

## **Assessment Linked to Middle School Science Learning Goals: Using Pilot Testing in Item Development**

**George E. DeBoer, Natalie Dubois, Cari Herrmann Abell  
AAAS Project 2061**

**NSTA National Conference  
Boston, MA**

**March 27–30, 2008**

### **Introduction**

In this paper we discuss a process for developing distractor-driven, multiple-choice test items (Sadler, 1998) that are closely aligned with science content standards. The work is part of a multi-year, NSF-funded project to develop assessment items aligned to middle school science content standards from AAAS's *Benchmarks for Science Literacy* (American Association for the Advancement of Science [AAAS], 1993) and the NRC's *National Science Education Standards* (National Research Council [NRC], 1996). We focus in particular on (1) how feedback obtained from students through pilot testing aids in the development of items that are effectively aligned to content standards and (2) how that feedback provides information on students' thinking about the targeted ideas. The goal of this work is to (1) promote the development of assessment tools and models that can be used by test developers, classroom teachers, and science education researchers to assess student understanding of key ideas in science and identify gaps in students' knowledge that stand in the way of their understanding phenomena in the world around them, and (2) use assessment to encourage instruction that emphasizes students' conceptual understanding of the natural world rather than instruction that asks students to memorize facts, definitions, and abstract principles disconnected from the world around them.

### **Aligning Assessment Items to Learning Goals**

Both *Benchmarks for Science Literacy* (AAAS, 1993) and the *National Science Education Standards* (NRC, 1996) are organized around ideas and skills that all students should learn by the end of various grade bands if they are to achieve the goal of science literacy by the time they graduate from high school. Although these national standards documents present the learning goals as lists of what students should know and be able to do, these learning goals are not simply disconnected facts about science. Rather, they are key elements in a series of conceptually coherent science narratives. In both the NRC's *National Science Education Standards* and AAAS's *Benchmarks for Science Literacy*, the individual ideas are framed by introductory essays that summarize the conceptual story those ideas are part of. The expectation is that students will develop richly interconnected mental models of objects, events, and processes in the natural world that will enable them to explain related events and make predictions about them.

**Key ideas.** The content standards provide useful direction to assessment developers regarding what students should know in science, but these statements also raise additional questions about exactly what students should be held accountable for. To provide further guidance to assessment

developers that will lead to a more precise identification of gaps in student knowledge, we further subdivide the content standards into finer-grained statements of knowledge, or *key ideas*. We then clarify each key idea by indicating what it is that we expect students to know about that idea and what the boundaries of that knowledge are for purposes of assessment. Consider the following key idea about the size of atoms and the clarification statement that follows:

*Key Idea: All atoms are extremely small.*

Students are expected to know that atoms are much smaller than very small things with which they are familiar—such as dust, blood cells, plant cells, and microorganisms. Students should know that the atoms are so small that many millions of them make up these very small things. They should know that this is true for all atoms. Students are not expected to know the actual size of atoms or the relative size of atoms compared to other atoms.

For purposes of assessment, this clarification statement provides guidance for item development by setting boundaries around the expectations for middle school students. Students are expected to know that atoms are much smaller than other small things with which they are familiar, and they are expected to know that atoms are millions of times smaller than those other very small things. Students are not expected to know the actual size of atoms or the relative size of atoms compared to other atoms. Decisions about the expectations that guide assessment development are made on the basis of the value that the idea will have for students in explaining and predicting real-world phenomena, and they are based on the importance of the idea in preparing students for ideas they will encounter later. In the case of the size of atoms, we want students to know, for example, that atoms are much smaller than cells so that they can appreciate the idea that a cell contains many millions of molecules (and atoms) that are involved in the work of the cell. It is widely known that many students, even when they are in college, think that cells are smaller than atoms or molecules (Tretter et al. 2006), making it impossible for them to have a mental image of molecular processes occurring inside of cells. Decisions about expectations are also based on a review of the literature on student learning (see, for example, Driver et al., 1996). Research on student learning provides information about the complexity of the mental models we can expect middle school students to develop, and it provides us with ideas that students find plausible, which we use as distractors in test items. Using misconceptions as answer choices also enables us to conduct studies to systematically validate and update existing research on student learning.

**Alignment.** After clarifying our expectations for students, we use two criteria to determine whether an assessment item is aligned to the content specified in a particular targeted idea. The necessity criterion addresses whether the knowledge specified in the idea is needed to successfully complete the task, and the sufficiency criterion addresses whether the knowledge in the targeted idea is enough by itself to successfully complete the task or if some additional knowledge is needed (Stern & Ahlgren, 2002; DeBoer, 2005). If the targeted knowledge is not needed to answer the question, then the item is obviously not a good indicator of whether students know that targeted idea. And if additional knowledge is needed to answer correctly, it is difficult to know if an incorrect response is due to not knowing the targeted idea or not knowing the additional idea. The purpose of such careful alignment is to help reduce errors in interpreting students' correct and incorrect responses with respect to the learning goals. But even if a

rigorous content alignment is achieved, items may still have features that make it difficult to determine whether a student's answer choice reflects what that student knows about the targeted idea. To further improve the validity of the claims we make from test item results, we identify and systematically eliminate as many problems with comprehensibility, test-wiseness, and inappropriate task context as we can. The criteria we use for judging alignment and identifying threats to validity are detailed in Project 2061's Assessment Analysis Procedure, available on the Project 2061 website at <http://www.project2061.org/assessment/analysis>. The use of the assessment analysis criteria in item development has been described in DeBoer, Lee, and Husic, in press; DeBoer, et al., in press; and DeBoer, Gogos, and Herrmann Abell, 2007.

### **Using Student Data to Inform the Design of Assessment Items**

Rigorously applying a set of criteria to determine the alignment of test items to learning goals and identifying features that obscure what students really know are important steps in the development of items that accurately measure the knowledge we want students to have. But findings from our research also indicate that this analytical approach is significantly enhanced when it is used in conjunction with one-on-one interviewing and pilot testing during which students are asked for their feedback on the items and their answer choices are compared to the reasons they give for those answer choices (DeBoer & Ache, 2005). Pilot testing and interviewing provide us with insights about the knowledge students have, how they interpret the wording of the test items, and how they reason through the answer choice options.

During pilot testing, we ask students to circle words they do not understand and to tell us if anything is confusing about a question, whether or not they guessed, whether each of the four answer choices is correct or incorrect, and why each answer choice is correct or incorrect. By comparing the answers that students select with their written reasons for their answer choices, we determine if an assessment item is measuring what we want it to measure or if students' answer choices reflect false negative or false positive responses to that item. Students can also indicate if they are unsure if something is correct or incorrect. A large number of unsure responses may indicate that there are structural problems with an item that need to be addressed, especially problems with terminology that may be unfamiliar to students.

Pilot tests are carried out in middle schools serving a wide range of students in urban, suburban, and rural communities. Approximately 100 to 150 students respond to each item during pilot testing. In some cases we set up comparative studies within classrooms to test specific hypotheses. For example, we may test the impact that a difference in wording between two answer choices has, or we may test whether an illustration has an effect on how students answer a question. Sometimes it is possible to give the two items being compared to the same students, but when one question is expected to have an effect on the answer to another question, we randomly distribute one form of the question to half of the students in a class and the other question to the other students, or we use other means of creating equivalent comparison groups. In most cases we simply observe how frequently students select each answer choice and carefully analyze the comments that they make, looking for problems they might have had interpreting the question we asked.

## General Methods

In 2007 we pilot tested items on the topics of Interdependence in Living Systems; Matter and Energy in Living Systems; and Substances, Chemical Reactions, and Conservation of Matter. We pilot tested items in 15 schools in 11 states that were located mostly along the East Coast and in the Midwest. The demographics of the school districts where we piloted were 63.6% White; 28.0% Black; 5.9% Hispanic; and 2.9% Asian, Pacific Islander, or Native American. Approximately 30% of the students in the districts where we pilot tested were eligible for free and reduced lunch or were otherwise categorized as economically disadvantaged. The schools differed widely in their ethnic and racial composition and in the number of students who were eligible for free and reduced lunch. Schools were located in rural, urban, and suburban communities. Students indicated their gender; whether they were in sixth, seventh, or eighth grade; and if English was their primary language. During pilot testing, students answered a series of questions about each test item. The pilot test questions appear in Figure 1.

*Figure 1: Questions asked during pilot testing*

- 
- |  |                  |
|--|------------------|
| 1. Is there anything about this test question that was confusing? Explain.   |                  |
| 2. Circle any words on the test question you don't understand or aren't familiar with.   |                  |
| 3. Is answer choice A correct?<br>Explain why or why not.  | Yes No Not Sure  |
| 4. Is answer choice B correct?<br>Explain why or why not.  | Yes No Not Sure  |
| 5. Is answer choice C correct?<br>Explain why or why not.  | Yes No Not Sure  |
| 6. Is answer choice D correct?<br>Explain why or why not.  | Yes No Not Sure  |
| 7. Did you guess?  | Yes No           |
| 8. Should there be any other answer choice?<br>If you think so, what are they?   | Yes No           |
| 9. Was the picture or graph helpful?<br>Why or why not? [Or, if there was no picture...]<br>Would a picture or graph be helpful? Why or why not? | Yes No<br>Yes No |
| 10. Have you studied this topic in school?   | Yes No Not Sure  |
| 11. Have you learned about it somewhere else?<br>Where? (TV, museum, etc)?   | Yes No Not Sure  |
- 

Because students could indicate that they were “Not Sure” about an answer choice or could indicate that more than one answer was correct, the percent correct that we are reporting for items in this paper is typically lower than it would have been if students had been forced to select just one answer choice. In some cases, more than half of the students said they were Not Sure for all answer choices rather than saying any of the answer choices was correct or incorrect. We considered both Not Sure and multiple answer selections to be incorrect.

In the following section, we provide examples of how we used pilot testing to identify problems with test items and to gain insight into student thinking about the ideas the items are targeting. Examples are from two of the three topic areas that were pilot tested in the spring of 2007—Matter and Energy in Living Systems; and Substances, Chemical Reactions, and Conservation of Matter. All of the items that are discussed in this paper are in various stages of development, not final items.

### **Using Pilot Test Results to Inform Item Revision: Examples from Two Topic Areas**

#### **Topic 1: Matter and Energy in Living Systems**

Items for the topic of Matter and Energy in Living Systems target student understanding of the nature of food and how food is used for growth in living organisms. In this paper we focus on two approaches we took to investigating students' understanding of food. First, we examined student responses to an item about the nature of food to probe whether students have difficulty interpreting answer choices in which they are asked to select a correct claim about the nature of food along with the correct reason for that claim. Second, we evaluated student responses to three versions of an item designed to test students' understanding of how food is used for growth to determine what knowledge students used to evaluate the correct answer choice across the different versions of the item.

**What is food?** The first item targets the idea that both plants and animals need food as a source of energy (see Figure 2). Students are expected to know that water is not food, and they are expected to know that the reason water is not food is because water is not a source of energy. We focused on water in this item because many students think that water is food for organisms (Lee & Diong, 1999) or is a source of energy for organisms (Horizon Research). The distractors targeted three misconceptions related to alternative definitions of food:

- Some students see food as any material (water, air, minerals, etc.) that organisms *take in* from their environment (Anderson, Sheldon, & DuBay, 1990; Roth & Anderson, 1987; Simpson & Arnold, 1982).
- Some students think that food is whatever *is needed* to keep animals and plants alive (Leach, Driver, Scott, & Wood-Robinson, 1992; Lee & Diong, 1999; Roth & Anderson, 1987) or grow (Anderson, Sheldon, & DuBay, 1990) without reference to any more specific function of food.
- Some students think that liquids cannot be food (Lee & Diong, 1999).

Figure 2: Pilot Test Results for Item ME1-6 (N=117)

---

Is water food for plants and animals?

- (14.5%) A. Yes, because water is taken in by plants and animals
  - (41.0%) B. Yes, because water is necessary for plants and animals
  - (12.0%) C. No, because liquids are not food for plants and animals
  - (10.2%) D. No, because water does not provide energy for plants and animals\*
- 
- (13.7%) Not Sure or Blank
  - (10.2%) Multiple Selections

Copyright © 2007 AAAS Project 2061

---

*Percentage of students choosing each answer is in parentheses.*

*\* Correct Answer*

Most students (41.0%) incorrectly selected answer choice B. Each of the remaining answers was chosen by about 10-15% of the students (see Figure 2). The item asks students to select a claim about water as food and a reason for the claim. We had expected students to select an answer choice based on their knowledge of a general principle and their ability to reason about whether the information in the answer choice was consistent with that general principle. If students selected the correct answer choice, we expected they would know the general principle that to be considered food a substance must be a source of energy for plants and animals. And we expected that students would reason that since water does not provide energy for plants and animals, water is not food. We had also expected that students who selected the distractors would do so because of their lack of knowledge of the general principle. For example, a student who selected answer choice A would hold the (incorrect) idea that “Anything that is taken in by plants and animals is food,” and would reason that since water is taken in by plants and animals, water must be food for plants and animals. However, when we examined the explanations students gave for their answer choices, we found that most students did not provide evidence that they were using the expected reasoning. For example, the explanations most students gave for selecting answer choices A or B did not cite a general principle or imply that they used a general principle in their reasoning. Only two students out of 72 provided written comments that showed that they thought water was food because it fit the rule/general principle that things that are taken in are food. Instead, most students stated that A or B was correct because the answer choice contained an accurate statement of fact, e.g., “water is taken in by plants and animals” or “water is necessary for plants and animals.” As a result, students did not appear to be answering the question we intended to ask, perhaps because they did not appreciate the difference between a correct statement and a correct reason.

This is not to say that all students were unable to see the difference between a correct statement and a correct reason. A number of students who said answer choices A or B were *not* the correct answers made comments that said that just because it was true that plants and animals take in water does not make water food. One student said clearly: “The answer is true but it is not the reason.”

Based on these observations, we revised the item for future testing. To avoid the problem of students' selecting answer choices because they contain correct statements of fact rather than because they are correct reasons for why water is or is not food, we followed the Yes-No claim with a statement of the general rule or principle that applies. By doing this, students could not avoid confronting the general rule or principle that justifies the Yes-No claim, and we could be more confident that their answer selections were based on their knowledge or lack of knowledge of that general rule or principle rather than just their knowledge of the truth of the claim. To further emphasize that we were interested in the reason why water was or was not food, we added the phrase, "Why or why not?" to the stem.

*Figure 3: Revised Stem and Sample Answer Choice for Item ME1-7*

---

REVISED STEM: Is water a source of food for plants and animals? Why or why not?

REVISED ANSWER CHOICE: Yes, because food is anything that is needed by plants and animals

---

We also noted that the two misconceptions in answer choices A and B—"food is anything that is taken in," and "food is whatever is needed"—seem to be conflated in the minds of students. Ten of 117 students said both A and B were correct answers, and in their written comments, a number of students either said both were true or they used the word "need" when explaining why A ("anything that is *taken in*") was true. Of 23 students who made written comments justifying why answer choice A was correct, nine used the word "need" in their explanation and only five used the words "taken in." Of 49 students who made written comments justifying why answer choice B ("whatever is *needed*") was correct, 37 explicitly used words such as "need," "necessary," "required," and "essential." None of the students used the phrase "taken in." From these written comments, it appears that for middle school students the stronger misconception is that water is food because it is needed, not because it is taken in. Results that we obtain from testing the revised item on a national sample will better answer the question of how strong these two misconceptions currently are among middle school students, especially when their attention is focused on the reason for something being food, not just on whether the answer choice included an accurate statement about a particular substance (water).

**Food for Growth.** Another learning goal targeted in our assessment development is the idea that the growth of organisms involves chemical reactions in which the atoms of molecules from food are used to make new molecules that become part of the organism's body structures. Many students think that food is a requirement for growth rather than a source of matter for growth (Leach et al., 1992; Smith & Anderson, 1986). They do not know that the atoms that make up the molecules from food are incorporated into the molecules that make up a growing organism's body. Some of these students may think that the food we eat somehow gets broken down and then those breakdown products are added to the body rather than that those breakdown products are used to make new molecules that are incorporated into the body structures. Other students think that food supplies an organism with energy for growth but not matter for growth (Smith & Anderson, 1986). To probe students' understanding of this idea, we formulated three versions of an item, each of which used identical distractors but a different correct answer to describe what happens to the molecules from food as an animal grows (Figure 4). One of the three items

appears in Figure 5. We administered the three versions to different students of the same teacher. Altogether, we used 18 classrooms taught by six teachers in five different schools.

*Figure 4: Correct Answer Choices for Items Written to Test Student Ideas about Food for Growth*

---

(15.8%)	Version 1, N = 76: An animal's body structures grow as the atoms of molecules from food are rearranged to form new molecules that become part of its body structures.
(43.6%)	Version 2, N = 55: An animal's body structures grow as molecules from food are broken down and reassembled into molecules of carbohydrates, fats, and proteins that make up its body structures.
(20.0%)	Version 3, N = 115: An animal's body structures grow as carbon atoms from food are linked together and to other atoms to make new molecules that become part of its body structures.

---

*Percentage of students choosing the correct answer is in parentheses. N is the number of students responding to the item.*

The proportion of students selecting the correct answer varied significantly across the three versions of the item ( $\chi^2=15.4$ ,  $df=2$ ,  $P < 0.001$ ), with more than twice as many students selecting the correct answer for Version 2 (46.3%) than for Versions 1 and 3 (15.8% and 20.0% respectively). From these results and the fact that the average percentage correct for all of the molecular-level items associated with this idea that we pilot tested is only 21.6%, it appears that the idea that an animal's body structures grow as atoms and molecules from food are incorporated into those body structures was unfamiliar to these students. The unusually high percentage of correct responses on Version 2 in comparison to the other items caused us to look closely at the written comments on the three versions.

*Figure 5: Version 1 of Three Items Written to Test Student Ideas about Food for Growth*

---

How do an animal's body structures—such as muscle, bone, and skin—grow?

- A. An animal's body structures grow as the molecules from food—including carbohydrates, fats, and proteins—accumulate inside the animal unchanged.
- B. An animal's body structures grow as the atoms of molecules from food are rearranged to form new molecules that become part of its body structures.\*
- C. An animal's body structures grow as the animal converts carbon dioxide molecules to form new molecules that become part of body structures.
- D. An animal's body structures grow as vitamins and minerals are added to its body structures unchanged.

Copyright © 2007 AAAS Project 2061

---

*Percent of students choosing each answer is in parentheses.*

*\* Correct Answer*

For Version 1 (15.8% correct), the correct answer says that growth occurs as “atoms of molecules from food are rearranged to form new molecules that become part of an animal's body structures.” Of the students who selected this as the correct answer, none mentioned atoms or molecules in their written comments. One student did say, “Yes, they are rearranged so the body can use them,” almost certainly referring to the atoms and molecules, but no other students did. For students who thought this was not the correct answer, their comments included: “We don't form new molecules, do we?” “Molecules from food are not rearranged to form new molecules.” One student said: “No, atoms and molecules don't make you grow.” Overall, the student responses to this version of the item suggest that very few of these students understood that growth in organisms occurs as atoms of molecules from food are rearranged to form new molecules that become part of body structures.

For Version 2 (43.6% correct), the correct answer says that growth occurs “as molecules from food are broken down and reassembled into molecules of carbohydrates, fats, and proteins that make up its body structures.” Of the students who selected this as the correct answer, many focused on the part of the answer choice that mentioned food being broken down, not on the part that mentioned the reassembly of those breakdown products into molecules that make up its body structures. Student comments included: “The food is broken down; that is how your body grows.” “Body changes food to muscle and fuel.” It appears that in this version of the item students looked past the idea that growth requires the assembly of “molecules” that become part of body structures and focused instead on a substance level “breaking down” of food, something their comments suggest they were familiar with, perhaps from what they had learned about digestion. What these students said is not incorrect, but their written comments reveal that selecting the correct answer did not require the understanding of growth at the molecular level that we were looking for. Finally, although we had some concern that students might be attracted to the correct answer choice by the words “carbohydrates, fats, and proteins,” this did not seem to be the case. In fact, some students who thought this was the incorrect answer said: “Food is already made up of carbs, fats, and proteins; it doesn't need to be changed into them.”

For Version 3 (20.0% correct), the correct answer says that growth occurs as “carbon atoms from food are linked together and to other atoms to make new molecules that become part of its body structures.” Many students focused on the word “carbon” and its association with living organisms but not on carbon as atoms that are linked together to make molecules that become part of body structures. Comments from students who thought this was a correct answer included: “All animal life is carbon based, so animal’s cells would have to be carbon based.” “Carbon is the main part of everything and to grow they must link together building onto you.” “Carbon chains are the base of all life.” For students who thought this was an incorrect answer, comments included: “Muscle builds from protein, not carbon.” “No, carbon has nothing to do with growing.” Although it is possible that these students may have had a mental model of carbon atoms becoming incorporated into the molecules that make up body structures, the comments seem to suggest that students chose this answer either because carbon was associated in their minds with living organisms or because they were thinking of carbon being “stuck” onto the body structures.

A few students focused on the idea of atoms being linked together, but even those students who mentioned atoms in their explanations provided very little evidence that they had a mental model in which carbon atoms become linked together to form molecules that make up an organism’s body structures. For students who used the language of atoms and molecules to explain their selection of the correct answer choice, comments included: “Yes, molecules do link.” “Yes, atoms have to do with structure growth.” “Atoms [from] food are linked together.” “Food and atoms help you grow.” For students who thought this was an incorrect answer, comments included: “No, because two atoms don’t form molecules.” “Carbon is a gas and it can’t link together to form other atoms.” “Carbon atoms don’t become part of its body structure.” “No, because they don’t link together.”

Taken together, pilot test results for these three items suggest that most middle school students that we tested do not have an atomic-molecular understanding of how growth occurs in organisms. Even when students did choose the correct answer, many of their explanations focused on substance-level ideas such as “breakdown of food” or “carbon [sticking] to bone structures.” Of the three versions of the item, Version 2, which allows students to answer correctly based on their knowledge that food is “broken down,” was the least effective item (i.e., it produces the greatest number of false positives) in this set for measuring student understanding of the targeted idea that growth in organisms involves chemical reactions in which molecules from food are used to make new molecules that become part of the body structures. Based on student comments to this item, the wording in the answer choice appeared to allow students to evaluate the correct answer choice using alternative knowledge of what happens to food *before* it is used as a source of building materials for growth. As a result, the item fails to target a molecular-level understanding of the use of food as a source of building materials for growth and suggests that the item could be improved by removing reference to food being “broken down of food.” This example illustrates how understanding the knowledge students use to evaluate an answer choice (as opposed to the knowledge we expected them to use to evaluate an answer choice) can guide the development of more informative assessment items.

## Topic 2: Substances, Chemical Reactions, and Conservation of Matter

In the Substances, Chemical Reactions, and Conservation of Matter topic, we targeted ideas related to chemical reactions and conservation of matter both at the substance level and the atomic-molecular level. At the substance level, students are expected to know that substances react to form new substances with different characteristic properties and that total mass is conserved. At the atomic level, they are expected to know that during chemical reactions atoms are rearranged to form new molecules and the number of each kind of atom is conserved.

**Chemical Reactions.** One of the questions we explored was how well middle school students would perform on test items written at the atomic-molecular level compared to how well they would perform on items written at the substance level. One hundred and fifty eight students in the seventh and eighth grades each received six of 11 items we had written to test ideas about chemical reactions at the substance level (substances react to form new substances with different characteristic properties) and five of the seven items we had written to test ideas about chemical reactions at the atomic level (atoms are rearranged to form new molecules). Different classes of students received different subsets of the items, but all students received six substance-level items and five atomic-level items. Students were compared on the number of items they answered correctly out of the total number they received using a paired samples t-test. Across all items testing the nature of chemical reactions, the average score was 37.0% correct on the substance-level items and 23.5% correct on the atomic-level items ( $t= 5.16, p<0.01$ ). In the following sections of the paper, we provide more detail about student understanding of chemical reactions at the substance and atomic-molecular levels.

*Chemical reactions at the substance level.* In one of the substance-level items (see Figure 8), students were told that two liquids were mixed together and that after mixing a solid substance formed. They were asked if this was a chemical reaction and why.

Figure 6: Pilot Test Results for Item SC67-2 (N=42)

---

A student mixes two different liquids together. After mixing the liquids, a solid substance forms. The student claims that a chemical reaction occurred. Is the student correct and why?

- (2.4%) A. Yes, because a solid is always formed during a chemical reaction.
- (4.8%) B. Yes, because the solid cannot be turned back into the starting liquids.
- (38.1%) C. Yes, because the solid is a new substance that was formed during a chemical reaction.\*
- (11.9%) D. Yes, because a chemical reaction always happens when two liquids are mixed together.
  
- (35.7%) Not Sure or No Response
- (7.1%) Multiple Selections

Copyright © 2007 AAAS Project 2061

---

Percent of students choosing each answer is in parentheses

\* Correct Answer

On this item, 38.1% answered correctly, and from their comments it was clear that many students knew that new substances are formed during chemical reactions. When asked why each answer choice was correct or incorrect, comments by students who chose the correct answer included: "Yes, because a new thing must form." "Yes, when chemical reactions happen you usually get something new. Example: new colors, substance, texture or smell." "Yes, the solid is a new substance from chemical reaction." "Yes, because new substances form when a chemical reaction occurs." "Yes, solid is a new substance created in a chemical reaction."

For students who thought the answer was not correct, no strong pattern emerged to suggest why they thought the answer was incorrect. One set of comments focused on an apparent misinterpretation of the item. A number of students said that "two liquids mixed together do not *always* form a solid," even though the answer choice does not use the word "always." Comments included: "No, because 2 liquids might not make a solid." "No, a solid doesn't have to form because of a chemical reaction." "No, because I don't know if a solid substance always formed after 2 liquids are mixed." To address this possible confusion about whether the question is asking if solids always form during chemical reactions, answer choice C was reworded so that it was clear that a new *substance* (not necessarily a new solid) is always formed during a chemical reaction. The answer choice now reads: "Because a new substance is always formed during a chemical reaction." Students also pointed out that there was no answer choice that said "No, the student is not correct." To address that concern, the question in the stem was changed from: "Is the student correct and why?" to "Why is the student correct?"

For another item involving the nature of chemical reactions at the substance level (see Figure 9), students were asked if chemical reactions involve liquids only; solids and liquids only; solids, liquids, and gases; or if they always produce a gas.

*Figure 7: Pilot Test Results for Item SC72-1 (N=49)*

---

Which of the following statements describes chemical reactions?

- (2.0%) A. Chemical reactions involve liquids only.
  - (4.1%) B. Chemical reactions always produce a gas.
  - (2.0%) C. Chemical reactions always involve a solid and a liquid.
  - (71.4%) D. Chemical reactions occur between solids, liquids, or gases.\*
- (14.3%) Not Sure or No Response  
(6.1%) Multiple Selections

Copyright © 2007 AAAS Project 2061

---

*Percent of students choosing each answer is in parentheses*

*\* Correct Answer*

Overall, 71.4% of the students answered correctly that chemical reactions can occur between gases, liquids, and solids. For the students who did not choose the correct answer, their reasons included: "No, because you can't get chemical reaction with a gas." "No, gases are what can be

produced.” “No, it doesn't involve solids.” “Not Sure. I know liquid and a solid but I'm not sure about gases maybe too.”

Overall, at the substance level students seemed comfortable with the idea that something changes during a chemical reaction, but they were less certain about the details of which substances can and cannot change.

*Chemical reactions at the molecular level.* At the molecular level, we wanted to find out if students knew that chemical reactions involve a rearrangement of the atoms to form new molecules. In one question, students were told that two substances were mixed together and that a chemical reaction occurred. Then they were asked what happens to the molecules of those substances and the atoms they are made of during the chemical reaction. The correct answer choice said that the atoms stayed the same but rearranged to form new molecules (see Figure 10).

*Figure 8: Pilot Test Results for Item SC35-2 (N=62)*

---

Two substances are mixed together. The molecules of the substances interact and a chemical reaction occurs. Which of the following describes what happens during the chemical reaction to the molecules and the atoms they are made of?

- (25.8%) A. The atoms change into new atoms and form new molecules.
  - (21.0%) B. The atoms stay the same but rearrange to form new molecules.\*
  - (11.3%) C. Some atoms rearrange to form new molecules and some atoms are destroyed.
  - (4.8%) D. The molecules and the atoms they are made of are the same after the chemical reaction.
- (30.6%) Not Sure or No Response  
(6.5%) Multiple Selections

Copyright © 2007 AAAS Project 2061

---

*Percent of students choosing each answer is in parentheses*

*\* Correct Answer*

Only 21.0% of the students answered correctly. Many students said they were unfamiliar with words such as “chemical reactions,” “atoms,” “molecules,” and “interact.” Nearly a third of the students did not pick any answer as correct, choosing instead “Not sure.” The most commonly chosen incorrect answer indicated that the atoms changed into new atoms (25.8%). The least popular answer choice was that the molecules and the atoms they are made of stayed the same after the reaction (4.8%). For students who thought the correct answer choice was not correct, comments included: “No, how are they going to stay the same but rearrange to new molecules.” “No, I don't think the atoms will stay the same.” “No, the atoms would not stay the same.” “No, you can't rearrange molecules.” “No, how can atoms stay the same and rearrange?” “No, I believe the atoms would change somehow.” “No, you need new atoms to make new molecules.” Students seem to know that something changes in a chemical reaction, but they are not sure what changes. At the atomic-molecular level, they are unaware that the atoms themselves stay intact and just rearrange to form new molecules. Many of the students thought that the atoms

themselves would have to change. For some students there was also the question of what “rearrange” means. Does it mean that the atoms change identity or change position? The student who said: “No, how can atoms stay the same and rearrange?” seems to be confusing change in position with change in identity. Future revisions of this item will focus on student confusion about the meaning of “rearrange.”

**Conservation of Matter.** One hundred and seventeen students in the sixth, seventh, and eighth grades each received six items (from a total of 12 that we had written on the topic) that were aligned to the idea of conservation of matter at the substance level (total mass is conserved) and five items (from a total of seven that we had written for this topic) that were aligned to the idea of conservation of matter at the atomic level (the number of each kind of atom is conserved). Not all students received the same subset of items, but all students received items on conservation at both the substance and atomic levels. Students were compared on the number of items they got correct out of the total number they received using a paired samples t-test. Across all items testing conservation of matter, the average score was 23.7% correct for the substance-level items and 16.9% correct for the atomic-level items ( $t=2.37$ ,  $p<0.05$ ). In the following sections of the paper, we provide more detail about student understanding of conservation of matter at the substance and atomic levels.

*Conservation of matter at the substance level.* At the substance level, students were tested on the idea that matter is conserved regardless of what changes are made to a substance. In one item involving sugar dissolving in water, students were asked if the mass of the water and sugar is more, less, or the same before and after the sugar dissolves, or if it depends on how much of the sugar dissolves (see Figure 11).

*Figure 9: Pilot Test Results for Item SC56-2 (N=37)*

---

A student adds a spoonful of sugar to a cup of water and covers the cup so that nothing can get out. The student then records the starting mass of the materials. Next the student shakes the cup until some of the sugar dissolves in the water. What happens to the mass of the materials after the sugar dissolves?

- (21.6%) A. The mass is the same as the starting mass.\*
- (10.8%) B. The mass is more than the starting mass.
- (8.1%) C. The mass is less than the starting mass.
- (8.1%) D. It depends on how much of the sugar dissolves.

(48.6%) Not Sure or No Response  
(2.7%) Multiple Selections

Copyright © 2007 AAAS Project 2061

---

*Percent of students choosing each answer is in parentheses*

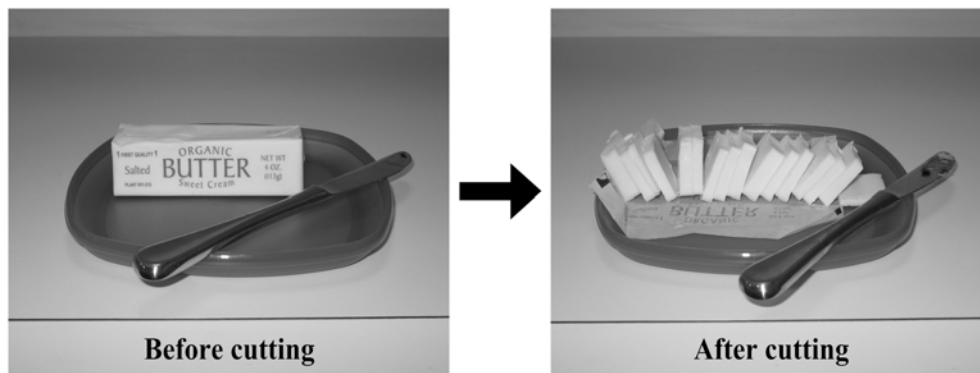
*\* Correct Answer*

Only 21.6% of students answered correctly, and nearly half of the students were unsure of their answer. Most of the incorrect answers seemed to be due to students' focusing on what was happening to either the solid sugar or to the liquid rather than the combination of the two. Some of the comments from students who thought the mass would not be the same before and after the sugar dissolved seemed to focus on the solid sugar weighing less because it had dissolved: "No. When the sugar dissolves, there is less in the cup." "No, because if the sugar dissolves the mass will probably go down." But most students focused on what happened to the weight of the liquid: "No, because the sugars mass has added to the waters mass." "No, because when the sugar dissolves it adds to the mass." "No, the sugar adds mass!" To counter this tendency to focus on either the solid sugar or the liquid rather than on the combination of the sugar and the water, we revised the item by adding a picture to show the sugar and the water in the sealed jar before and after mixing, and we simplified the language of the question. Instead of asking "What happens to the mass of the materials after the water dissolves?" we now ask: "What will happen to the weight of the jar containing the water and sugar after some of the sugar dissolves?"

Another substance-level conservation item asked students what would happen to the weight of butter after it was cut into smaller pieces (see Figure 10).

Figure 10: Pilot Test Results for Item SC78-2 (N=42)

A student cuts a stick of butter on a dish into smaller pieces using a knife. The student weighs the materials before and after cutting the butter into pieces. Will the butter weigh more, less, or the same when it is in small pieces and why?



- (2.4%) A. The butter will weigh less because the small pieces are smaller than the stick.
  - (2.4%) B. The butter will weigh more because there are lots of small pieces but only one stick.
  - (4.8%) C. The butter will weigh less because some of the weight disappears when the butter is cut into small pieces.
  - (69.0%) D. The butter will weigh the same because the small pieces are the same amount of butter as the stick just in a different form.\*
- (21.4%) Not Sure or No Response  
(4.8%) Multiple Selections

Copyright © 2007 AAAS Project 2061

*Percent of students choosing each answer is in parentheses*

*\* Correct Answer*

Students were much more successful on this item than they were on the dissolving sugar item, with 69.0% of the students choosing the correct answer. The question seemed to be less confusing for students, probably because they were asked to focus on just the butter rather than having to consider the combination of two substances (sugar and water). Still, there were issues that came up in student comments that caused us to make revisions to the item. For students who thought answer choice C was correct (*The butter will weigh less because some of the weight disappears when the butter is cut into small pieces*), a number of them focused on the possibility that some of the butter would now be on the knife. “Yes, because some of the butter is on the knife (after being cut) so it would weigh less.” “Yes, I think this is right because butter does come off and the less butter, the less it weighs.” “Yes, maybe because some of the butter was still stuck on the knife and plate or package.” To avoid confusion about whether the butter we were asking about included the butter that was stuck to the knife and paper, the stem now reads: “A student cuts a stick of butter on a dish into smaller pieces using a knife. The student weighs the dish, knife, wrapper, and butter before and after cutting the butter into pieces. Will the dish,

knife, wrapper, and butter weigh more, less, or the same when the butter is in small pieces and why?"

After examining the results for the butter and the dissolving sugar items, it appears that for the most part students have a solid grasp of the idea of conservation of mass on the substance level. Their written responses to the butter item showed a clear understanding of the fact that mass stays the same when something is cut into smaller pieces. Incorrect answers were due mainly to a confusion about which part of the butter was to be considered. And incorrect answers to the dissolving sugar item seem to have more to do with the students' understanding of the nature of dissolving than their understanding of conservation of mass.

*Conservation at the atomic level.* At the atomic level, we expect students to know that both the identity and number of atoms are conserved during phase change, dissolving, and chemical reactions. In one question (see Figure 11), students are given a situation in which bubbles form when two liquids are added together. They are told that this reaction occurs in an open container.

*Figure 11: Pilot Test Results for Item SC45-2 (N=83)*

---

A student finds the total mass of two liquids. When she mixes the liquids in an open container, she observes bubbles. After the bubbling stops, she finds that the total mass of the liquids in the container is less than the total mass of the liquids before they were combined. How can her observation be explained?

- (16.9%) A. Some atoms escaped from the container.\*
- (13.3%) B. Some atoms in the container were destroyed.
- (4.8%) C. Some atoms in the container spread out more.
- (7.2%) D. Some atoms in the container became smaller and lighter.

(53.0%) Not Sure or No Response  
(4.8%) Multiple Selections

Copyright © 2007 AAAS Project 2061

---

*Percent of students choosing each answer is in parentheses*

*\* Correct Answer*

This was a difficult item for students. Only 16.9% chose the correct answer and over half selected Not Sure or did not select any answer choice. Another 13.3% thought that atoms could be destroyed, and 7.2% thought some of the atoms got smaller and lighter. Students who thought atoms could be destroyed wrote: "Yes, its possible that some got destroyed." "Yes, the atoms could have not been powerful enough to survive." "Yes, they could have been destroyed when the liquids were mixed." "When two liquids come together like that they will be destroyed." "After another liquid was added it could kill some of the atoms." Many of the students insisted that the atoms could not escape. "No, because atoms don't escape." "No, because atoms can't escape from a container." "No, if its sealed then nothing can get out." "No, they cant escape unless they cut holes in the top." Apparently, these students missed the fact that the container



Figure 13: Pilot Test Results for Item SC89-3 (N=88)

	A	B	C	D	Not Sure / No Response	Multiple Selections
% of students choosing each answer choice	2.3%	0%	25.0%	3.4%	67.0%	2.3%

About two-thirds of the students selected Not Sure or did not select any answer choice, and only 25% of the students chose the correct answer (see Figure 13). Most of the students who answered correctly made comments that suggested that they added up the white, gray, and black balls to get the correct answer. Written comments from students who chose the correct answer included: “Yes, it adds up with the molecules that there was before.” “Yes, it is the right amount of everything.” “Yes, I think choice C will be the correct choice because it looks like the right amount of molecules.” “Yes, every atom is a counted for.” “Yes, molecules are equal on each side.” The comments of students who did not select the correct answer focused on the number of molecules before and after the reaction. Comments included: “No, it does not have the right amount of molecules.” “No, because they are only two different molecules.” “No, all of the molecules aren't there.” “No, because if you add them together it equals 7.” (There were six molecules to start.) A number of students confused atoms and molecules in their answers, and many of the students said they had not studied this topic before and were unfamiliar with atoms or molecules. From these results, it is clear that most of the middle school students we tested do not know that during chemical reactions the number and identify of atoms is conserved. Comments also indicate considerable unfamiliarity with the language of chemical reactions, including the meaning of words such as atom, molecule, or chemical reaction.

Because this item did not include a reason why the molecule represented in each answer choice is the correct answer to the question, we created a new item to accomplish that. In addition to providing reasons for each answer choice, the new item also provides students with support in interpreting the language of atoms and molecules. They are explicitly told that atoms are represented by circles and that molecules are represented by circles that are connected to each other. Field testing this item will give us a much better idea of student understanding of conservation at the atomic level.

The results presented here show that for the idea of conservation, the context in which the question is asked is important in how the students answer. Our pilot test items were written on the assumption that students already knew what chemical reactions, dissolving, and phase change were so that the idea of conservation could be tested in any of those contexts. But, in fact, most of the middle school students we tested did not have a solid grasp of those prior ideas, and that as much as a lack of understanding of conservation may be the reason for their incorrect answers.

*The use of actual names of compounds vs. generic names for those compounds.* Finally, another question we asked for the Substances, Chemical Reactions, and Conservation topic is how well students would respond to items in which named chemical substances are used, e.g., butanol and butanoic acid, compared to substances with generic names, e.g., liquid 1, liquid 2, etc. We thought students might be helped by having the names of specific chemical compounds even if the students were unfamiliar with those substances. We piloted two versions of the same item. One version used the chemical names of the liquids and the other used generic “liquid 1, 2...”

labels. Students in sixth, seventh, and eighth grade classrooms from four different schools responded to the items. Half of the students were randomly selected to respond to the version with the chemical names and the other half responded to the version with the generic labels. The percent correct was 20.7% (N=87) for the version with the chemical names and 17.0% (N=88) for the version with the generic labels ( $\chi^2=0.39$ ,  $p>0.1$ ). However, over half (~58%) of the students that responded to the version with the chemical names circled the names of the chemical substances to indicate that they were unfamiliar with the words. During one-on-one interviewing, all of the students who were interviewed had trouble reading the chemical names out loud and said that they would prefer having the generic labels. As a result, the chemical names of the liquids were not used in the revised version of the item.

### **Summary**

A recent report by the National Mathematics Advisory Panel noted that there are many flawed items on existing tests, including NAEP and state tests, often related to the wording of an item. Their report suggests that: “Careful attention must be paid to exactly what mathematical knowledge is being assessed by a particular item and the extent to which the item is, in fact, focused on that mathematics” (2008, p. 60). We have found similar flaws in many of the existing science items. To improve on the quality of assessment items in science, we place a great deal of emphasis on the qualitative alignment of assessment items to learning goals. Our use of the criteria of necessity and sufficiency ensures a high degree of accuracy in this alignment. We also pay close attention to possible construct-irrelevant features of the test items that might produce false negative or false positive responses on the part of students. Low comprehensibility of the items and easy application of test-wiseness strategies by students can increase the number of students whose answers are not valid indicators of their actual knowledge. In pilot testing we ask students to explain why answer choices are correct or incorrect, and what they say is then matched with the answer choice they select to validate their answer choices or to uncover flaws in the items that lead to invalid interpretations of those answer choices.

Although this qualitative analysis is just one part of our item development process, and it is followed up by rigorous psychometric testing following field testing of approximately 2000 students per test item, the qualitative judgments are an essential part of the work we do. Whereas statistical procedures can be used to find out how difficult test items are, how they discriminate among high and low performing students, and whether or not they function differently for different subpopulations, a rigorous examination of the qualitative alignment of the items to the targeted learning goals is also needed to create items that measure what it is that we want to measure.

Multiple-choice tests are often criticized for assessing student knowledge of just the facts of science, but multiple-choice tests also can be constructed that ask students to think through more complex situations and to analyze, explain, and predict phenomena. And when clusters of items targeting the same idea are used together, they can be particularly helpful in assessing students’ understanding of key science ideas. Although a considerable amount of effort is required to construct such test questions, when done well they provide educators with important information about what students know and can do. The report of the National Mathematics Advisory Panel also noted that many educators believe that constructed-response items are more effective than

multiple-choice items and that they are a more authentic measure of mathematical skill. After examining the literature on the psychometric properties of constructed-response items as compared to multiple-choice items, they concluded that “the evidence in the scientific literature does not support the assumption that a constructed response format, particularly the short-answer type, measures different aspects of mathematics competency in comparison with the multiple-choice format” (p. 60). In other words, the current situation is that both kinds of items need improvement.

The procedures that we have developed can uncover many of the flaws in items that the Math Panel noted in their report. By drawing heavily on written feedback from students we can find out what they find confusing about an item or what words they are not familiar with. Seemingly simple words such as “unchanged” or “rearranged” may present problems in the context of the science being described, even though they are familiar words to students. Confusion often can be reduced simply by inserting a picture in an item or substituting a more familiar word. Feedback from students also allows us to probe their understanding of abstract symbols that are used to represent physical entities (such as variables in an experiment or organisms in a food web), and it allows us to probe how well they can interpret questions in which they are asked to select a reason why a claim is correct. The expectation is that after these flaws are detected and corrected, the items will perform well when subjected to the scrutiny of quantitative psychometric analysis. Multiple choice items have the advantage of being able to carefully define the space in which students are asked to think about a problem or task and thereby limit the irrelevant responses that often come with open-ended test questions. In particular, multiple choice items can deliberately focus on the misconceptions that students have and assess how common those misconceptions are.

In our work, we use a two-year development cycle that begins with clarifying the learning goals that we intend to target and ends with national field testing and an analysis of the psychometric properties of the items. We can validate our development process by predicting the effect that changes made following pilot testing will have on student performance on the field tests. Ideally, we would also pilot test the items multiple times, making revisions based on student feedback each time. Although we cannot do this repeated pilot testing with the resources we have, we strongly recommend that others make an effort to get this kind of feedback from their students as they use our sample items or other items they develop. We have found middle school students to be an invaluable resource for providing information needed to improve test items as well as to improve the validity of the conclusions we draw about their knowledge and their ability to reason clearly in science.

---

*This work is funded by the National Science Foundation ESI-0352473.*

## References

- American Association for the Advancement of Science. (1993). *Benchmarks for science literacy*. New York: Oxford University Press.
- 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.
- DeBoer, G. E., & Ache, P. (2005, April). *Aligning assessment to content standards: Applying the Project 2061 analysis procedure to assessment items in school mathematics*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada. Retrieved September 29, 2006, from <http://www.project2061.org/research/assessment/aera2005.htm>.
- DeBoer, G. E., Gogos, A., & Herrmann Abell, C. (2007, April). *Assessment linked to science learning goals: Probing student thinking during item development*. Paper presented at the annual conference of the National Association for Research in Science Teaching, New Orleans, LA.
- DeBoer, G. E., Herrmann Abell, C., Gogos, A., Michiels, A., Regan, T., & Wilson, P. (in press). Assessment linked to science learning goals: Probing student thinking through assessment. In R. Douglas, J. Coffey, & C. Stearns (Eds.), *Linking science and assessment*. Arlington, VA: NSTA Press.
- DeBoer, G. E., Lee, H. S., & Husic, F. (accepted). Designs for assessing coherent understanding of science. In Y. Kali, J. E. Roseman, & M. C. Linn (Eds.), *Designing coherent science education*.
- Driver, R., Leach, J., Millar, R., & Scott, P. (1996). *Young people's images of science*. Bristol, PA: Open University Press.
- 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, 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.

- National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.
- 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.
- Stern, L., & Ahlgren, A. (2002). Analysis of students' assessments in middle school curriculum materials: Aiming precisely at benchmarks and standards. *Journal of Research in Science Teaching*, 39(9), 889–910.
- Tretter, T. R., Jones, M. G., Andre, T., Negishi, A., & Minogue, J. (2006). Conceptual boundaries and distances: Students' and experts' concepts of the scale of scientific phenomena. *Journal of Research in Science Teaching*, 43(3), 282–319.
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.