CHAPTER 12

Assessment Linked to Science Learning Goals: Probing Student Thinking Through Assessment

George E. DeBoer and Cari Herrmann Abell
Project 2061, American Association for the Advancement of Science

Arhonda Gogos
Sequoia Pharmaceuticals

An Michiels
Leuven, Belgium

Thomas Regan
American Institutes for Research

Paula Wilson
Kaysville, Utah

Standards-based reform of K–12 science education is built on the idea that fundamental improvement begins with the development of a well-articulated and coherent set of learning goals to guide instruction and assessment. This vision for reform has been supported for over a decade by Project 2061, the education initiative of the American Association for the Advancement of Science (AAAS), through its *Science for All Americans* (1989) and *Benchmarks for Science Literacy* (1993), and by the National Research Council (NRC) through its *National Science Education Standards* (1996).

The focus of this chapter is on how to design science assessment items that are linked to the content standards in *Benchmarks for Science Literacy* and the *National Science Education Standards*. It outlines AAAS Project 2061's approach to the design of assessment items that can be used with students in middle and early high school. The chapter describes (1) our criteria for determining that the assessment items test the targeted ideas and not some other ideas, (2) precautions we take to ensure that each assessment item is a fair and accurate measure of student knowledge, and (3) the use we make of student feedback during item development rather than relying solely on psychometric features to judge the suitability of the items. Drawing on a set of example items and students' responses to them, we show how data obtained from students through pilot-testing provide important insights about what students do and do not know and how those insights can be used to improve the effectiveness of each assessment item as a measure of student learning.

Background

With the adoption of the federal No Child Left Behind Act (NCLB) in 2002, the need for assessments that are well aligned to content standards became greater than ever. NCLB placed a new emphasis on standards-based accountability for students, teachers, schools, school districts, and states. Beginning in the 2007 academic year, testing in science became part of the NCLB mandate, and each state is now required to assess students' achievement of its science standards at least once at the elementary, middle, and high school levels. This emphasis on accountability through assessments linked to content standards has the potential to bring all parts of the instructional system in science together with the common goal of improving student learning in science. The importance of science content standards is also noted in a 2006 report from the National Research Council:

> *To serve its function well, assessment must be tightly linked to curriculum and instruction so that all three elements are directed toward the same goals. Assessment should measure what students are being taught, and what is taught should reflect the goals for student learning articulated in the standards.* (p. 4)

But, for standards-based accountability to produce the desired results, assessment and instruction must in fact be aligned to the most important ideas in science, and assessment must in fact probe student understanding

of those ideas. If that can be accomplished, assessment can provide direction for instruction and provide valuable feedback to teachers about their teaching and their students' learning. However, there is growing concern on the part of policy makers, educators, and the public about the quality of state assessments, which are key to the success of the NCLB strategy. A 2006 study from the American Federation of Teachers, for example, concludes that for reading, mathematics, and science, the three subject areas that are the focus of NCLB testing, only 11 states have strong content standards and tests that are aligned to them.

Although much of the debate about assessment has focused on the high-stakes testing required at the state level by NCLB, the tests that teachers have available to them also warrant attention. Classroom testing should enable teachers to probe students' understanding of the science ideas specified in the content standards and get feedback about their own teaching of those ideas. According to the National Research Council,

> *…classroom teachers are in the position to best use assessment in powerful ways for both formative and summative purposes, including improving classroom practice, planning curricula, developing self-directed learners, reporting student progress, and investigating their own practices…. Teachers need not only…interpret the assessment-generated information, they also must use the information to adapt their teaching repertoires to the needs of their students.* (2001, p. 15)

Whether formative or summative in nature, classroom-based or statewide, high-quality assessments can contribute to student learning and promote science literacy for all as it is envisioned in the national and state science standards. At present, however, most educators agree that there are too many poorly written assessment items that do not align properly with the content standards. These problems are prevalent in other content areas as well, not just in science. A recent report by the National Mathematics Advisory panel (2008) notes that there are many flawed items on existing tests and 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" (p. 60).

To help address these problems in the context of science assessment, AAAS's Project 2061 is developing test questions for middle school and

early high school science that are aligned to core ideas in national and state content standards and designed to be highly effective probes of students' understanding of those ideas. Because these assessment items are aligned to the ideas in the content standards and not to any particular curriculum material, to get an item correct students must demonstrate real understanding of those ideas rather than merely reciting words they have memorized from their textbooks or heard in their classrooms. The items are filling other needs as well:

- They enable *teachers* to keep track of their students' understanding of specific ideas over time and to conduct classroom research on the effects of various instructional strategies on student learning of those ideas. Many of the assessment items that are being developed expect students to use their knowledge to explain and predict phenomena that they may not have encountered before in school. Items that are embedded in real-world contexts accessible to students but different from those commonly used in textbooks or classroom lessons enable teachers to gauge more precisely their students' knowledge of the science ideas. The items also provide diagnostic information to help teachers determine what misconceptions or other problems may be impeding their students' learning.
- They provide *test developers and test administrators*, particularly those at the state and district levels, with models for items that are well aligned to the science ideas targeted in state and national standards and that also conform to rigorous psychometric, linguistic, and cognitive requirements. Tests that are developed from such items can reliably inform education policy and decision making and ensure that the consequences of decisions made based on those tests are fair for students, teachers, administrators, and schools.
- They provide *curriculum developers and researchers* with high-quality assessment items that are aligned to content standards to compare the effectiveness of various instructional materials objectively. Existing assessment items are not focused enough on the specific ideas in the content standards to provide precise and replicable measures of student understanding of the ideas and skills included in those content standards. High-quality assessment items linked to content standards can also be integrated into instructional materials themselves. Although instructional materials

often include embedded questions and summary assessment activities, they are rarely linked to specific content standards and are rarely presented as probes to help teachers uncover their students' thinking so that instruction can be adjusted based on how students respond.

- They give *parents* and other members of the public specific information about what it is that children, teachers, and schools are being held accountable for with respect to the content standards of their state and local communities and what alignment of assessment to those content standards means. Clear statements of the standards themselves, as well as assessment items that measure understanding of the ideas in the standards, are essential if parents are to contribute meaningfully to their children's education.

Developing Assessment Items Aligned to Standards

The AAAS Project 2061 procedure for developing assessment items involves three stages: (1) clarifying the targeted content standard, (2) designing assessment tasks that are precisely aligned to the specific ideas in the targeted content standards, and (3) using data derived from one-on-one interviewing and pilot-testing items with students to improve the items' effectiveness. By focusing on the ideas that an item targets and on the item's likely effectiveness as an accurate probe of students' knowledge of those ideas, the process helps to articulate what is being tested by a particular item, thus improving the validity of interpretations that can be made from test results.

1. Clarifying the Content Standards

Both *Benchmarks for Science Literacy* and *National Science Education Standards* are organized around ideas and skills that all students should have learned by the end of each grade band if they are to achieve the goal of science literacy by the time they graduate from high school. The learning goals in these documents, and the organization of these learning goals, provide guidance for developing instruction and curriculum as well as assessment.

Key ideas. Although state and national content standards provide important guidance to assessment developers regarding what students should know in science, to increase the precision of the content alignment, 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 for a benchmark from the topic of plate tectonics:

> *The outer layer of the Earth—including both the continents and the ocean basins—consists of separate plates.*

Clarification statements. Clearly, there are concepts in this statement about Earth's plates that need to be elaborated. What knowledge, for example, should students have of what a plate is? Our clarification specifies the following:

> *Students are expected to know that the solid outer layer of the earth is made of separate sections called plates that fit closely together along the entire surface where they are in contact such that each plate touches all the plates next to it. They should know that any place where two plates meet is called a plate boundary. They should know that plates are continuous solid rock, miles thick, which are either visible or covered by water, soil, or sediment such as sand. They should know that the exposed solid rock of mountains is an example of plate material that is visible. Students are not expected to know the term bedrock. Students should know that 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), which together are large enough to make up almost the entire outer layer of the earth.*
>
> *.....They should also know that there are additional smaller plates that make up the rest of the outer layer, but they are not expected to know the size of the smaller plates or how many there are. Students are expected to know that the boundaries of continents and oceans are not the same as the boundaries of plates. They should know that some boundaries between plates are found in continents, some in the ocean floors, and some in places where oceans and continents meet. Students are not expected to know the names of specific plates or the exact surface areas of plates. Students are not expected to know the terms lithosphere, crust, or mantle; the difference between lithosphere and crust; or that a plate includes the crust and the upper portion of the mantle.* (DeBoer 2007)

This clarification statement was written in response to three questions that are central to the design of assessments that target a key idea: (1) Is the description of plates that is specified here what is needed for students of this age to form a mental model that allows them to predict and explain phenomena involving plates? (2) Is the description of plates that is specified here needed for students to understand *later ideas* and the accompanying phenomena that they will encounter? (3) Will the specified terminology contribute enough to students' ability to communicate about the targeted ideas to make that terminology worth learning?

In the case of Earth's plates, we judged that students should know what the plates are made of, approximately how thick they are, approximately how many there are, that the plates are not all the same size and shape, and that each plate directly touches all the plates next to it. These elaborations of the term *plate* are needed so that students can be helped to develop a mental model of a plate that enables them to understand ideas about plate motion and the consequences of plate motion. We made a judgment that although the term *lithosphere* is often used when communicating about plates, neither the term nor the concept (i.e., that it is made of both crust and the upper portion of the mantle) contributes enough to explaining phenomena related to plate motion to hold students accountable for knowing the term. At other times, we may judge that the idea behind the term is important even if the term itself is not. We recognize that individual teachers may choose to show students the relationship between lithosphere, upper mantle, and plates, but our assessment items will not include the term for reasons explained above.

Misconceptions. Research on student learning also plays an important part in helping us clarify the learning goals. It provides information about the age-appropriateness of the ideas we are targeting and the level of complexity of the mental models we can expect students to develop. The research on student learning also identifies many of the misconceptions that students may hold, which we include as distracters in the items so that we can test for these ideas alongside the ideas we are targeting.

Connections among ideas. Once the ideas to be assessed have been identified and clarified, our next step is to see how those ideas relate to other ideas within a topic and across grade levels. The objective here is to be as clear as possible about what ideas we are explicitly testing and what prior knowledge we can assume students will have. For example, if we expect stu-

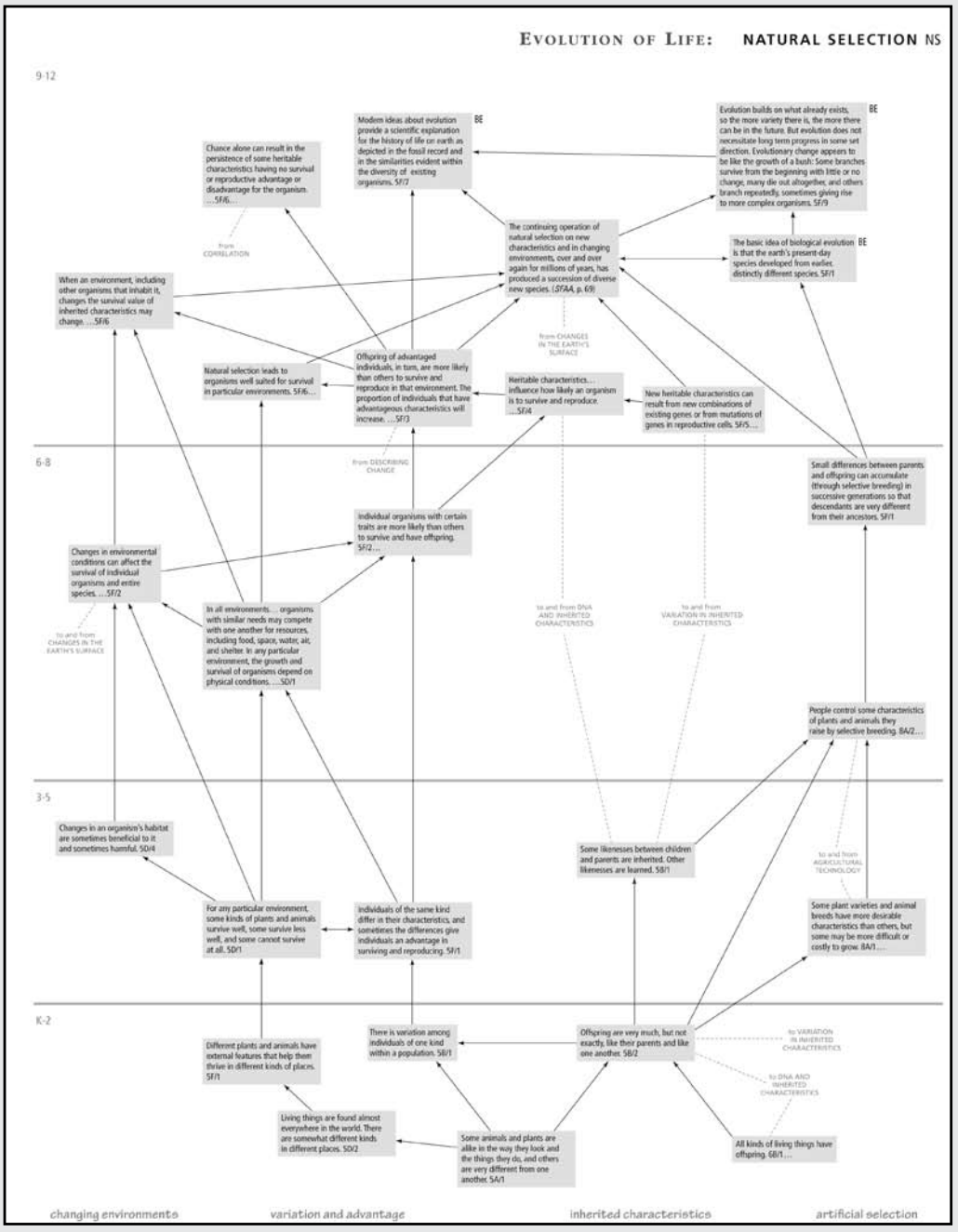
dents to know that digestion of food involves a process in which the atoms of molecules from food are rearranged to form simpler molecules, should we assume that students already know that molecules are made of atoms? If not, are test questions on chemical digestion really just testing whether students know the relationship between atoms and molecules?

We use the conceptual strand maps published in the *Atlas of Science Literacy* (AAAS 2001; 2007) to make judgments about when students can be expected to know certain ideas along a K–12 learning progression for a particular topic. The *Atlas* map for the topic of natural selection in Figure 12.1, for example, has four strands: changing environments, variation and advantage, inherited characteristics, and artificial selection. The interconnections among the ideas in these strands are visually represented—or mapped—to show the progression of ideas within each conceptual strand through four grade bands as well as the links between ideas across strands. In the *variation and advantage* strand, a benchmark at the 6–8 grade level says: "In all environments…organisms with similar needs may compete with one another for resources, including food, space, water, air, and shelter" (AAAS 2001, p. 83). This is preceded on the *Atlas* map by a benchmark at the 3–5 grade level that says: "For any particular environment, some kinds of plants and animals survive well, some survive less well, and some cannot survive at all" (AAAS 2001, p. 83). In testing whether students know that organisms in ecosystems compete with each other for resources, we would assume that students already know that not all organisms in an ecosystem survive. As a rule, unless we have reason to believe otherwise, we assume that students already know the ideas listed at an earlier grade band and use the ideas and language from those earlier ideas in item development for the grade band that follows.

2. Aligning Assessment Items to Content Standards

We use two criteria to determine whether the content targeted by an assessment item is aligned to the content in a particular key idea. The *necessity* criterion addresses whether the knowledge specified in the learning goal is *needed* to successfully complete the task, and the *sufficiency* criterion addresses whether the knowledge specified in the learning goal is *enough by itself* to successfully complete the task (Stern and Ahlgren 2002). If the targeted knowledge is not needed to answer the question, then the item is obviously not a good indicator of whether or not students know that

Figure 12.1 Conceptual Strand Map for the Topic of Natural Selection from the *Atlas of Science Literacy* (AAAS 2001), p. 83



targeted idea. Also, if additional knowledge is needed, it is difficult to know if an incorrect response is due to students' not knowing the targeted idea or that additional idea. The purpose of such careful alignment is to help reduce errors in interpreting students' correct and incorrect responses. When items are well aligned to the targeted content, students' responses are more likely to provide accurate insights into their understanding of that content. But improving content alignment is not enough. There are other factors that can also affect the validity of an item, and Project 2061's assessment analysis procedure takes those factors into account as well (AAAS 2008).

Improving validity. Test items should be written in such a way that teachers and researchers can draw valid conclusions from them about what students do and do not know about the ideas being tested. Unfortunately, many test items have features that make it difficult to determine if a student's answer choice reflects what the student does and does not know about an idea. When an item is well designed, students should choose the correct answer only when they know an idea, and they should choose an incorrect answer only when they do not know the idea. They should not be able to answer correctly by using test-taking strategies that do not depend on knowing the idea (a false positive response) or be so confused by what is being asked that they choose an incorrect answer even when they know the idea being tested (a false negative response). To improve an item's validity, we identify and eliminate as many problems with comprehensibility and test-wiseness as we can.

3. Using Student Data to Improve 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 indicate that this approach works much more effectively when used in combination with one-on-one interviews with students or pilot tests of items in which students' answer choices are compared to the explanations they give for their answers (DeBoer and Ache 2005).

By comparing the answer choices that students select with their oral or written explanations, we determine if an assessment item is measuring what we want it to measure or if something about the item makes it more likely for students to give false negative or false positive responses to the

item (DeBoer et al. 2007, 2008). In the pilot-testing, students are asked the questions that appear in Figure 12.2. For questions 3 through 6, students are asked to explain why an answer choice is correct or not correct or why they are "not sure."

Figure 12.2 Questions Posed to Students in the Pilot Testing of Test Items

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? Yes No Not Sure
4. Is answer choice B correct? Yes No Not Sure
5. Is answer choice C correct? Yes No Not Sure
6. Is answer choice D correct? Yes No Not Sure
7. Did you guess when you answered the test question? Yes No
8. Please suggest additional answer choices that could be used.
9. Was the picture or graph helpful? If there was no picture or graph, would you like to see one? Yes No
10. Have you studied this topic in school? Yes No Not Sure
11. Have you learned about it somewhere else? (TV, museum visit, etc.)? Where? Yes No Not Sure

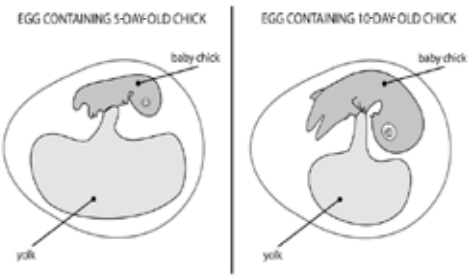
The following six examples illustrate the kinds of information we are able to derive from pilot testing the items. The pilot tests were carried out in both urban and suburban middle and high schools serving a wide range of students. The examples show how we use what we have learned to improve the items' alignment to the key ideas and their validity as measures of student learning.

Example 1: Matter and Energy Transformations in Living Systems

Key Idea: Organisms use molecules from food to make complex molecules that become part of their body structures.

This item was intended to find out if students know that food contains molecules used by organisms to make other molecules that become incorporated into their body structures. The distracters test commonly held misconceptions about what happens to the food that organisms consume.

When a baby chick develops inside an egg, the yolk in the egg is its only source of food. As the chick grows, the yolk becomes smaller. Why does the yolk become smaller?



A. The yolk enters the chick, but none of the yolk becomes part of the chick.
B. The yolk is broken down into simpler substances, some of which become part of the chick.
C. The yolk is completely turned into energy for the chick.
D. The yolk gets smaller to make room for the growing chick.

Pilot testing with middle school students showed that 21.6% of those who were tested got this question correct. Only 4 of the 16 students who chose the correct answer (B) gave any indication that they knew that the yolk becomes incorporated into the chick's body. Those students said that

the yolk "becomes the chick" or "becomes part of the chick," but none of those students used the word *molecule* in their explanation or in any other way suggested that they know that there is a chemical transformation of the yolk into new substances that become part of the chick's body. Other students said the chick "needs" the yolk to grow or said that the yolk "helps" the chick to grow.

Students who chose each answer:

	A	**B**	C	D	Not Sure/Blank	Total
#	8	16	23	20	7	74
%	10.8	21.6	31.1	27.0	9.5	100

Based on these results, the following revisions were considered:

1. Change answer choice B to read: "The yolk gets smaller because some of the atoms of the molecules from the yolk are assembled into new molecules that become part of the chick's body."
2. To more explicitly test if students have the less sophisticated idea that food is *needed* but not that it becomes incorporated into a growing animal's body, include the following distracter: "The yolk gets smaller because some of the molecules from the yolk are used by the chick to live and grow, even though none of the atoms of the molecules from the yolk become part of the chick's body."

These changes should make it more likely that students will not get the test item correct without understanding the idea being tested. In addition, the suggested revisions remove the idea that there is an intermediate stage in which "simpler substances" are formed.

Example 2: Atoms, Molecules, and States of Matter

Key Idea: For any single state of matter, increasing the temperature typically increases the distance between atoms and molecules. Therefore, most substances expand when heated.

This item was written to test whether students know that molecules get farther apart when they are heated and that this molecular behavior explains why most substances expand when heated. The item also includes common misconceptions related to thermal expansion and the behavior of molecules, especially the existence of "heat molecules."

The level of colored alcohol in a thermometer rises when the thermometer is placed in hot water. Why does the level of alcohol rise?

A. The heat molecules push the alcohol molecules upward.
B. The alcohol molecules break down into atoms which take up more space.
C. The alcohol molecules get farther apart so the alcohol takes up more space.
D. The water molecules are pushed into the thermometer and are added to the alcohol molecules.

Pilot testing showed that 25.9% of the students answered this question correctly. The most common response was that heat molecules push the alcohol molecules upward. Pilot testing also revealed that eight of the students were not familiar with the terms *alcohol* or *colored alcohol*, at least not in the context of a thermometer. A number of students wrote comments that confirmed that they held the misconception that there are *heat molecules*.

Students who chose each answer:

	A	B	**C**	D	Not Sure/Blank	Total
#	48	7	28	5	20	108
%	44.4	6.5	25.9	4.6	18.5	100

Based on the results of pilot testing, the following revisions were considered:

1. Because answer choice A is the only one that has the word *heat* in it, students may choose answer choice A because they connect the liquid rising in the thermometer with heat rising. Therefore, add *heat* to one or more of the answer choices.
2. Change the word *alcohol* to *liquid*.

These changes remove a word that students may find confusing in the context of thermometers. The changes also make it less likely that students will be drawn to answer choice A, in which they associate the liquid *rising* with their knowledge that "heat *rises*."

When we interviewed students about this item, we found that they had difficulty reconciling what they expected to be a very small expansion of the liquid in the bulb into what appears to be a very large expansion of the liquid in the narrow tube of the thermometer. One student who knew that substances expand when heated did not believe the liquid could expand that much and, therefore, chose A. Even though her commitment to "heat molecules" did not appear to be strong during the interview, it seemed to her that something besides thermal expansion had to explain such a large increase. Because of developmental issues regarding younger children's ability to engage in proportional reasoning, the thermometer context may be a difficult context for general testing of beginning middle school students' understanding of thermal expansion. But these results also point to an opportunity to provide focused instruction to help students see that a small change in the volume of a liquid is amplified in a narrow tube.

Example 3: Atoms, Molecules, and States of Matter

Key Idea: All atoms are extremely small.

This item was developed to test students' knowledge of the size of atoms. The clarification of this idea says that students should know that atoms are millions of times smaller than other small things such as cells or the width of a hair.

Approximately how many carbon atoms placed next to each other would it take to make a line that would cross this dot? .
A. 6
B. 600
C. 6000
D. 6,000,000

The results of our pilot testing with middle and early high school students showed that 22.8% of the students answered this item correctly. Approximately 25% of the students said they did not know what a carbon atom was and a number of students had difficulty imagining a line of carbon atoms across the dot.

Students who chose each answer:

	A	B	C	**D**	Not Sure/Blank	Total
#	19	23	30	51	101	224
%	8.5	10.3	13.4	22.8	45.1	100

Based on the results of pilot testing, the following revisions were considered:
1. Change "carbon atom" to "atom" so that students who do not yet know about specific atoms will not be disadvantaged.
2. Because imagining a line of carbon atoms across a dot may create an unnecessary cognitive load on students, try to create a simpler context for this item.

Example 4: Force and Motion

Key Idea: If an unbalanced force acts on an object in the direction opposite to its motion, the object will slow down.

This item was developed to test students' understanding of the relationship between forces on an object and the object's motion, focusing specifically on what happens when a force acts in the direction opposite to the object's motion.

A ball is kicked straight up. The ball's speed decreases as it moves upward. Why does the ball's speed decrease?

A. Because the ball gets heavier as it gets farther away from the ground.
B. Because the force of the kick diminishes as the ball moves upward.
C. Because the force of the ball's motion decreases.
D. Because the force of gravity is in the opposite direction to the ball's motion.

Results of pilot testing the item with middle school students showed that 76.9% of the students got this item correct. However, 42% indicated that they were not familiar with the word *diminished*.

Students who chose each answer:

	A	B	C	**D**	Not Sure/Blank	Total
#	1	1	2	20	2	26
%	3.8	3.8	7.7	76.9	7.7	100

Based on the results of pilot testing, the following revision was considered:

Change the wording of answer choice B from "…the force of the kick diminishes as the ball moves upward" to "…the force of the kick runs out as the ball moves upward."

Example 5: Plate Tectonics

Key Idea: The solid outer layer of the Earth—including both the continents and the ocean basins—consists of separate plates.

Two items were piloted to test students' knowledge of the term *bedrock* to help decide if it should be used in assessment items. The two items are identical except that one uses the term *bedrock* and the other uses the descriptive phrase *solid rock*.

Item without the term *bedrock*.
Which of the following are part of Earth's plates?
A. Solid rock of continents but not solid rock of ocean floors.
B. Solid rock of ocean floors but not solid rock of continents.
C. Solid rock of both the ocean floors and the continents.
D. Solid rock of neither the ocean floors nor the continents.

Students who chose each answer:

	A	B	**C**	D	Not Sure/Blank	Total
#	5	3	19	0	6	33
%	15.2	9.1	57.6	0.0	18.2	100

Item with the term *bedrock*.
Which of the following are part of Earth's plates?
A. Bedrock of continents but not bedrock ocean floors.
B. Bedrock of ocean floors but not bedrock of continents.
C. Bedrock of the ocean floors and the continents.
D. Bedrock of neither ocean floors nor continents.

Students who chose each answer:

	A	B	**C**	D	Not Sure/Blank	Total
#	1	2	17	3	11	34
%	2.9	5.9	50.0	8.8	32.4	100

The results show that 50% of the middle school students were able to correctly answer this question when the term *bedrock* was used compared to 57.6% of the students when *solid rock* was used. However, there were

also a greater number of *not sure* responses when *bedrock* was used (32.4% compared to 18.2%). In addition, 32 of the 34 students wrote responses indicating that they did not know what bedrock is. Without understanding the term, students were apparently translating *bedrock* to mean *rock* even though they were not sure what *bedrock* is. Based on the results of pilot testing, it was decided not to include the term *bedrock* in assessing student knowledge about the composition of Earth's plates.

Example 6: Control of Variables

Key Idea: 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.

The following item was developed to determine if students understand that the way to determine if one variable is related to another is to hold all other relevant variables constant. The item was also designed to test a number of common misconceptions that students have regarding the control of variables, including the idea that all of the variables should be allowed to vary in a controlled experiment.

A student wants to test this idea: The heavier a cart is, the greater its speed at the bottom of a ramp. He can use carts with different numbers of blocks and ramps with different heights.
Which three trials should he compare?

Students who chose each answer:

	A	B	C	**D**	Not Sure/Blank	Total
#	20	1	6	41	8	76
%	26.3	1.3	7.9	53.9	10.5	100

The results of pilot testing the item showed that 53.9% of the students answered correctly. In addition 26.3% chose answer choice A, which targets the misconception that both variables should vary at the same time. Answer choices B and C, however, were less successful distracters. Answer choice B was chosen by only one student. Of the six students who chose C, three students said they rejected answer choices A and B because there were no weights in one of the carts for those answer choices. Also, three students thought the word *trials* in the stem referred to the answer choices and circled three answer choices as correct. Six students (even some of those who chose the correct answer) thought that the word *blocks* in the stem referred to the parts of the ramp rather than the weights in the cart.

Based on the results of pilot testing, the following revisions were considered:
1. Replace the blocks with metal balls, and increase the number of balls in each cart by 1 so that there are no empty carts.
2. Replace answer choice B with three carts with the same weights on different height ramps.
3. Replace the question in the stem with: "Which 3 sets of experiments should the student do?" Label the 3 images for each answer choice "Experiment 1," "Experiment 2," and "Experiment 3."

Conclusion

As the examples above illustrate, many factors can affect how well a test item measures what students know about a particular idea. In our work, we have analyzed hundreds of items covering more than a dozen science topics. The items have come from a wide range of sources including international, national, and state tests; curriculum materials; and a variety of item banks. Nearly all of these items had problems that compromised how well they measure what students know. Most item developers depend on sophisticated quantitative analyses to judge the suitability of their test items. The

items are selected if they produce reliable results and discriminate among categories of test takers. This approach is inadequate, according to the National Mathematics Advisory Panel (2008), which called for items that are also designed to measure "specified constructs" as a way to reduce the number of flawed items (p. 61). This recommendation makes sense in science as well, and we offer intensive qualitative analysis described in this chapter, along with more quantitative approaches, as a way to produce items that are effective measures of what we want to measure.

References

American Association for the Advancement of Science (AAAS). 1989. *Science for all Americans*. New York: Oxford University Press.

American Association for the Advancement of Science (AAAS). 1993. *Benchmarks for science literacy*. New York: Oxford University Press.

American Association for the Advancement of Science (AAAS). 2001. *Atlas of science literacy*. Washington, DC: AAAS and NSTA.

American Association for the Advancement of Science (AAAS). 2007. *Atlas of science literacy, Vol. 2*. Washington, DC: AAAS and NSTA.

American Association for the Advancement of Science (AAAS). 2008. Project 2061's approach to assessment alignment: Assessment analysis utility. Retrieved April 12, 2006, from *www.project2061.org/research/assessment/assessment_form.htm*.

American Federation of Teachers. 2006. Smart testing: Let's get it right (Policy Brief No. 19). Washington, DC: American Federation of Teachers (July).

DeBoer, G. E. 2007. Developing assessment items aligned to middle school science learning goals. Presented at the Knowledge Sharing Institute of the Center for Curriculum Materials in Science. Washington, DC, July 22–25.

DeBoer, G. E., and P. Ache. 2005. Aligning assessment to content standards: Applying the Project 2061 analysis procedure to assessment items in school mathematics. Paper presented at the meeting of the American Educational Research Association, Montreal, Canada (April). Retrieved September 29, 2006, from *www.project2061.org/research/assessment/aera2005.htm*

DeBoer, G. E., N. Dubois, C. Herrmann Abell, and K. Lennon. 2008. Assessment linked to middle school science learning goals: Using pilot testing in item development. Paper presented at the annual conference of the National Association for Research in Science Teaching, Baltimore, MD (March).

DeBoer, G. E., C. Herrmann Abell, and A. Gogos. 2007. Assessment linked to science learning goals: Probing student thinking during item development. Paper presented at the annual meeting of the National Association for Research in Science Teaching, New Orleans (April).

National Mathematics Advisory Panel. 2008. *Foundations for success: The final report of the National Mathematics Advisory Panel*. Washington, DC: U.S. Department of Education.

National Research Council (NRC). 1996. *National science education standards*. Washington, DC: National Academy Press.

National Research Council (NRC). 2001. *Classroom assessment and the* national science education standards. Washington, DC: National Academy Press.

National Research Council (NRC). 2006. *Systems for state science assessments*. Washington, DC: National Academies Press.

Stern, L., and A. Ahlgren. 2002. Analysis of students' assessments in middle school curriculum materials: Aiming precisely at benchmarks and standards. *Journal of Research in Science Teaching* 39(9): 999–910.