students should also know that it is possible for a single substance to undergo a chemical reaction, such as when the substance is heated or an electrical current flows through the substance. They should know that it is not true that all chemical reactions are irreversible.
>
> Students are not expected to know that chemical reactions involve the rearrangement of atoms into new molecules. This idea is addressed Idea D. Students are also not expected to know that nuclear reactions are not chemical reactions nor why nuclear reactions are not chemical reactions. Nuclear reactions are addressed in later ideas (4E/H6* and 4G/H6*).

During the development of the assessment items, student misconceptions were incorporated into the distractors. The following is a list of some of the misconceptions that were tested.

- Weight and volume are characteristic properties of a substance (Smith, Wiser, Anderson, Krajcik, & Coppola, 2004).
- Temperature is a characteristic property of a substance (Thomaz, Malaquis, Valente, & Antunes, 1995).
- A chemical reaction occurs during a change of state (Ahtee & Varjola, 1998; BouJaoude, 1992; Hall, 1973; Novak & Musonda, 1991; Stavridou & Solomonidou, 1998).

- A chemical reaction occurs when a substance dissolves (Abraham, Williamson, & Westbrook, 1994; Ahtee & Varjola, 1998; BouJaoude, 1992; Eilks, Moellering, & Valanides, 2007; Novak & Musonda, 1991; Stavridou & Solomonidou, 1998; Valanides, 2000).
- A chemical change is irreversible (Calik & Ayas, 2005; Cavallo, McNeely, & Marek, 2003).
- Chemical reactions require two reactants (Cavallo et al., 2003; Eilks et al., 2007).
- The atoms and molecules of the reactants of a chemical reaction are transformed into other atoms and molecules (Andersson, 1986).
- The products of a chemical reaction, though unseen, must have somehow existed from the start in another location, like the air or inside the starting materials (Andersson, 1986; Solomonidou & Stavridou, 2000).
- Matter is not conserved during a change of state (Hall, 1973; Lee, Eichinger, Anderson, Berkheimer, & Blaskeslee, 1993; Stavy, 1990).
- Mass is not conserved during processes in which gases take part (Berkheimer, Anderson, Lee, & Blaskeslee, 1988; Hesse & Anderson, 1992; Mas, Perez, & Harris, 1987; Ozmen & Ayas, 2003).
- When a tree grows in a closed system, the total mass of the system increases (Mitchell & Gunstone, 1984). Or when mold grows in a closed system, the total mass of the system increases (AAAS pilot testing, 2007).
- When a chemical reaction occurs, atoms just disappear. For example, the atoms burn up (Andersson, 1986), or the number of atoms decreases when wood burns in a closed system (Mitchell & Gunstone, 1984).
- The formation of a gas during a chemical reaction is evidence of new atoms or molecules being produced (AAAS pilot testing, 2007).
- The number of molecules is always conserved. Some students confuse conservation of atoms and conservation of molecules (Mitchell & Gunstone, 1984).

Findings

Rasch Modeling. For all of the items combined, the data had a good fit to the Rasch Model, with a separation index for the items of 10.83 which corresponds to a test reliability of 0.99, and a person separation index of 1.61 which corresponds to a person reliability of 0.72. (The separation index represents the spread of student abilities or item difficulties and indicates the approximate number of different levels of difficulty or ability that can be reliably differentiated. A separation index greater than 1.5 is generally considered acceptable.) The point-measure correlation coefficients for the items ranged from 0.02 to 0.54 with a mean of 0.35. Figure 1 shows the item-person map for the 52 items included on the field tests. The map shows the range of person abilities on the left side of a vertical line and item difficulties on the right side of the line. Low ability/difficulty is represented at the bottom of the map and high ability/difficulty at the top. As can be seen on the map, the student abilities and the item difficulties have approximately normal distributions, and there is a high degree of overlap between ability and difficulty. The mean item difficulty is slightly higher than the mean student ability, which means that the average student score is slightly less than 50% for this set of items. Along most of the range of student ability, there are test items that provide a reliable measure of what they know about this topic, but for the lowest part of the ability distribution there are no items.

Figure 1: Item-person map showing the distribution of student abilities on the left and item difficulties on the right. Where item difficulty and person ability match, the person has a 50% chance of answering the item correctly. Item difficulties are shown for the 52 items included on the field tests.

Summary of What Students Know and the Misconceptions They Have

The average percent correct for all of the items was 39%. The percent correct was 41.6% for items aligned to Idea A, 46.6% for Idea C, 39.2% for Idea D, 34.6% for Idea G, and 35.9% for Idea H. The average Rasch item difficulties were -0.03 for Idea A, -.027 for Idea C, 0.04 for Idea D, 0.35 for Idea G, and 0.19 for Idea H. The results indicate that middle school students had the most difficulty with ideas related to conservation of mass (Ideas G and H) and were most successful with questions testing the idea that substances react chemically to form new substances with different characteristic properties (Idea C).

Grade-to-Grade Differences. We conducted an analysis of covariance (ANCOVA) to investigate if there were any differences in students' knowledge from grade to grade. Because our sampling procedures did not ensure that the sixth, seventh, and eighth grade students could be considered equivalent (see Table 1), an analysis of covariance was performed controlling for whether the students identified English as their primary language. (English as the primary language was significantly correlated with student performance.) ANCOVA showed that differences in performance by grade are significant at the 0.01 level of significance (F=4.308). The estimated marginal means for the overall percent correct for all items combined are reported in Table 2. A Bonferroni post-hoc test showed that the eighth grade students performed significantly better than both the sixth and seventh grade students, but there was not a significant difference between the sixth and seventh grades.

Table 1
The percentage and number of students in each grade by gender and whether they indicated that English was their primary language

	Female	Male	English	Non-English
Grade	% (N)	% (N)	% (N)	% (N)
6th Grade	52% (430)	48% (396)	93% (776)	7% (54)
7th Grade	51% (713)	49% (681)	88% (1220)	12% (173)
8th Grade	50% (528)	50% (526)	89% (943)	11% (111)

Table 2
Overall percent correct by grade (Estimated Marginal Means)

Grade	6^{th} Grade	7^{th} Grade	8^{th} Grade
% correct	38.5%	37.2%	40.5%

We found this same pattern for each of the ideas that were tested, with significant increases occurring between seventh and eighth grade but not between sixth and seventh. It is hardly an unexpected finding that students in eighth grade would do better than students in seventh or sixth grade, but for some topics that we have tested, this is not the case. In the case of chemistry, the improvement could be due to more students being taught chemistry in the eighth grade (something that teachers in our sample said was true for these students) or the greater maturity of eighth grade students that made them more likely to understand the ideas being taught and tested.

Idea A: Characteristic properties of substances

For Idea A (characteristic properties can be used to identify substances), about half of the students knew that boiling point (49.7%), melting point (55.2%), and density (56.0%) can be used to identify substances. Fewer students knew that odor could be used to identify substances, either for a liquid (29.4%) or for a gas (22.7%), or that solubility was a characteristic property (35.2%). Also, 22.9% of the students

incorrectly thought that mass or weight could be used to identify a substance, and 27.4% incorrectly thought that volume could be used to identify a substance. Approximately 20% of the students incorrectly thought that temperature was a characteristic property of a substance.

Idea C: Chemical reactions (substance level)
For Idea C (the products of a chemical reaction have different characteristic properties than the reactants), approximately 65% of the students knew that the product of a chemical reaction can be a solid, liquid, or gas, and a little more than half (55.3%) recognized the general principle that when reactants and products have different characteristic properties, this is an indication that a chemical reaction has taken place. But the students also held several misconceptions related to chemical reactions. A number of the students thought that dissolving is a chemical reaction (28.7%), that a change of state is a chemical reaction (19.1%), that a chemical reaction requires starting with at least two substances (65.9%), and that chemical reactions are never reversible (36.0%).

Idea D: Chemical reactions (atomic level)
For Idea D, a number of students knew that atoms rearrange to form new molecules during a chemical reaction and that they do not change into other atoms (39.5%). However, about half of the students incorrectly thought that at least some of the atoms *do* change into new atoms during a chemical reaction. In another item, about a third of the students incorrectly thought that the atoms of the reactants involved in a chemical reaction broke down and released the molecules of the products. This may be related to the confusion some students have about whether atoms are made of molecules or molecules are made of atoms.

Idea G: Conservation of mass (substance level)
For Idea G, students are expected to know that when substances mix, undergo chemical reactions, change state, or dissolve, or when objects are cut or broken into smaller pieces, the total mass of all the matter will always remain the same. The items aligned to this idea covered a variety of examples of these substance-level contexts. If the students know the rule that no matter what you do to a substance its mass always stays the same, they should be able to answer correctly no matter what the context. However, we found that students' performance on the set of items varied from 17.0% correct to 69.9% correct. Instead of correctly applying the rule that mass is conserved, they thought incorrectly that when mold grows on a piece of bread in a sealed bag, the system weighs more than before the mold grew (56.4%); that when liquid is heated in a closed system, the mass of the liquid increases (47.9%); and that the mass of a sealed plant-jar system decreases as the plant dies (41.2%). The only conservation task that a majority of students responded to correctly involved a stick of butter cut into pieces (69.9%).

Idea H: Conservation of mass (atomic level)
For Idea H, we developed two items that used circles to represent atoms and groups of circles to represent molecules. The students were able to identify the representation that had the same number of each kind of atom before and after a chemical reaction (45.5% and 52.3% on the two items). However, for an item that asked why the total mass of a system is the same before and after a chemical reaction, 35.1% of the students incorrectly thought it was because the number of *molecules* remains the same, not the number of each kind of *atom*. Additionally, 28% of the students thought that when a gas is produced during a chemical reaction, the total mass of the system increases because new atoms are produced, and 27% of the students thought that the total mass would decrease because atoms are destroyed. For one item that involved a liquid turning into a gas in a sealed jar, the stem explicitly stated that the number of atoms in the jar stayed the same and asked what happened to the mass of the jar and everything inside it. Even with the statement that the number of atoms remained the same, 17.8% of the students still thought the mass would increase and 30.3% still thought the mass would decrease.

Differential Item Functioning
For all of the items, we conducted a differential item functioning (DIF) analysis to investigate whether the items performed similarly by gender and by whether or not English is the student's primary language. We found only a few items on which one group or another either under or over performed on that item with respect to their expected score ($p<0.01$).

Gender. Girls over performed on an item that required them to know that a new substance always results from a chemical reaction. Boys under performed on this item. Boys scored better than expected on an item that tested the idea that the mass of a system would not change if a gas is formed or if a solid is formed. Girls scored worse than expected on this item. On a conservation of mass item involving cutting a stick of butter into pieces, girls performed better than expected, and boys performed worse than expected. Boys exceeded their expected performance on a conservation of mass item involving a chemical reaction in which a gas is produced. On all other items, boys and girls performed as expected given their ability and difficulty of the items. Overall, we were unable to find any patterns in the results for the DIF analysis for gender.

English as Primary Language. There were four items on which students for whom English was not their primary language either over or under performed. They performed not as well as expected on an item that asked them to identify an example of a chemical reaction. They exceeded their expected performance on a conservation of mass item involving a chemical reaction taking place in a sealed container. They also exceeded their expected performance on an item where they were given a table of properties and asked to identify which liquids were the same substance. The fourth item included a table of properties for three solids, and students were asked if any of the solids could be the same substance. We developed two versions of this item in our field test. The first had answer choices that simply listed the two solids (e.g. "Solids 2 and 3 could be the same substance.") The answer choices for the second version listed the two solids and then included a reason why the two solids could be considered the same substance (e.g. "Solids 2 and 3 could be the same substance. Even though they do not have the same mass, they have the same melting point and color.") We thought that the extra sentence in the answer choices of Version 2 might cause students who did not have English as their primary language to struggle with the item because of the extra words. However, our DIF results show that students for whom English was not their primary language scored better than expected for the item with the extra words in the answer choices, and they scored as expected on Version 1 without the extra words. This indicates that sometimes adding more words can actually clarify the item. Other than this, we did not identify any patterns in the DIF analysis for the items.

Assessing Higher Level Students as a Way of Discovering Issues with items
We expected the high school graduates and the college students to perform significantly better on all of the items compared to the middle school students and, for most of the items, this was true. Data for the high school graduates and the college students appears in Table 3.

Table3
Percent correct by idea for the middle school students compared to the high school graduates and college students

	Idea A	Idea C	Idea D	Idea G	Idea H
Middle School students	41.6%	46.6%	39.2%	34.6%	35.9%
High School graduates	64.1%	68.9%	65.8%	53.1%	59.4%
College students	80.6%	84.8%	87.5%	77.5%	88.2%

However, there was one item that the college students did not perform better on, so we took a closer look at the item to see why the college students did not perform well. Did the college students just not know the targeted idea, or is the item itself flawed?

The item, shown in Figure 2, dealt with comparing two samples of gases to see if they were the same substance. The students were asked to choose whether comparing the temperatures, volumes, weights, or odors of the gases would help them decide if the gases are the same substance. The intended correct answer was answer choice D: comparing the odors of the gas. We expected that students who chose that answer would do so because odor is a characteristic property and the other options are not characteristic properties. For all of the samples we tested, students did not do well on this item. The percent correct was 22.7% for middle school students, 20.6% for high school graduates, and 26.2% for students enrolled in various college chemistry courses. We knew from another item that most of the college students did know that odor is a characteristic property, with 71.1% of them answering correctly on that item. This suggested to us that the item had structural problems that needed to be addressed. We began by looking at the most popular answer choice that the college students selected.

Figure 2: Item SC96-3

You have two balloons. Balloon 1 is filled with a gas, and Balloon 2 is filled with a gas. The gases are pure substances. What could you do to help decide if the gases in the balloons are the same substance?

A. Compare the temperature of the gas in Balloon 1 to the temperature of the gas in Balloon 2.

B. Compare the volume of the gas in Balloon 1 to the volume of the gas in Balloon 2.

C. Compare the weight of the gas in Balloon 1 to the weight of the gas in Balloon 2.

D. Compare the odor of the gas in Balloon 1 to the odor of the gas in Balloon 2.

Table 4
Number and percentage of students who selected each answer choice for Item SC96-3

	A (Temp.)	B (Volume)	C (Weight)	D* (Odor)
Middle School students	295 (17.9%)	492 (29.9%)	485 (29.5%)	374 (22.7%)
High School graduates	41 (22.8%)	54 (30.0%)	48 (26.7%)	37 (20.6%)
College students	22 (8.0%)	30 (10.9%)	151 (54.9%)	72 (26.2%)

We found that the most popular answer choice selection made by the college students was answer choice C (comparing the weights of the gases). We also knew from the results of several other items, that most college students (> 90%) do not think that weight or mass are characteristic properties of substances. So we looked for reasons why they may have chosen weight over odor in this case. We speculated that the students may have assumed that the balloons were the same size and, therefore, had the same volume. If this is the case, then comparing the weights would be a very effective, and probably the best, way to decide if the gasses in the balloons are the same substance because equal volumes of different substances

usually have different weights. Testing the college students showed us that this item needs to be changed if it is to be used effectively.

Topic 2: Plate Tectonics

Description of the Sample Tested. In Spring 2008 we sent field tests assessing understanding of middle-school level ideas for the topic Plate Tectonics to 2051 middle school students across the United States. For all students tested, 48% were male and 52% were female; 93% of the students said that English was their primary language, and 8% said it was not (Table 5).

Table 5
The percentage and number of students in each grade by gender and whether they indicated that English was their primary language

	Female	Male	English	Non-English
Grade	% (N)	% (N)	% (N)	% (N)
6th Grade	52% (352)	47% (319)	89% (600)	9% (58)
7th Grade	51% (357)	48% (336)	93% (654)	5% (34)
8th Grade	52% (355)	47% (322)	90% (619)	9% (60)
Total	52% (1064)	48% (977)	92% (1873)	8% (152)

We also administered the test questions to one class of 17 college students to determine if the items would be answered correctly by a group of students who were expected to have the knowledge these items are targeting. This class included 11 male and 6 female upper-division geology students.

Description of the Target Learning Goals. The ideas being assessed in this topic are drawn from Chapter 4, Section C of *Benchmarks for Science Literacy* (AAAS, 1993) and Earth Science Content Standard D of *National Science Education Standards* (NRC, 1996). In this paper we report on assessment items focus on five key ideas:

Idea A**:** The outer layer of the earth – including the continents and the ocean basins - consists of separate plates.

Idea B: The earth's plates sit on a hot, slightly softened layer of the earth.

Idea C: The plates move very slowly, pressing against one another in some places and pulling apart in other places.

Idea D: When two of earth's plates press against each other, if the plates differ in density, the denser plate will sink beneath the other and the less dense plate may fold upward forming mountains, but if they are about equal in density both plates will fold upward.

Idea E: Melted rock material rises up between plates that are pulling apart, creating new plate material.

Each idea was further clarified in order to state precisely what students would be expected to know and what students would not be assessed on. These clarification statements act as item writing specifications; we strictly follow the boundaries of the stated learning goals to ensure a close alignment between the items and the learning goals. For example, the clarification statement for Idea A says:

Students are expected to know that the solid rock layer that lies under the water, soil, and loose rock on the surface of the earth is divided into massive sections of rock, called "plates." The plates fit closely together such that all of the edges of a plate touch the plate next to it. Students should know that plates are miles thick, and that the solid rock sometimes visible at the surface of the earth is only a very small fraction of the entire plate, in terms of how deep and how wide the plates are. Students should know that today there are about 12-15 very large plates, each of which encompasses large areas of the earth's outer layer (e.g., an entire continent plus adjoining ocean floor, or a large part of an entire ocean basin). Together these plates are large enough to make up most of the entire outer layer of the earth. Students should also know that there are additional smaller plates that make up the rest of the outer layer. Students should know that the continents and ocean basins are part of the plates, and that the exposed solid rock of mountains is an example of plate material, which is visible when it is not covered by water, soil, or loose rock such as sand. Students should also know that the boundaries of continents and ocean basins are not necessarily the same as the boundaries of plates. Some boundaries between plates are found on continents, some on ocean floors, and some in places where oceans and continents meet.

Students are not expected to know the term "bedrock." Students are not expected to know the names of specific plates, the size of the smaller plates, or how many small plates there are. Students are not expected to know that the term "plates" refers to the lithosphere, or to know that two main components (crust and upper mantle) make up the plates. Students are not expected to know that the plates are not uniform in composition or thickness. Students are not expected to know the terms lithosphere, crust, or mantle.

During the development of the assessment items, student misconceptions were incorporated into the distractors. The following is a list of some of the misconceptions that were tested.

- The plates include ocean basins but do not include the continents (Horizon, 2005).
- The plates are only found deep within the earth; no part of a plate can be seen at the earth's surface (Libarkin et al., 2005a,b; Horizon, 2005).
- Plates are under the continents and hold the continents up (AAAS pilot testing, 2006).
- Plates are arranged like a stack of layers in the earth (Marques and Thomson, 1997; Horizon, 2005).
- Plates are made of melted rock (AAAS pilot testing, 2006).
- There are seven plates (because there are seven continents) (AAAS pilot testing, 2006).
- Plates are separated by empty gaps (Libarkin et al., 2005).
- The layer below the plates is magma (i.e., melted rock) (AAAS pilot testing, 2006).
- There is nothing but solid rock below earth's plates (AAAS pilot testing, 2006).
- Continents and ocean basins move, but they do not move with the plates (after Libarkin, 2005).
- Ocean basins do not move (AAAS pilot testing, 2006).
- The rock material below the plates does not move (AAAS pilot testing, 2006).
- Earth's plates do not move (AAAS pilot testing, 2006).
- Plates move, but they will only move a little bit even over millions of years (AAAS pilot testing, 2006).
- Mountains are made of piles of loose rock, not folded plate material (AAAS pilot testing, 2006).

Findings

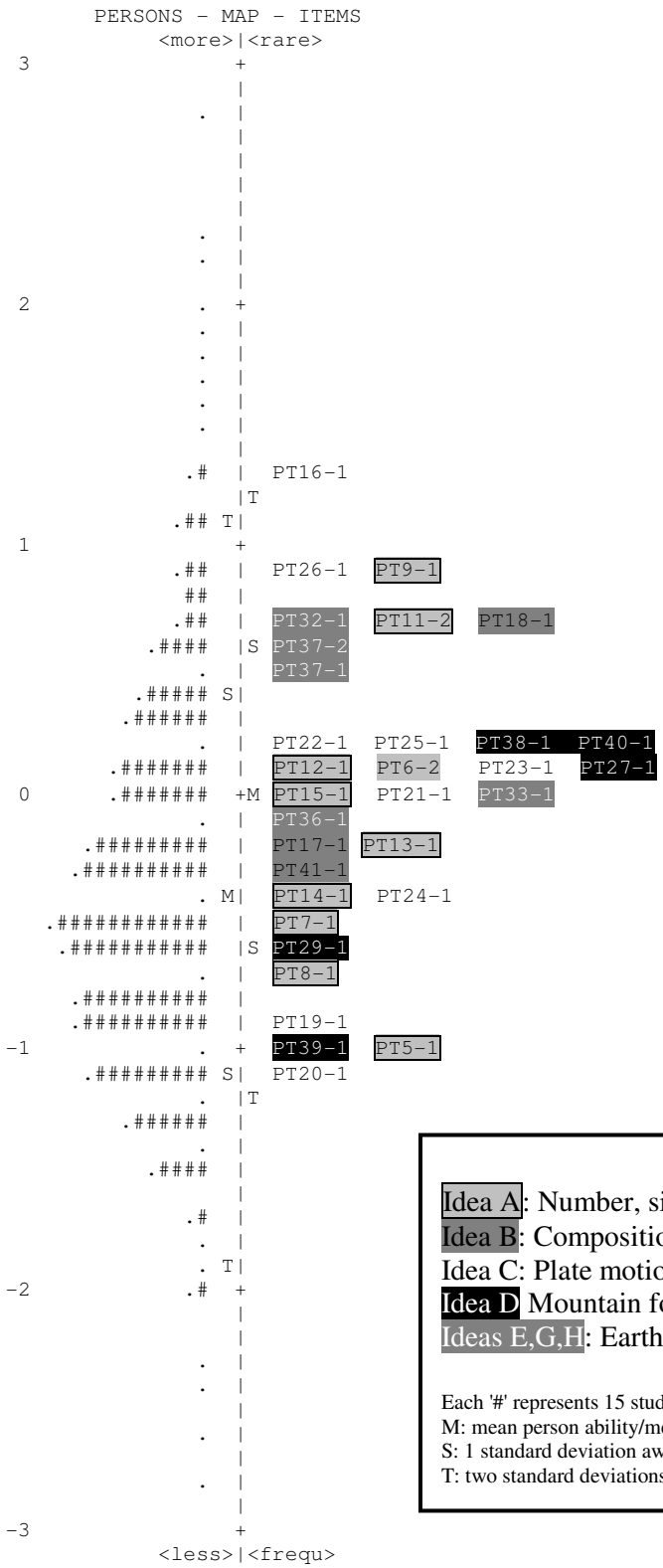
Summary of What Students Know and the Misconceptions They Have. Table 6 shows the average percent correct for sixth, seventh, and eighth grade students for each idea tested. The results indicate that middle school students had the most difficulty with the idea that the earth's plates are directly above a slightly softened layer of rock material (Idea B), and were most successful with the ideas that the plates move very slowly, and that when two of earth's plates press against each other, mountain formation occurs (Ideas C and D).

Table 6
Percent correct by idea for the middle school students in grades six, seven, and eight

	n	Idea A	Idea B	Idea C	Idea D	Idea E	Overall
Number of items		10	3	8	5	6	32
6th grade	676	41%	32%	44%	43%	38%	41%
7th grade	701	44%	33%	46%	48%	37%	43%
8th grade	689	47%	34%	50%	48%	39%	45%
Total % correct	2084	44%	33%	47%	46%	38%	43%

Rasch Difficulty Measure
As can be seen in the item map (Figure 3), the mean difficulty of the items is slightly higher than the mean ability level of the middle school students who were tested. The distribution of item difficulty vs. person ability for the middle school students indicates that there are items to assess the knowledge of the highest ability students, but there are very few items to discriminate effectively among the lower ability students and none to test the very lowest ability students. The item map also shows that items for the different ideas appear at different places on the student ability scale. For some ideas, the items cover the full range of student ability and others concentrate at one end of the spectrum or the other.

Figure 3: Item-person map showing the distribution of student abilities on the left and item difficulties on the right. Where item difficulty and person ability match, the person has a 50% chance of answering the item correctly. Item difficulties are shown for the 32 items included on the field tests.



Key

Idea A: Number, size, and composition of earth's plates
Idea B: Composition of the layer beneath the plates
Idea C: Plate motion
Idea D Mountain formation as plates press together
Ideas E,G,H: Earthquakes, volcanoes, new plate material forming

Each '#' represents 15 students.
M: mean person ability/mean difficulty
S: 1 standard deviation away from the mean
T: two standard deviations away from the mean

Grade-to-Grade Differences. There is a small but statistically significant difference in scores for the Plate Tectonics items when comparing scores by grade ($F=17.143$, $p<0.001$); the eighth grade students (45% correct) outperform all other middle school students ($p<0.01$), and the seventh grade students (43% correct) outperform sixth grade students (41% correct) ($p<0.05$). Although most schools begin teaching plate tectonics in sixth grade, it is not until the end of eighth grade that all students have had instruction in the topic, so the improvement in scores through middle school could be attributed to greater numbers of students having had instruction on the topic by the end of eighth grade.

Summary of What Students Know and the Misconceptions They Have

Answer choice selection on the field tests demonstrated that there are many ideas within this topic that are well understood by middle school students, but there remain many common misconceptions. Selected results from field testing Ideas A, B, C, and D are presented below.

Idea A: Number, size, and composition of earth's plates

For the most part, students knew that the outer layer of the earth is made up of plates (93%), but they had varying ideas about where the plates are, how they fit together, and what they are made of. About 50% of the students knew that plates fit together so that each plate touches all the plates around it, but some students thought that plates do not touch but are separated by oceans (15%) or by melted rock (15%), or that the plates are stacked on top of each other (18%). On another item that targets the same knowledge but uses drawings to depict how plates fit together, 47% of students chose the correct representation, but 26% of students chose a drawing that showed the plates separated by an empty gap that extends to the bottom of the plate, and 25% of students selected drawings that showed the plates stacked on top of each other.

We also found that many students were unfamiliar with the idea that the plates make up the outer solid rock layer of the earth so that they can be seen when they are not covered by soil, loose rock, or water. Only 28% of students recognized that a photograph that shows the solid rock layer that makes up a cliff is showing part of a plate. About 40% of students thought that plates could never be seen because they are always deep within the earth. In responding to another item, 40% of students thought that plates are made of melted rock. Students also showed confusion about the relationship between continents and plates. Although 63% of students knew that plates are composed of continents and ocean basins, when students were shown a photograph of part of a continent (a rock cliff extending into a body of water), 42% of the students said that the continent is part of a plate, but 22% said that the continent is on top of but not part of a plate, and 25% said the continent is on top of a layer of water that is above a plate.

Idea B: Composition of the layer beneath the plates

Most students knew that the layer directly beneath the plates is moving (84%). We also found that 70% of them knew that this layer moves slowly, and 47% knew that it moves slowly in different directions in different places. The composition of this layer, however, was not as well understood. Only 17% of students knew that it is made mostly of slightly softened rock, and 51% thought that the layer is mostly liquid or a combination of liquid and solid rock.

Idea C: Plate motion

As noted above, from their answers to questions under Idea A, we learned that most students knew that continents are part of plates and that fewer students knew that ocean basins are part of plates. This difference in understanding in the relationship between plates and continents, versus plates and ocean basins, is reflected again in Idea C, where we address the movement of plates. For example, 81% of students knew that continents are moving, and 66% of students knew that continents move along with the plate they are a part of, but only 69% of students knew that ocean basins are moving, and only 42% knew that ocean basins move along with the plate they are a part of.

For the most part, students knew that plates move (92%) and that they move several inches per year (62%), but 18% thought they move several feet per year, and 8% thought they don't move at all. However, students had a much harder time using their knowledge about the rate of plate motion to estimate how far plates could move over longer periods of time. Even though 62% of the students knew that plates move several inches per year, and 66% knew that continents move along with the plate they are a part of, when asked about the motion of continents, only 38% of students knew that a continent could move about 10 feet in 100 years. Many students vastly underestimated the distance a plate could move over that time period, with 34% thinking the continent would move only 10 inches. Others (19%) overestimated the distance it could move, thinking that a continent could move as far as 10 miles in 100 years. Similarly, when the question was asked in terms of the motion of plates directly (not the movement of continents), 31% of students recognized that over a million years two plates could become separated by 40 miles. Other students (31%) thought that they could become separated by 40 inches, and 17% thought the distance they moved would be so small it would not even be measurable.

Students also had difficulty with the idea that the plates move along with the layer beneath the plates. We found that only 24% of students knew that the plates and the rock material directly beneath the plates move together, 28% thought that the layer beneath the plates does not move at all, and 35% of students thought that the plates and the layer directly beneath the plates move separately.

Idea D: Mountain formation as plates press against each other

We learned from our analysis of student responses to items testing Idea D that students seemed to know more about what causes mountain building to occur than what mountain building is. For example, only 38% of students knew that when two plates push together and one plate crumples upward, the plate is bending and folding to form a mountain, whereas many students thought that the plate is breaking into pieces of rock the size of tennis balls (21%) or huge boulders (26%). These misconceptions are consistent with the incorrect idea that mountains are piles of loose rock, not solid rock that is continuous with earth's plates. We also found that the majority of students (63%) knew that mountains have been developing continuously throughout the history of the earth.

Discussion of What Students Know. The prevalence among middle school students of strongly held misconceptions about the nature of plates (Idea A) and where they are located suggests that despite high performance on several items targeting knowledge about the movement of plates (Idea C) and the results of plate motion (Ideas D and E), many of the students' models of plate activity are built upon incorrect ideas of what the plates are. For example, we know from results of testing not presented here that 72% of students know that earthquakes can occur where two plates are pulling away from each other, but we also noted earlier that 40% of students thought that the plates are made of melted rock. Furthermore, 76% of students knew that mountains form where plates push together, yet 66% of students said that plates are always deep within the earth and are never seen on earth's surface. It is difficult to imagine what mental model the students have that explains earthquakes in terms of plates made of melted rock pulling away from each other, or mountains being formed by plates that are deep within the earth pressing together. It seems that students are learning bits and pieces of knowledge without making the connections between them. They have mental models for the composition and location of plates but do not realize the inadequacy of their models for explaining natural phenomena.

Student performance on items that target the idea that plates move was very high. In fact, only about 10% of students thought that plates do not move. However, despite the fact that most students knew that plates move an inch or two per year, they struggled with the idea that plates can move great distances over long time periods. When asked to apply their knowledge to predict how far a continent would move in 100 years, given the options of 10 inches, 10 feet, or 10 miles, 34% of students thought that the continent would move only 10 inches in 100 years. The mathematics needed to know that an inch or two per year times 100 years is greater than 10 inches is certainly within the grasp of most middle school students.

One possible explanation for their failure to better estimate how far a continent would travel over time is that some of them think that continents are above, but not part of plates. Therefore, they may think that continents move more slowly than the plates move. But on another item, which asks about the movement of mountains on plates and clearly defines the mountains as being part of two separate plates, students still have problems estimating the distance they would move. The item asks students to estimate how much the distance between the two mountains would increase after 1 million years if they moved at the same rate as plates are moving today. Only two quantities are given as options, 40 miles and 40 inches. The other two options are that the plates move an immeasurable amount and that the distance would not change at all. Despite the fact that 40 inches vastly underestimates the distance plates could move apart over a million years if they are both moving several inches a year, 31% of students chose that answer instead of the correct answer that the distance between the plates would increase by about 40 miles (which was selected by 40% of the students). Apparently some students have a hard time believing that plates can actually move as far as miles apart, even over a very long time period.

Differential Item Functioning. As part of Rasch modeling, we conducted a differential item function (DIF) analysis of each of the items. We examined the items for differential functioning by gender and whether or not English is the student's primary language. We found only a small number of significant results ($p<0.05$) in which one group either under or over performed on that item with respect to their expected score, and we were unable to detect patterns in the results that would cause us to modify any of the items.

For example, we found that girls scored better than expected on an item that asks whether volcanic eruptions or earthquakes can occur where two of earth's plates are pulling apart. Boys exceeded their expected performance on three items. One item tests the idea that a cliff is part of a plate that is exposed at the surface of the earth, another tests how fast a plate moves in a year, and the third tests the idea that continents move along with plates that they are a part of.

When we looked at the performance of students for whom English was not their primary language, we found that they performed better than expected on three items. One shows a beach cliff and asks if any part of a plate is visible in the photograph, one asks how the plates and the layer directly beneath the plates move relative to each other, and a third item targets knowledge that the layer beneath the plates is made of slightly softened rock and that this layer moves with the plates. Students whose primary language was English performed as expected on all items.

Responses From College Students

The 17 college students performed much better on the test as a whole (73% correct), and for each key idea, than the middle school students (Table 7).

Table 7
Percent correct by idea for the college students

	Idea A	Idea B	Idea C	Idea D	Idea E	Overall
Number of items	10	3	7	5	5	30
College students	82%	57%	71%	85%	70%	73%

A closer look at the few items on which college students had low scores helped us determine if these scores reflect flaws in the item or if they indicate persistent misconceptions about fundamental ideas for Plate Tectonics that continue to be held by advanced college students.

One idea a number of the college students had trouble with is that the solid rock layer that we sometimes see protruding through the ground can be the top part of a plate. In the college student sample, 24% said that no part of a plate can ever be visible at the surface of the earth, and 35% said that parts of plates can be visible at the surface of the earth but that the solid rock of a cliff is not an example of part of a plate. Responding to a different item, 38% of the college students again answered that plates are always deep within the earth and can never be seen at the earth's surface. Because a significant number of the college students selected this distractor in the two different items, and because it is unclear how those distractors could be considered correct under any circumstance, we are confident that the items are not flawed but rather that these results indicate that these college students are unaware of the fact that mountains, cliffs, and other outcroppings of rock are continuous with earth's plates.

In another example, however, a large number of college students selected an incorrect answer choice, and on closer examination we decided that the item may not be a fair measure of what students know. We found that a surprising number of college students (75%) chose a distractor that suggests they have the misconception that there is empty space between plates. However, in responding to another item, only 18% of the college students chose distractors that showed pictures of empty gaps forming between two plates that are moving apart. This inconsistency suggests that there could be a problem with the phrasing of the first item. The item asks how you could tell where plate boundaries are. One of the distractors says that you could identify the location of plate boundaries "by using the location of mountains and empty spaces. The location of the mountains shows where two plates push together and empty spaces form between two plates that are moving apart." It is possible that these students are reading "empty space" to mean "rift valley," which would make that answer choice correct. We will follow up with the college students to find out exactly how they were interpreting the language of this item.

Similarly, both the college students and the middle school students had low scores on an item about the plates moving along with the layer beneath the plates. 47% of the college students chose a distractor that said that the plates and the layer beneath the plates move separately. Because the layer beneath the plates moves through convecting cells, only the upper part of this layer moves along with the plates and the rest of this layer does, in fact, move separately from the plates. Although we do not expect middle school students to have this level of knowledge of the process of plate motion, we will have to re-examine this item and make changes to it.

Finally, the performance of the college students on questions about the rate of motion of the plates mirrors the pattern of the responses from the middle school students. Every college student knew how fast plates move over a year, but only 24% knew that the plates could move about 10 feet over 100 years, with 76% of students thinking that plates could move only about 10 inches over 100 years. On the other hand, 71% of the college students knew that the plates could move forty miles over a million years, compared with only 40% of the middle school students. The college students seem to have a similar problem as the middle school students in accepting that plates can move significant distances over moderate time periods, despite clearly knowing the rate that the plates are moving. But the college students are more comfortable with the idea that plates can move very long distances over very long time periods.

Topic 3: Control of Variables

Description of the Sample Tested. Field testing was conducted on a sample of 2830 sixth, seventh, and eighth grade students from around the country. Approximately 51% of the sample was male and 49% female. About 55% of the students self-identified as White, and about 45% self-identified as students of color, including Black, Hispanic, Asian, Pacific Islander, Native American, or various combinations of those, including White. Approximately 85% said that their primary language was English and 15% said their primary language was something other than English.

Description of the Target Learning Goals. Our approach to writing multiple choice questions begins with a very clear definition of the knowledge we are testing. For controlling variables (CV), we started with the following statement from *Benchmarks for Science Literacy* (AAAS, 1993).

> *If more than one variable changes at the same time in an experiment, the outcome of the experiment may not be clearly attributable to any one of the variables (1B/M2a)* (p. 12).

We then elaborate the original statement into a set of supporting and corollary statements that define the boundaries of the expectations we have for students in middle school. The following clarification statement for controlling variables defines the range of ideas we are testing and, therefore, the questions that are acceptable to ask students:

> Students should know that by varying more than one variable at a time it is not possible to determine the relationship between either variable and the outcome of an experiment. Students should also know that by changing one variable at a time and holding all other relevant variables constant it is possible to determine whether that one variable is related to the outcome or not. Students are not expected to know which variables out of all possible variables could be related to the outcome. Nor are students expected to know that it may not be possible to control or even identify all relevant variables. These ideas are included in Benchmark 1B/M2b and 1B/H3. However, when given the set of variables to take into account, students are expected to know that in order to determine if there is a relationship between a particular variable and an outcome, all other variables in the set must remain constant.
>
> Students should know that the reason for controlling a particular variable (holding it constant) in an experiment is because it may have an effect on what is being tested. Students are expected to apply this knowledge in contexts that involve individual events or objects as well as groups of people, events, or objects. Students are not expected to know when they can or cannot generalize beyond the given experimental and control groups.
>
> Students should know that a variable is an entity that may assume different values, either quantitative or qualitative. Students should know that "variable" does not refer to a particular value of a variable. For example, they should know that when "types of liquid" is defined as a variable, water and juice are "values" of the variable and not variables themselves. Students should know what a controlled experiment is, and they should know what a control group and an experimental group are in a controlled experiment. Students are not expected to know the terms "independent" and "dependent" variables.

For our work on controlling variables, we have designed four types of assessment items to examine students' understanding in slightly different ways. The four types are:

I. Given an idea to be tested (hypothesis) and an experimental setup, explain why certain variables are (or should be) kept constant.
II. Select an experimental setup to test the effect of a variable on the experimental outcome, when all relevant variables are provided.
III. Identify the variable(s) being tested in a given controlled experimental setup.
IV. Given an experiment with two variables changing at the same time, determine that no conclusion can be drawn regarding the effect of each individual variable.

For each of the four types, we developed assessment items in different contexts (physical science, life science, and consumer products). The contexts are ones that the students should be familiar with from earlier grades (evaporation, sinking and floating, and the effects of environmental changes on the behavior of living organisms) or from everyday living (fuel efficiency, comparing the effectiveness laundry detergents, etc.). Although students were almost certainly familiar with these contexts, they were

not necessarily contexts that had been used during instruction on CV. In other words, the assessment tasks were not designed around any particular curriculum material and should be considered to be remote transfer tasks for most students (Chen & Klahr, 1999).

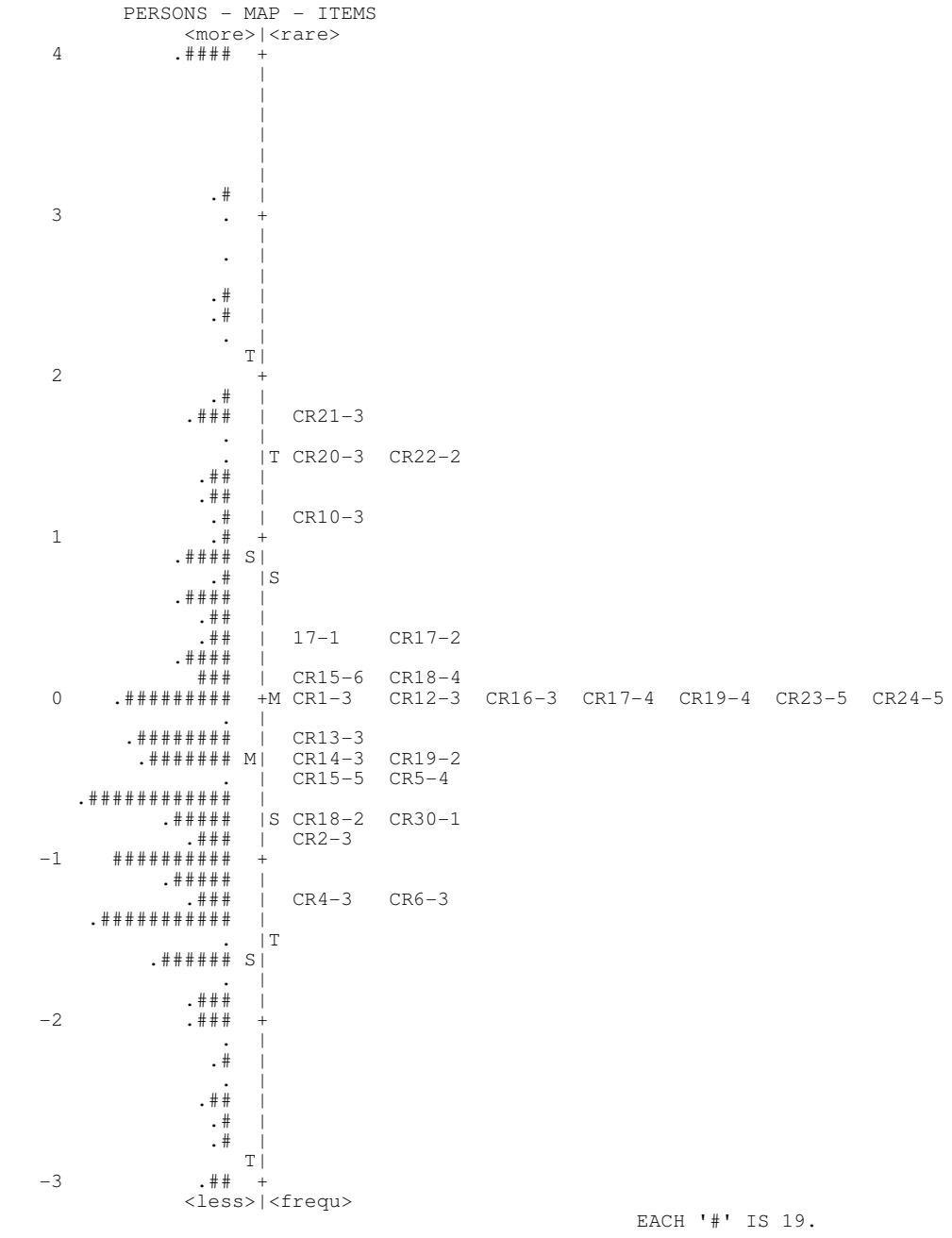
We imbedded various misconceptions regarding CV into the distractors. The list is included below, but can be summarized into three main misconceptions: (1) The first is the idea that when an experiment is being conducted, you can learn something about everything that is included in the experiment. The idea is that scientists would not include something in an experiment if they did not expect to learn something about it, so if they measure the temperature or height of something, they do so because they hope to learn something about the effect of that variable on the outcome. The students miss the idea that a variable can be included as a control variable. (2) Related to this is the idea that when an experiment is performed with two variables changing at the same time (a confounded study), you can learn something about the effect of each individual variable. (3) Finally, some students know that something has to be controlled and something has to vary, but they reverse which is which. The list of misconceptions that capture these themes, as they appear in the research literature or as they were identified in our own pilot testing, are listed below:

- In order to determine whether a certain variable has an effect on the outcome, one must vary ALL variables at the same time (Tschirgi 1980).
- In order to determine whether a certain variable has an effect on the outcome, one must keep this variable constant and change other related variables. (Tschirgi 1980, Zimmerman & Glaser 2001)
- A given experiment tests for the effects of ALL related variables, regardless of whether they are allowed to vary or are held constant (Project 2061 student data).
- When testing the effect of a variable on the outcome of the experiment, it does not matter if other relevant variables change at the same time (Project 2061 student data)
- If two variables change at the same time, one can learn about the effect of each variable on the outcome (Project 2061 student data).
- When asked what an experiment is testing, students may respond based on their prior experiences and ideas regarding the particular context, not based on the experiment presented to them (Project 2061 student data).
- A given experiment tests for the effect of a variable that remains constant while other variables change (Project 2061 student data).
- When testing for the effect of a particular variable on an outcome, it is important for that variable to vary, but only one other relevant variable has to be controlled and the rest can vary (Project 2061 student data).
- A given experiment tests for the effect of more than one variable that remains constant while other variables change (Project 2061 student data).
- If two variables change at the same time, one can learn about the effect of at least one of the variables on the outcome (Project 2061 student data).

Findings

Rasch Modeling. The item reliability of the set of items was .99 (separation index, 11.84) and the person reliability was .67 (separation index, 1.41).The range of point-measure correlation coefficients for the set of items was between .42 and .56, suggesting that the items are measuring a single factor. The person-item map shows that the items adequately measure the middle range of students who were sampled, and the ability of the tested students is just slightly below the mean difficulty of the test items. The map also shows that there are no items that test either the top range of student ability or the bottom range. Because there were not enough items to reliably discriminate between the performance of either the high or low performing students, the person reliability is only marginally acceptable.

Figure 4: Item-person map showing the distribution of student abilities on the left and item difficulties on the right. Where item difficulty and person ability match, the person has a 50% chance of answering the item correctly. Item difficulties are shown for the 25 items included on the field tests.



Summary of what Students Know and the Misconceptions They Have

Although the same basic idea about controlling variables was tested in each of the four types of questions, some differences appeared in student responses, depending on the type of situation that was presented to students in the question. Students were most successful with Type II questions (50.9%), in which they had to select a correct experimental setup. They were also successful with Type I questions (48.1%), in which they were given an idea to be tested (hypothesis) and an experimental setup, and they had to say

why certain variables were (or should be) kept constant. They were least successful with Type IV questions (21.0%) in which they were given an experiment with two variables changing at the same time and had to determine that no conclusion could be drawn regarding the effect of each individual variable.

Table 8
Overall percent correct by item type

Item Type	Type I	Type II	Type III	Type IV
% Correct	48.1%	50.9%	44.2%	21.0%

We also found that seventh and eight grade students performed somewhat better than 6th grade students on the CV items.

Table 9
Overall percent correct by grade

Grade	6th Grade	7th Grade	8th Grade
% correct (N)	40.9% (934)	45.3% (1028)	43.4% (868)

The most common misconceptions that students had appear below. The percentage in parentheses represents how often students chose answers containing that misconception divided by the total number of opportunities students had to chose that misconception.

- If two variables change at the same time, one can learn about the effect of both variables on the outcome (29.8%).
- If two variables change at the same time, one can learn about the effect of at least one of the variables on the outcome (10.3%).
- A given experiment tests for the effects of ALL related variables, regardless of whether they are allowed to vary or are held constant (14.2%).
- In order to determine whether a certain variable has an effect on an outcome, one must keep this variable constant and change other related variables (9.9%).

The main problem students had was with the idea that if two variables change at the same time, no conclusion can be drawn regarding the effect of each individual variable. Their misunderstanding is represented in the first two misconceptions listed above, and, because both of these appeared as answer choices in each question of this type, the results can be added together and show that 40.1% of the students believed that you can learn something from at least one of the variables in a confounded study. Related to that misconception is the idea that you can learn something about *all* the variables in a study, whether they are held constant or allowed to vary. Apparently, in the students' minds, if a variable was included in a study, the researcher must have wanted to learn something about it, and, therefore, something *would* be learned about it. This idea was held by 14.2% of the students. Only 9.9% of the students held the idea that the variable that is to be tested should be held constant and all other variables allowed to vary, the opposite of what is actually the case for a controlled study.

Differential Item Functioning (DIF)
A differential item function analysis was conducted to determine if the items performed similarly for males and females, and if the items performed similarly for students who said that English was their primary language and those who said English was not their primary language.

The results for gender showed significant under or over performance compared to expectations on two items. Boys significantly over performed on an item involving carts and ramps. Girls significantly under performed compared to expectations on the same item. The reverse result was found on the second part of a two-part question on which the students had to choose the answer that explained why they had selected the answer choice that they did on the first part of the question. The context of the item was an experiment dealing with factors affecting the behavior of fish in a fish bowl. Both boys and girls performed as expected on the first part of the item, but they differed on the portion of the item that asked them to explain their answer choice. But on two other two-part items of that type, those gender differences were not observed.

The DIF results for language showed significant under or over performance for students who said that English was not their primary language on three items. There was no significant over or under performance on any items for students who said English was their primary language. On both of the items on which the non-English students over performed, an illustration was provided. On the one item on which they under performed, no visual was present. This alone is not enough to explain the observation because there were many other items on which they performed as expected, regardless of whether an illustration was provided or not. However, what makes this interesting is that two of these items can be thought of as a pair, one with an illustration and one without. For both of them, the context was a farmer growing crops and wanting to find out what factors affected the growth of the crops. On the item on which the non-English students did better than expected, the context was supported by an illustration of the crops with labels for soil type and fertilizer type. On the item on which the non-English students did worse than expected, the item did not have an illustration to support it. It may be that in selected cases, a well placed illustration is particularly helpful to students whose first language is not English.

Responses from College Students

We tested 1134 college students at a major university in a southeastern state on 12 of the 24 control of variables test items. We found that the college students performed much higher on all four types of items and on all individual items than the middle school students did. As with the middle school students, they scored lowest on the Type IV items, which presented them with a confounded study and asked what they could learn from it. However, the performance of the college students did not reveal any anomalies in the data that might be explained by structural problems with the items.

Table 10
Overall percent correct by item type for middle school students and college students

Item Type	Type I	Type II	Type III	Type IV
% Correct for Middle School Students	48.1%	50.9%	44.2%	21.0%
% Correct for College Students	90.3%	88.0%	95.0%	82.7%

Topic 4: Matter and Energy in Living Systems

Description of the Sample Tested
Field testing was conducted on a sample of 2967 sixth, seventh, and eighth grade students from around the country. The sample was 49.6% male and 50.4% female. Of the students in the sample, 56.9% self-identified as White, and about 43.1% self-identified as students of color, including Black, Hispanic, Asian, Pacific Islander, Native American, and various combinations of those, including White. The sample included 89.7% students who said that their primary language was English, and approximately 10.3% who said their primary language was something other than English.

Description of the Target Learning Goals
This topic deals with the transformation of matter and energy in living systems, although the ideas we tested focus on only the matter transformations. Matter transformation involves chemical reactions in which carbon atoms are rearranged to make new molecules that serve as building materials for organisms. This topic's key ideas are based on benchmarks from Chapter 5, Section E of *Benchmarks for Science Literacy*, AAAS, 1993.

> Idea A; All organisms need food as a source of molecules that provide chemical energy and building materials.
>
> Idea B: Plants make their own food in the form of sugar molecules from carbon dioxide molecules and water molecules. In the process of making sugar molecules, oxygen molecules are produced as well.
>
> Idea E: Animals use carbon-containing molecules from food to make a variety of other carbon-containing molecules that become part of their body structures.
>
> Idea G: All organisms, including plants and animals, have mechanisms for storing molecules from food for later use.

Each key idea is further clarified to provide very clear specifications for item writing. The clarification for Idea E includes the following statements:

> Students should know that growth of body structures, including repairing and replacing body structures, involves using carbon-containing molecules (carbohydrates, fats, and proteins) from food to make other carbohydrate, fat, and protein molecules that become part of their body structures. Students should know that muscles are made largely of protein molecules, fat tissue is made largely of fat molecules, the skeletons of insects, lobsters, and crabs are made largely of carbohydrate molecules ("Body structures" include any organ, tissue, or part of an organism with which students are likely to be familiar.)
>
> The processes by which carbon-containing molecules from food are used to make the carbohydrates, proteins, and fats that become part of the body structures involve chemical reactions, not the simple addition of substances from food to body structures. The carbohydrates, fats, and proteins that animals eat do not get incorporated into body structures without first going through a chemical reaction.

Carbohydrates, fats, and proteins from food contribute other kinds of atoms to the molecules that make up body structures. Students are not expected to know the identity of these other atoms.

During the development of the assessment items, student misconceptions were incorporated into the distractors. The following is a list of some of the misconceptions that were tested.

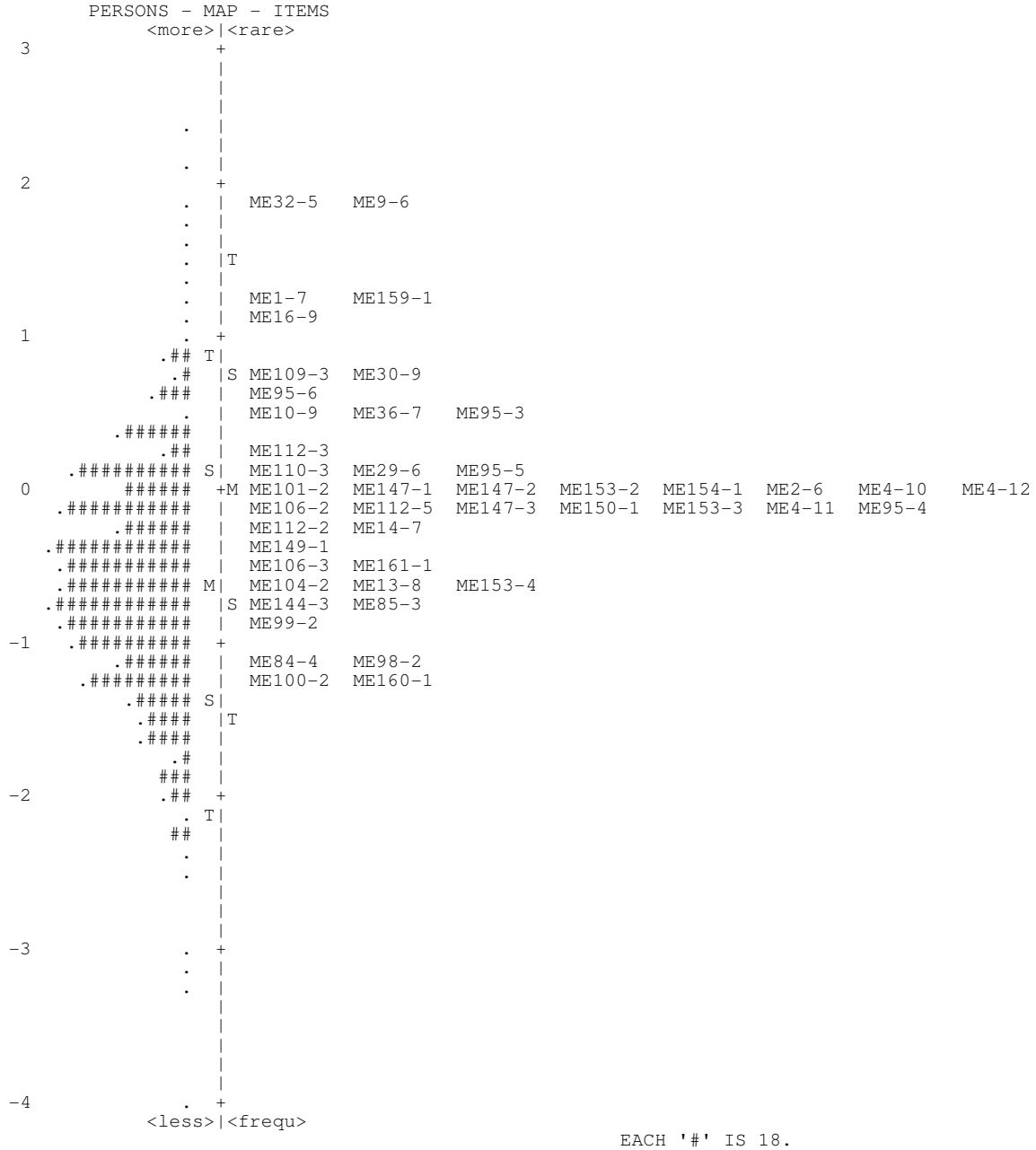
- Many children identify food as any substance that is edible or can be eaten (Lee & Diong, 1999).
- Students see food as any material (water, air, minerals, etc.) that organisms take in from their environment (Anderson et al., 1990; Simpson & Arnold, 1982; Roth & Anderson, 1987).
- Some students think that soil is food for plants (Leach et al., 1992, 1996; Vaz et al., 1997; Wandersee, 1983) or that substances in the soil are food for plants (Kuech et al., 2003; Leach et al., 1992; Simpson & Arnold, 1982; Stavy et al., 1987; Tamir, 1989, Wandersee, 1983).
- Some students think that water is a source of energy (Horizon Research) or is food (Lee & Diong 1999), particularly for plants (Wandersee, 1983; Vaz et al., 1997).
- Although students of all ages identify food as necessary to promote growth and health, many do not recognize that it is the source of material which becomes either part of their bodies in growth and repair or the source of energy (Driver 1994, p. 60).
- Students do not think that food becomes part of the body. Instead food is used for energy and the part that is not used for energy is eliminated as waste (AAAS Project 2061 Pilot testing, 2007).
- Students do not understand that substances have to be changed (undergo chemical reactions) before they become part of an organism's body (AAAS Project 2061 Pilot testing, 2007).
- Students do not recognize that food consists of carbon-containing molecules in which carbon atoms are linked to other carbon atoms (AAAS Project 2061 Pilot testing, 2007).
- Molecules from food that are not used immediately for energy or for building materials are eliminated as waste; they are not stored (AAAS Project 2061 Pilot testing, 2007).

Findings

Rasch Modeling. The item reliability was .99 (separation index, 9.69) and the person reliability was .45 (separation index, 0.90). The person reliability and separation index are unacceptably low, suggesting that this set of items does not effectively discriminate among middle school students in their understanding of this topic. In addition, the range of point-measure correlation coefficients for the set of items was between .03 and .45, suggesting that some of the items with low point-measure correlation coefficients are measuring a specific aspect of the larger idea not captured in the other items, or they are measuring another idea altogether. For a group of college juniors that we tested, the person reliability was .80 (separation index, 2.01), and the range of point measure correlation coefficients was between .14 and .59 and averaged

The person-item map for the middle school students shows that the mean of the item difficulty is considerably above the mean of the student ability. Items adequately measure the middle and upper range of students who were sampled, but there are no items that test the bottom range of student ability.

Figure 5: Item-person map showing the distribution of student abilities on the left and item difficulties on the right. Where item difficulty and person ability match, the person has a 50% chance of answering the item correctly. Item difficulties are shown for the 45 items included on the field tests.



Summary of what Students Know and the Misconceptions They Have
The average percent correct for all of the items was 35.2%. The percent correct was 43.8% for items aligned to Idea A, 31.8% for Idea B, 17.8% for Idea E, and 47.3% for Idea G. The results indicate that middle school students had the least difficulty with items testing the definition of food (Idea A) and food storage (Idea G) and the most difficulty with items testing the link between matter transformation and growth (Idea E).

We examined the extent to which students held various misconceptions, including the following misconception for Idea E:

> Food does not become part of the body. Instead food is used for energy, and the part not used for energy is eliminated as waste (AAAS Project 2061 Pilot Testing, 2007)

On the field test, two items provided students with the opportunity to select this misconception as a distractor. In one item, students were asked what happens to the food that an animal eats as the animal grows, and in the other item they were asked what happened to the grass that a cow eats as it grows. For these two items, the distractor was chosen 60% of the time. Only 9% of students selected the correct response—that some of the food (or grass) is changed into new substances that become part of the animal's (or cow's) body.

Differential Item Functioning (DIF)
A differential item function analysis was conducted to determine if the items performed similarly for males and females, and if the items performed similarly for students who said that English was their primary language and those who said English was not their primary language.

The results for gender showed significant under or over performance compared to expectations on six of the 45 items. Four of the items tested ideas about food, and two of them tested ideas about storage of food. On three of the items testing ideas about what food is, the boys under performed compared to expectations, and the girls over performed on those same three items. On the fourth item, boys over performed compared to expectations, and the girls under performed. On the item on which the boys over performed compared to expectations, the question was about food for plants. The others addressed the question of food for animals. On all other items addressing food for plants or food for animals, boys and girls performed as expected. Overall, there was no pattern that appeared in the DIF data, and, on close inspection, we could identify nothing in any of the items that would bias the items toward boys or girls.

The results for language showed significant under or over performance compared to expectations on three of the 45 items. On all of the items, the students who said that English was their primary language performed as expected. On two of the three items, students who said English was not their primary language over performed compared to expectations, and on one of the three items they under performed. They over performed on an item that had lengthy answer choices but also had an illustration to describe the context, and they over performed on an item that asked them what happened to the food that an animal eats as the animal grows. That item did not include an illustration. But on another item asking whether plants and animals store molecules from food for later use, the non-English students under performed compared to expectations. On all of the other 42 items, students who said English was their primary language and those who said it was not performed as expected. As with the gender DIF analysis, we were unable to identify any features of any of the items that would bias them toward one group or the other.

Results of Testing College Students
We tested three groups of college students at a major urban comprehensive university in the Northeast. This included 284 students enrolled in a chemistry course who took the test during the first week of class, prior to any college-level instruction. These students had taken high school biology. We also administered the test to 84 junior chemistry majors, and 30 biochemistry graduate students. There were

three items in particular that yielded results that made us take a closer look at the items to see if there was anything in the way they were written that would yield false negative results.

One item asked students what happens to the food that an animal eats as the animal grows. The correct answer is that "some of the food is changed into new substances that become part of the animal's body." The answer choice says nothing about energy or waste. On this item, only 7% of the entering college students answered correctly, 15% of the junior chemistry majors answered correctly, and 48% of the graduate students answered correctly. The most popular answer choice for these groups of students was that some of the food that an animal eats as it grows is changed into energy and the rest leaves the body as waste. As we look more closely at the item, it may be that because the correct answer said nothing about energy, this may have suggested to students that we meant that food is used to make new substances alone and is not used for energy. Although the correct answer is correct as written, and the most popular answer choice among the students is incorrect as written, adding an item that includes the word "energy" along with "new substances" would most likely yield more correct answers.

The other two items on which college students did poorly dealt with the source of the mass of wood that comes from a tree, and what plants use for food. The problem for students in both of these items is that they have learned and they believe that minerals in the soil are a source of food for plants. On the "What do plants use for food" item, 21% of entering college students answered correctly, 21% of junior chemistry majors answered correctly, and 19% of biochemistry graduate students answered correctly. The overwhelmingly most popular answer choice among all of these groups was that "Plants use both sugars that they make and substances that they take in from the soil as food." Between 62% and 74% of the students chose that as the correct answer. As long as we insist that food includes only those things that can be used both as a source of chemical energy and as a source of building materials, there seems little we can do to modify this item. The idea that minerals in the soil are food for plants seems to be just something that persists as a misconception even among college students.

Using Item Clusters to Validate Hypothesized Relationships among Key Ideas

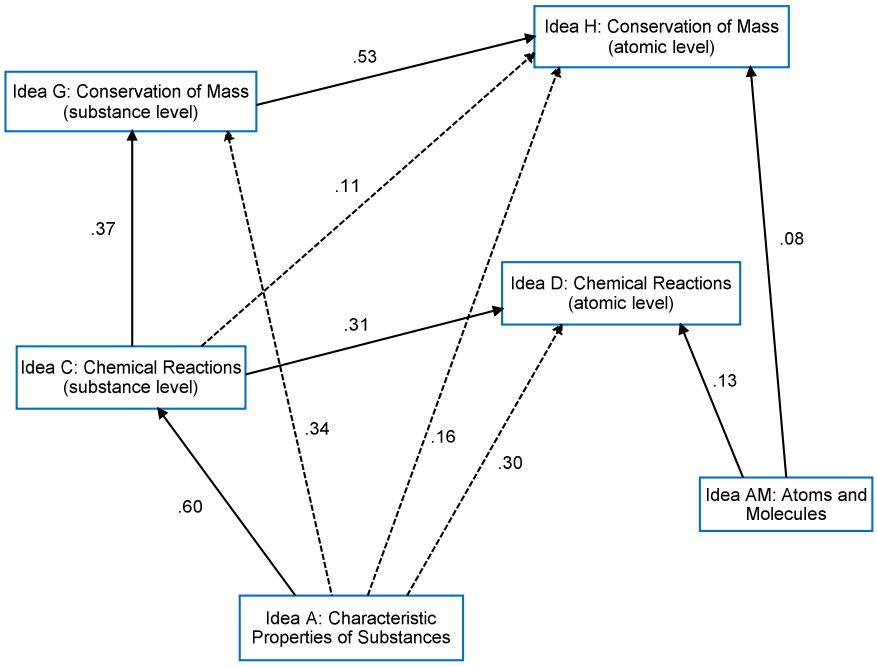
Finally, in our work, we also develop assessment maps to show the relationship among ideas. The assessment maps help us keep track of prerequisite ideas so that we can make judgments about whether it is reasonable to assume that students already have those ideas or if those prerequisite ideas should be separately tested along with the other ideas in that topic. For example, even though students may not know that it is oxygen that is breathed in and that it is carbon dioxide that is breathed out of the lungs, or that gas exchange takes place at the lining of the lungs, we do assume that middle school students know that the lungs are used in breathing. As we test students' understanding of the more sophisticated ideas, our items are written based on the assumption that the students know that the lungs are used in breathing. In the discussion that follows, we show how we used path analysis to test the hypothesized relationships among the five key ideas for the Substances, Chemical Reactions, and Conservation topic and a cluster of four items from the Atoms, Molecules, and States of Matter topic (we will call this Idea AM).

In our hypothesized model, Ideas AM, A, C, and G are precursors (either directly or indirectly) to Idea H; Ideas A and C are precursors to Idea G; Ideas AM, A, and C are precursors to Idea D; and Idea A is a precursor to Idea C. The model was tested twice using two versions of the field test. Each version had 30 items that covered all six ideas.

A series of linear regression analyses were performed, and standardized coefficients (beta weights) were calculated with Ideas H, G, D, and C successively used as the dependent variable. Figure 6 shows the hypothesized path along with the beta weights using Form B. The hypothesized model fit the data well, with R square values of .98 for Form A and .99 for Form B. Significant beta weights (at the 0.001 level

of significance) were observed for all hypothesized relationships in the model. Beta weights ranged from .07 to .59 for Form A and from .08 to .60 for Form B. The largest beta weights for both forms occurred for the relationship between Idea A (Characteristic properties of substances) and Idea C (Chemical reactions on the substance level). The smallest beta weights for both forms occurred for the relationship between Idea AM (Atoms and molecules) and Idea H (Conservation of mass on the atomic level).

Figure 6: Path analysis of the chemistry ideas included on the field tests (Beta weights calculated using Form B are shown next to the arrows)



The results are not surprising, both where the relationships are strong and where they are weak. It is not surprising, for example, that there would be a very strong path coefficient (.60) between Idea A and Idea C. Idea C requires that students use knowledge of the properties of substances to make decisions about whether or not a chemical reaction has occurred. They decide if a chemical reaction has occurred on the basis of whether the properties of the products are different than the properties of the reactants. Idea A is about the fact that substances have characteristic properties that can be used to identify them. It is highly unlikely that a student could answer questions about Idea C without having knowledge of Idea A. A weaker link (.11) between Idea C and Idea H is also understandable. Student understanding of chemical reactions at the substance level (that the properties of the products of a reaction are different than the properties of the reactants) is not needed to answer questions about the conservation of mass at the atomic level.

These analyses are proving to be a promising way for us to test the strength of the relationships between various ideas for the different topics that we are studying.

Acknowledgements

This work is funded by the National Science Foundation (Grants # ESI 0227557 and ESI 0352473).

In addition to the authors of this paper, Dr. Natalie Dubois, Dr. Kristen Lennon, Dr. Arhonda Gogos, and Dr. Paula Wilson also contributed to the work.

References

Abraham, M. R., Williamson, V. M., & Westbrook, S. L. (1994). A cross age study of the understanding of five chemistry concepts. *Journal of Research in Science Teaching*, 31(2), 147-165.

Ahtee, M., & Varjola, I. (1998). Students' understanding of chemical reaction. I*nternational Journal of Science Education*, 20(3), 305-316.

Anderson, C. W., Sheldon, T. H., & DuBay, J. (1990). The effects of instruction on college nonmajors' conceptions of respiration and photosynthesis. *Journal of Research in Science Teaching, 27*(8), 761-776.

Andersson, B. R. (1986). Pupils' explanations of some aspects of chemical reactions. *Science Education*, 70(5), 549-563.

Berkheimer, G. D., Anderson, C. W., Lee, O., & Blaskeslee, T. D. (1988). Matter and Molecules Teacher's Guide: Science Book. East Lansing, Michigan: Michigan State University.

BouJaoude, S. B. (1992). The relationship between students' learning strategies and the change in their misunderstandings during a high school chemistry course. *Journal of Research in Science Teaching,* 29(7), 687-699.

Calik, M., & Ayas, A. (2005). A comparison of level of understanding of eighth-grade students and science student teachers related to selected chemistry concepts. *Journal of Research in Science Teaching*, 42(6), 638-667.

Cavallo, A. M. L., McNeely, J. C., & Marek, E. A. (2003). Eliciting students' understandings of chemical reactions using two forms of essay questions during a learning cycle. *International Journal of Science Education*, 25(5), 583-603.

Chen, Z. & Klahr, D. (1999). All other things being equal: Acquisition and transfer of control of variables strategy. *Child Development*, 70, 1098-1120.

DeBoer, G.E., Herrmann Abell, C.F., and Gogos, A., (2007, March-April). *Assessment Linked to Science Learning Goals: Probing Student Thinking During Item Development*. Paper presented at the National Association for Research in Science Teaching Annual Conference, New Orleans, LA.

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

DeBoer, G.E., Lee, H.S., & Husic, F. (2008b). Assessing integrated understanding of science. In Y. Kali, M.C. Linn, & J.E. Roseman, (Eds.), *Coherent Science Education: Implications for Curriculum, Instruction, and Policy* (pp. 153-182). New York, NY: Columbia University Teachers College Press.

Driver, R., Squires, A., Rushworth, P. and Wood-Robinson, V. (1994). *Making sense of secondary science: Research into children's ideas*. New York, NY: Routledge.

Eilks, I., Moellering, J., & Valanides, N. (2007). Seventh-grade students' understanding of chemical reactions: Reflections from an action research interview study. *Eurasia Journal of Mathematics, Science & Technology Education*, 3(4), 271-286.

Hall, J. R. (1973). Conservation concepts in elementary chemistry. *Journal of Research in Science Teaching*, 10(2), 143-146.

Hesse, J. J., & Anderson, C. W. (1992). Students' conceptions of chemical change. *Journal of Research in Science Teaching*, 29(3), 277-299.

Horizon Research, Inc. (2005). [misconceptions developed for distracters for multiple-choice Plate Tectonics items]. Unpublished data.

Kuech, R., Zogg, G., Zeeman, S. & Johnson, M. (2003) Technology rich biology labs: effects of misconceptions. *Annual Meeting of the National Association from Research in Science Teaching* (Philadelphia, PA)

Leach, J., Driver, R., Scott, P., & Wood-Robinson, C. (1992). *Progression in understanding of ecology concepts by pupils aged 5 to 16*. Leeds: Children's Learning in Science Research Group, Centre for Studies in Science and Mathematics Education, University of Leeds.

Lee, O., Eichinger, D. C., Anderson, C. W., Berkheimer, G. D., & Blaskeslee, T. D. (1993). Changing Middle School Students' Conceptions of Matter and Molecules. *Journal of Research in Science Teaching*, 30(3), 249-270.

Lee, Y. J., & Diong, C. H. (1999). Misconceptions in the biological concept of food: results of a survey of high school students. In M. Waas (Ed.), *Enhancing Learning: Challenge of Integrating Thinking and Information Technology into the Curriculum* (pp. 825-832). Singapore: Education Research Association.

Libarkin, J. C. and Anderson, S. W. (2005a). Assessment of learning in entry-level geoscience courses: results from the Geoscience Concept Inventory. *Journal of Geoscience Education* 53 (4), 349-401.

Libarkin, J. C., Anderson, S. W., Dahl, J., Beilfuss, M., & Boone, W. (2005b). Qualitative analysis of college students' ideas about the Earth: interviews and open-ended questionnaires. *Journal of Geoscience Education*, 53, 17-26.

Marques, L., and Thomson, D. (1997). Misconceptions and conceptual change concerning continental drift and plate tectonics among Portuguese students aged 16-17. *Research in Science & Technological Education*, 15, 195-222.

Mas, C. J., Perez, J. H., & Harris, H. (1987). Parallels between adolescents' conception of gases and the history of chemistry. *Journal of Chemical Education*, 64(7), 616-618.

Mitchell, I., & Gunstone, R. (1984). Some student conceptions brought to the study of stoichiometry. *Research in Science Education*, 14, 78-88.

Novak, J. D., & Musonda, D. (1991). A twelve-Year Longitudinal Study of Science Concept Learning. *American Educational Research Journal*, 28(1), 117-153.

Ozmen, H., & Ayas, A. (2003). Students' difficulties in understanding of the conservation of matter in open and closed-system chemical reactions. *Chemistry Education Research and Practice*, 4(3), 279-290.

Roth, K. J., & Anderson, C. W. (1987). *The power plant teacher's guide. (Occasional Paper No. 112).* East Lansing, MI: Institute for Research on Teaching, Michigan State University.

Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching,* 35(3), 265-296.

Simpson, M., Arnold, B. (1982) The inappropriate use of subsumers in biology learning. *European Journal of Science Education* 4:173-182.

Smith, C., Wiser, M., Anderson, C. W., Krajcik, J., & Coppola, B. (2004). *Implications of Research on Children's Learning for Assessment: Matter and Atomic Molecular Theory.* Center for Education (NRC).

Solomonidou, C., & Stavridou, H. (2000). From inert object to chemical substance: students' initial conceptions and conceptual development during an introductory experimental chemistry sequence. *Science Education,* 84, 382-400.

Stavridou, H., & Solomonidou, C. (1998). Conceptual reorganization and the construction of the chemical reaction concept during secondary education. *International Journal of Science Education*, 20(2), 205-221.

Stavy, R., Eisen, Y., and Yaakobi, D. (1987). How students aged 13-15 understand photosynthesis. *International Journal of Science Education*, 9 (1), 105-115.

Stavy, R. (1990). Children's Conceptions of Changes in the State of Matter: From Liquid (or Solid) to Gas. *Journal of Research in Science Teaching*, 27(3), 247-266.

Tamir, P. (1989) Some issues related to the use of justifications to multiple-choice answers. *J. Biol. Ed*. 23 (4): 285-292.

Thomaz, M. F., Malaquis, I. M., Valente, M. C., & Antunes, M. J. (1995). An Attempt to Overcome Alternative Conceptions related to heat and temperature. *Physics Education*, 30(1), 19-26.

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development*, 51, 1-10.Valanides, N. (2000). Primary Student teachers' understanding of the particulate nature of matter and its transformations during dissolving. *Chemistry Education Research and Practice*, 1(2), 249-262.

Vaz, A.N., Carola, M.H. & Neto, A.J. (1997) Some contributions for a pedagogical treatment of alternative conceptions in biology: an example from plant nutrition. *Annual Meeting of the National Association for Research in Science Teaching* (Oak Brook, IL).

Wandersee, J. H. (1983). Students' misconceptions about photosynthesis: a cross-age study. Paper presented at the Proceedings of the International Seminar: Misconceptions in Science and Mathematics, Ithaca, NY. June, 1983.

Zimmerman, C., & Glaser, R. (2001). Testing positive versus negative claims: A preliminary investigation of the role of cover story in the assessment of experimental design skills (Tech. Rep. No. 554). Los Angeles, CA: UCLA National Center for Research on Evaluation, Standards, and Student Testing (CRESST).