Assessment Linked to Science Learning Goals: Probing Student Thinking During Item Development

National Association for Research on Science Teaching
Annual Conference, New Orleans, LA April 15-18, 2007

George E. DeBoer, Cari Herrmann Abell, and Arhonda Gogos, AAAS/Project 2061

Standards-based reform is built on the idea that all parts of the K-12 educational system—including curriculum, instruction, and assessment, as well as pre- and in-service teacher education—will have their greatest impact when built around a set of agreed upon learning goals that specify what students should know and be able to do to achieve the goal of science literacy by the time they leave school. The importance of learning goals is described in the policy statement of the Center for Curriculum Materials in Science: "[A focus on learning goals] allows all parts of the system to be rationally connected and provides a basis for curriculum coherence within and across grades. Aligning all parts of the system to learning goals fosters the development of instructional tools and resources, educational experiences for teachers, research studies, and policies that are focused on the same important ends. Without this focus, it will be difficult for all segments of the education community to work toward common purposes while at the same time ensuring the variation needed to help all students succeed" (CCMS, 2004).

In a standards-based environment, assessment is used to provide accurate information on how well students are meeting pre-established learning goals, but assessment must also provide guidance regarding what students already know, what they are capable of learning, and the appropriateness of the learning goals themselves. Besides measuring outcomes, assessment is also a tool in the definition and modification of the learning goals.

This symposium addresses how multiple choice assessment items can be developed that precisely measure student understanding of the ideas specified in the AAAS *Benchmarks for Science Literacy* (BSL) and the NRC's *National Science Education Standards* (NSES). It focuses particularly on how the pilot testing of multiple choice assessment items, during which students are asked to provide written comments about the test items themselves, can offer useful information about what students do and do not know and the appropriateness of the target learning goals for students at the middle school level.

Targeting the Learning Goals

Benchmarks and *Standards* as a Starting Point

In our assessment work, the national science standards documents—*Benchmarks for Science Literacy* (AAAS, 1993) and the *National Science Education Standards* (NRC, 1996)—are the source of the science ideas that we target. *Benchmarks* and *NSES* were organized around ideas that students should know at various grade bands if they are to achieve the goal of science literacy by the time they graduate from high school. The ideas in these documents, and the organization of these ideas, provide guidance for

instruction, curriculum development, and assessment. To help users make connections between ideas, each section of *Benchmarks* refers the reader to other chapters and sections where related ideas are addressed.

Teasing out Key Ideas

In order to achieve the level of precision we desire in the definition of learning goals and in the diagnosis of learning difficulties, we begin with the statements in *Benchmarks* and *Standards*, but then we create an even finer-grained subdivision of knowledge. Here are some examples of how we derive key ideas from the benchmarks and standards:

From Plate Tectonics:

- The outer layer of the earth—including both the continents and the ocean basins—consists of separate plates (BSL 4C/M11**). Note: Benchmarks with ** are new benchmarks developed during the preparation of *Atlas of Science Literacy 2* (AAAS, 2007).
- The plates move very slowly, pressing against one another in some places, pulling apart in other places (BSL 4C/M12**).

From Chemical Reactions

- A substance has characteristic properties, such as density, a boiling point, and solubility, all of which are independent of the amount of the sample and can be used to identify the substance (From NSES 5-8B:A1a).
- Since different substances have different characteristic properties, a mixture of substances can often be separated into the original substances using one or more of these characteristic properties (From NSES 5-8B:A1b).

From Laws of Motion

- If a force acts on an object in the direction opposite to its direction of motion, the object will slow down (From BSL 4F/M3a).

Clarifying the Key Ideas

We clarify each key idea by indicating what it is that we expect students to know for assessment purposes. Consider the key idea:

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

Clearly, there are concepts in this statement that need to be elaborated. What, exactly, do we expect students to know about earth's plates? The clarification statement that we developed says:

> *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.

The writing of the clarification statement is guided by three questions: (1) Is the description of plates that is specified what is needed for students to form a mental image that allows them to predict and explain phenomena involving plates? (2) Is the description of plates that is specified what is 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 it 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 they are not all the same size and shape, and that separate plates fit tightly together. These elaborations of the term "plate" are also intended to guide instruction that will lead to a mental model of a plate that can be used to understand subsequent ideas about plate motion and the consequences of plate motion, which come later in the learning sequence. This mental model should help students understand such things as mountain building and where earthquakes and volcanoes form when they are introduced to those ideas. With respect to terminology, we also decided not to expect students to know the term lithosphere because we did not think that it would contribute significantly to explaining phenomena related to plate motion.

Misconceptions Related to the Learning Goals

Research on student learning is important in our work because it provides ideas for plausible distractors in test items, which we use to validate earlier research on misconceptions. For each of the topics we are working on, we have identified ideas that students have about those topics and the variety of ways that students think about the phenomena related to those ideas. This also helps us determine whether the complexity of the mental models we are aiming toward is age-appropriate. Although students may be capable of developing very sophisticated ideas at any age, the time required to do so at a young age may argue for waiting until later to try to develop certain ideas.

Part I: How we Ensure Alignment of Items to Content Standards

Applying the Criteria of Necessity and Sufficiency

For all of the assessment items that we develop, two criteria are used to determine if the item is content aligned to the learning goal: The *necessity* criterion addresses whether the learning goal is needed to complete the task and the *sufficiency* criterion addresses whether the learning goal is enough by itself to complete the task. For the necessity criterion, expert reviewers determine separately for each answer choice if the knowledge specified in the learning goal or its clarification statement is needed to evaluate the truth and relevance of that answer choice. (The correct statement of a relevant misconception is also considered to be part of a learning goal.) In judging whether the necessity criterion is met, reviewers focus on whether the targeted content knowledge (as opposed to some other content knowledge) is what is needed to answer correctly. They do not consider whether test-wiseness or general cognitive abilities instead of the targeted content can be used to determine the correct answer. The focus here is strictly on content knowledge. For sufficiency, reviewers determine separately for each answer choice if any more knowledge than is specified in the learning goal and its clarification statement is required to complete the task. Knowledge that can be considered within the scope of general knowledge for students of this age is not considered to be additional knowledge even when it is not specified in the learning goal (e.g., the number of hours in a day, days in a week, etc.). What can and cannot be taken for granted as general knowledge and abilities for middle school students is often revealed during interviews and pilot testing.

In summary, clearly stated learning goals whose boundaries have been carefully defined in clarification statements provide an exact definition of the construct being measured. The criteria of necessity and sufficiency can then be applied with precision to ensure alignment of the knowledge specified in the learning goal to the knowledge required by the test item. Validity is improved by having a very clear statement of what is included and what is not included in the construct being measured. This also allows us to identify a spread of items that define the construct.

Example 1: Analyzing *Necessity* and *Sufficiency* for an Item from the Topic of Atoms, Molecules, and States of Matter

Key Idea: <u>*All atoms are extremely small*</u> (from BSL 4D/M1a).

Clarification Statement:

Students are expected to know that atoms are much smaller than very small items with which they are familiar, such as dust, blood cells, plant cells, and microorganisms, all of which are made up of atoms. Students should know that the atoms are so small that many millions of them make up these small items with which they are familiar. They should know that this is true for all atoms. The comparison with very small objects can be used to test students' qualitative understanding of the size of atoms in relation to these objects. Students will not, however, be expected to know the actual size of atoms.

Which of the following is the smallest?

A. An atom
B. A bacterium
C. The width of a hair
D. A cell in your body

AAAS PROJECT 2061 COPYRIGHT © 2006

The item meets the *necessity* criterion. The knowledge in the learning goal is needed to answer correctly. There is no other content that could substitute for the ideas specified in the learning goal.

The *sufficiency* criterion is not met. Students also need to know the term "bacterium," which is additional knowledge. In pilot testing, 25% of 193 students indicated that they did not know what a bacterium was (even though most said they knew what bacteria were). The item should say "microorganism" or "bacteria" to match the clarification statement and what is generally understood terminology for students of this age. Although "hair" is not explicitly mentioned in the clarification statement, it is a very small item with which students are familiar.

Example 2: Analyzing *Necessity* and *Sufficiency* for an item from the topic of Substances, Chemical Reactions, and Conservation

Key Idea: *Substances may react chemically in characteristic ways with other substances to form new substances with different characteristic properties* (From NSES 5-8B:A2a).

Clarification Statement:

This idea deals with chemical reactions at a macroscopic level. Students should know that most substances can react chemically with some other substance and that during a chemical reaction one or more substances change into one or more new substances. They are expected to know that the original substances in a chemical reaction are called reactants and the resulting substances are called products. They should know that the products can be identified as new substances because they have different characteristic properties from the original substances. They should know that chemical reactions can occur between liquids, solids, or gases and that they can also occur when substances are heated or when electrical energy is added to the original substances. They should know that some chemical reactions occur very rapidly and dramatically while others occur much more slowly over longer periods of time.

Misconceptions:

- A chemical reaction occurs during a change of state.
- A chemical reaction occurs when a substance dissolves.

Which of the following is an example of a chemical reaction?

A. A piece of metal hammered into a tree
B. A pot of water being heated and the water evaporates
C. A spoonful of sugar dissolving in a glass of water
D. An iron railing developing an orange, powdery surface after standing in air

AAAS PROJECT 2061 COPYRIGHT © 2006

The *necessity* criterion is not met. The knowledge in the learning goal is not needed. Answer choice D, the correct answer, is a specific instance of a general principle (SIGP). The student can get the item correct by knowing that rusting is a chemical reaction without knowing the general principle that new substances are being formed that have different characteristic properties.

The *sufficiency* criterion is met. The knowledge in the learning goal is sufficient to answer correctly. Phase change and dissolving are addressed in the misconceptions. Hammering a piece of metal is common knowledge. Changing properties of reacting materials is addressed in the learning goal.

Ensuring Construct Validity

Test items often contain construct irrelevant features that make it difficult to have faith in the answers that students choose. Comprehensibility and test-wiseness issues are some of the most common threats to construct validity. Students should choose the correct answer when they know the idea and they should choose an incorrect answer when they do not know the idea, not use test taking strategies that do not depend on knowing the content or be so confused by what is being asked that they choose the incorrect answer even when they know the idea being tested. Our goal is to eliminate as many factors as possible that are not related to the knowledge being measured (construct irrelevant factors) so that the number of false negative and false positive responses is minimized. This ensures a more accurate measure of student knowledge.

There are three criteria that we apply to the design of test items to improve construct validity. These involve comprehensibility, test-wiseness, and appropriateness of task context.

Analyzing Comprehensibility: *To make valid interpretations of what student know, a test question must be understandable to students.*

To ensure that valid interpretations of student knowledge can be made, we look for a number of threats to comprehensibility:

1. It is not clear what question is being asked.
2. The task uses unfamiliar general vocabulary that is not clearly defined. (Note: This is referring to general language usage, not technical scientific or mathematical terminology, which is addressed under *sufficiency*.)
3. The task uses unnecessarily complex sentence structure or ambiguous punctuation that makes the task difficult to comprehend when plain language could have been used.
4. The task uses words and phrases that have unclear, confusing, or ambiguous meanings. This may include commonly used words that have special meaning in the context of science. For example the word "finding" could be unfamiliar to students when referring to a scientific "finding."
5. There is inaccurate information (including information in the diagrams and data tables) that may be confusing to students who have a correct understanding of the science.
6. The diagrams, graphs, and data tables may not be clear or comprehensible. (For example, they may include extraneous information, inaccurate or incomplete labeling, inappropriate size or relative size of objects, etc.).

Example 3: **Analyzing *Comprehensibility* for an item from the topic of Atoms, Molecules, and States of Matter**

Key Idea: *For any single state of matter, a change in temperature typically changes the average distance between atoms or molecules. Most substances or mixtures of substances expand when heated and contract when cooled* (from BSL 4D/M3b).

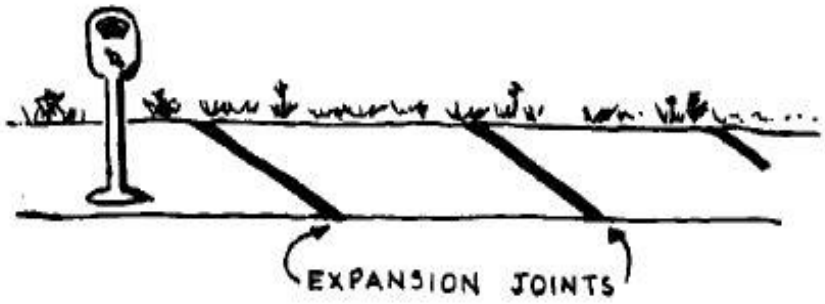
Clarification Statement:

Students should know that as the temperature of a substance increases the average distance between the atoms/molecules of the substance typically increases, causing the substance to expand. Students should also know that as the temperature of a substance decreases the average distance between the atoms/molecules typically decreases, causing the substance to contract. Students are expected to know that this expansion or contraction can happen to solids, liquids, and gases. They are also expected to know that expansion or contraction due to changes in temperature can also happen to mixtures of substances. Students are expected to know that expansion or contraction due to changes in temperature is not permanent (e.g., objects that expand when heated then contract when cooled). They are expected to know that the number of atoms and the mass of the atoms do not change with changes in temperature. Students should also know that different substances expand and contract differently.

Students are not expected to know the details of the relationship between the speed of the atoms or molecules and thermal expansion. Students are also not expected to know the substances that violate this rule and shrink when heated or that water will

shrink when heated anywhere between 0ºC and 4ºC. Students are not expected to know or apply gas law equations. Because the definition of the size of an atom is varied and complex, we only expect students to know that the size of an atom or molecule does not decrease when the temperature increases or the size does not increase when temperature decreases.

Most sidewalks made out of concrete have cracks every few yards as shown in the diagram below. These are called expansion joints as labeled in the diagram below. What happens to the width of the cracks during a hot day in the summer and why?



A. The cracks get wider because the concrete shrinks.
B. The cracks get wider because the concrete gets softer.
C. The cracks get narrower because the concrete expands.
D. The cracks get narrower because the ground underneath the sidewalk shrinks.

AAAS Project 2061 Copyright © 2006

The item contains two inaccuracies and a term students may not be familiar with. The spaces between the sections of concrete are not "cracks" and they do not appear "every few yards." The word "expansion joint" may not be familiar to students and it is used in the stem and illustration but not in the answer choices. Although these may be minor issues to someone who knows the idea being tested, they could potentially interfere with students whose knowledge is uncertain.

Analyzing Appropriateness of Task Context: A task context sets up situations or stories involving people, objects, or events in real-world occurrences or under idealized conditions. Any context should be understandable, interesting, and sensible to students. As much as possible, one group of students should not be advantaged or disadvantaged relative to the other students. A task context should be chosen that can reasonably be assumed to be familiar to most if not all students. Although judgments about the suitability of task contexts are difficult in the absence of knowledge of students' particular life experiences, we avoid using contexts that we think may be inappropriate. In addition, all situations that are described should include plausible quantities and

dimensions and should be accurate and credible. Students should not have to spend time wondering if there is a mistake in the question. Task contexts should be scientifically and mathematically accurate, and idealizations (such as a frictionless world) should be clearly noted.

In order to make valid interpretations of what student know from test questions that use real-world or idealized contexts, those contexts must be accessible to students.

To ensure fairness, we look for the following ways in which a task may not be accessible to students:

1. The context may be unfamiliar to students.
2. The context may advantage or disadvantage one group of students because of their interest or familiarity with the context.
3. The context is complicated and not easy to understand so that students might have to spend a lot of time trying to figure out what the context means.
4. The information and quantities that are used are not reasonable or believable.
5. The context does not accurately represent scientific or mathematical realities or make clear when idealizations are involved.

Analyzing Test-Wiseness: A major threat to construct validity is test-wiseness. This criterion addresses characteristics of the task that might (1) allow students to make a satisfactory response using only general test-taking skills without understanding the idea being tested or (2) mislead students into choosing an incorrect answer. For example, in multiple choice questions, distractors should be plausible to students and one answer (especially the correct answer) should not be distinctly different from the other answers. One answer choice should not be significantly longer or more elaborate. It should not be more qualified than the others, using terms such as "always," "never," or "everyone." It is also important to pay attention to the use of logical opposites that may make it easy to eliminate answer choices. Lists of common student misconceptions (and interview and pilot test data, if available) can be used to identify more plausible distractors.

To make valid interpretations of what student know, a test question must not be answerable by using test-taking strategies that do not depend on knowing the ideas being tested.

To ensure that valid interpretations can be made of students' answer choice selections, we look for a number of common test-wiseness issues*:*

1. Some of the distractors are not plausible.
2. One of the answer choices differs in length or contains a different amount of detail from the other answer choices.
3. One of the answer choices is qualified differently from the other answer choices, using words such as "usually" or "sometimes," or an answer choice uses different units of measurement.
4. The use of logical opposites may lead students to eliminate answer choices.

5. One of the answer choices contains vocabulary at a different level of difficulty from the other answer choices that may make it sound more scientific.
6. The language in one of the answer choices mirrors the language in the stem.

Example 4: Analyzing *Test-Wiseness* for an item from the Topic of Matter and Energy Transformations in Living Systems:

Key Idea: *Food is a source of molecules that serve as fuel and building material for all organisms*.

Clarification:

Students are expected to know that food consists of carbon-containing molecules in which carbon atoms are linked to other carbon atoms in complex molecules such as carbohydrates (including simple sugars), fats, and proteins. They are also expected to know that other molecules (besides carbohydrates, fats, and proteins) that have carbon atoms linked together might also be food for organisms, but they are not expected to know which carbon-containing molecules are or are not food for any particular organism. They are expected to know that these carbon-containing molecules serve as building material that organisms use for growth, repair, and replacement of body parts (such as leaves, stems, roots, bones, skin, muscles, cells, and parts of cells) and provide the chemical energy needed to carry out various life functions. They are expected to know that substances that do not provide both chemical energy and building material are not food. They are not expected to know what chemical energy is other than that it resides in the molecules of substances. They are expected to know that light is not food because it does not provide building material and they are expected to know that even though, water, carbon dioxide oxygen, minerals, and vitamins may provide some building materials, they are not food because, they cannot be used as fuel. The idea that minerals and vitamins are essential for growth and repair of body parts and that they are present in small amounts in food is included in Benchmark 6E/E1.

Misconceptions related to this idea:

- Many children associate the word *food* with anything that they identify as being edible .
- Students see food as substances (water, air, minerals, etc.) that organisms take [directly] in from their environment.
- Some students think that food is anything that is needed to keep animals and plants alive.

Is the oxygen that animals breathe a kind of food?

A. Yes, because oxygen enters the body.
B. Yes, because all animals need oxygen to survive.

C. **No, because animals do not get energy from oxygen.**
D. No, because oxygen can enter an animal's body through its nose.

AAAS Project 2061 Copyright © 2006

To most students, answer choice D (No, because oxygen can enter an animal's body through its nose) is probably not a plausible explanation for why oxygen is not food. In pilot testing, 5 of 29 students selected this answer choice, which is intended to test the misconception that the point of entry is what determines if something is food, but many other students questioned how the nose was relevant in a question about food and eliminated it on that basis. Answer choice D could be revised and still test the misconception by changing it to say that oxygen is not food because it is not edible or because it does not enter through an animal's mouth.

Part II: Using Student Data to Inform the Design of Assessment Items

Applying a set of criteria to determine the alignment of test items to learning goals and to seek out threats to construct validity is an important part of developing items that accurately measure the knowledge we want students to have. But we have learned from studies we have conducted that this approach works much more effectively when used in combination with interviewing and pilot testing in which students' answer choices are compared to the reasons they give for their answers (DeBoer and Ache, 2005). Pilot testing and interviewing are an important part of our item development work and are also used to inform revisions to the learning goals and clarification statements.

1. We use pilot testing and interviewing to probe student thinking about the targeted ideas and the test items.
2. We compare student answer choices to their explanations.
3. When answer selections and explanations don't match, we look for structural problems with the item that could produce these mismatches.

In the spring of 2006, pilot testing was conducted in 112 classrooms across five content areas involving about 2700 students: Atoms and Molecules (726 students); Force and Motion (610 students); Flow of Matter and Energy (312 students); Plate Tectonics (568 students); Control of Variables (462 students). Pilot testing was conducted in schools from different parts of the country having varying demographic characteristics: (1) A middle school and a high school in a northeastern suburb having a student population that is 40% White, 48% African American, and 8% Hispanic, with 25% of the students eligible for Free and Reduced Lunch. (2) A middle school in a northeastern suburb having a student body that 95% White, where 10% of the students are eligible for Free and Reduced Lunch. (3) A K-8 school in a rural northeast community having a student body that is 98% White, in which 49% of the students are eligible for Free and Reduced Lunch. (4) A middle school in a small southern town with a student body that is 70% White and 24% African American, where 33% of the students are identified as economically disadvantaged, and (5) a middle school in a small southwestern town,

having a student body that is 95% Hispanic, and in which 95% of the students are eligible for Free and Reduced Lunch.

By comparing the answer choices that students select with their oral or written explanations, we can determine if an item is measuring what we want it to measure or if it is likely to produce false negative or false positive responses on the part of students. Students also point out words that they do not understand, as well as confusing language that we can correct during the item revision process. During pilot testing, students are asked the following questions. For questions 3-6, students are asked to explain why an answer choice is correct or not.

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?
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

What we learn from pilot testing: Examples from Plate Tectonics.

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

Clarification Statement: (See page 2 for a clarification statement for this key idea.)

Example 5: Plate Tectonics

Original Item

Which of the following is TRUE about the earth?

A. It has one plate.
B. It has seven plates.
C. It has about fifteen plates.
D. It has over a hundred plates.

AAAS Project 2061 Copyright © 2006

Is Answer Choice Correct?	A	B	C	D	% Correct (n=61)
Yes	2	12	24	8	37.7%
No	51	34	18	42	
Not Sure	6	13	18	9	
No Response	2	1	1	2	
Multiple selections for single answer choice	0	1	0	0	

Revised Item

Which of the following is TRUE about the earth?

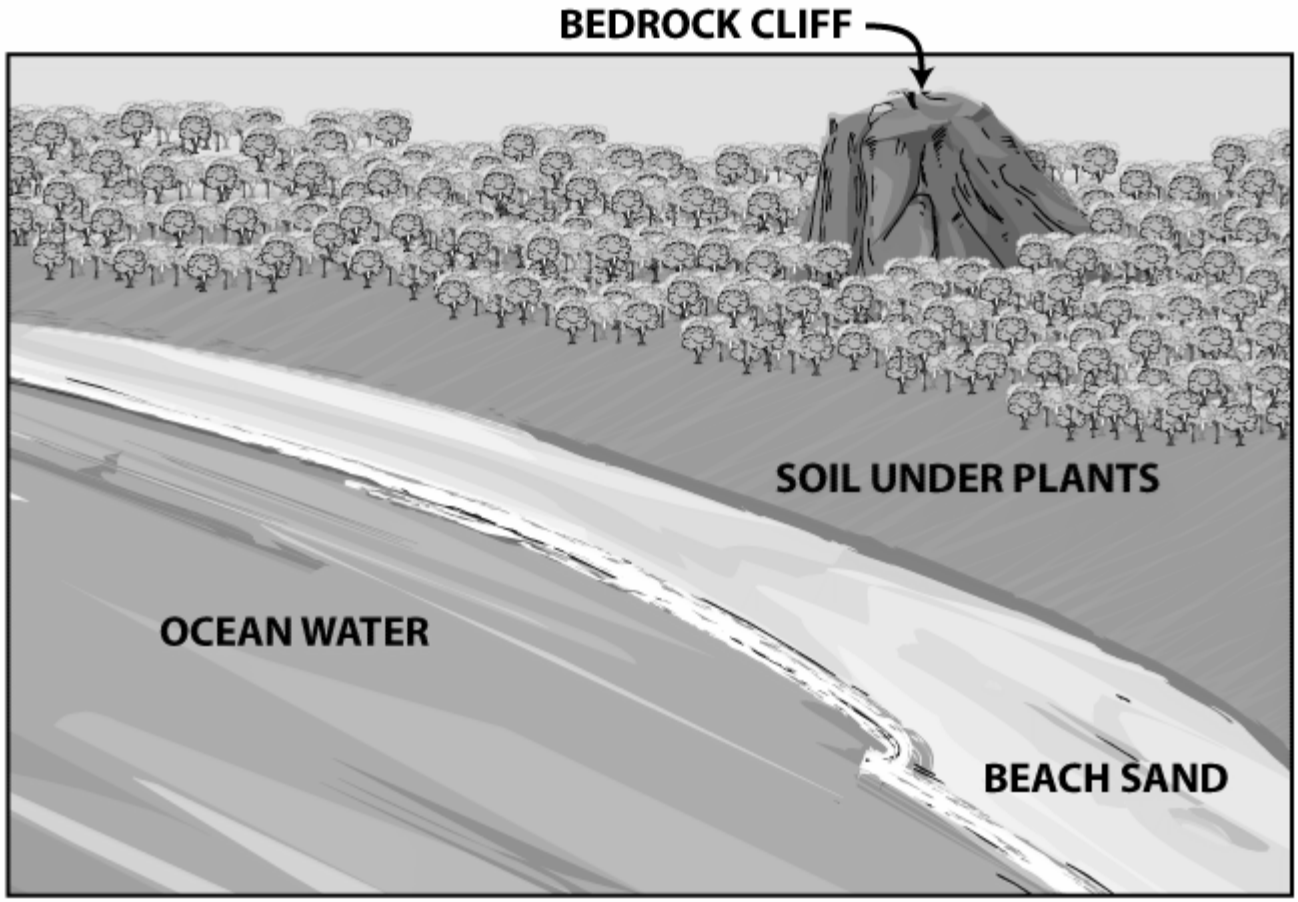
A. It has one very large plate that covers the entire earth.
B. It has seven very large plates, one for each continent.
C. It has about twelve to fifteen very large plates.
D. It has over a hundred very large plates.

AAAS Project 2061 Copyright © 2006

Reasons for revisions: Some students said that they had learned there are 12 major plates and some that there are 15, so we included an answer choice with a 12-15 range. Also, some students said there are seven plates because there are seven continents, and they equated plates with continents. Because we wanted to explicitly test this link between number of continents and number of plates, we included a statement to that effect in answer choice B. Also, D in the original item could be correct if very small plates are counted.

Example 6: Plate Tectonics

Original Item:



Which of the things in the picture above is material that the earth's plates are made of?

A. Beach sand
B. Bedrock cliff
C. Everything that can be seen
D. The plate material cannot be seen.

AAAS Project 2061 Copyright © 2006

Is Answer Choice Correct?	A	B	C	D	% Correct (n=51)
Yes	5	14	13	8	23.5%
No	32	23	26	27	
Not Sure	12	13	10	14	
No Response	2	1	2	2	
Multiple selections for single answer choice	0	0	0	0	

Revised Item:

Is any part of one of earth's plates visible in the picture?

A. Yes. Everything in the picture is part of a plate.
B. Yes. The solid rock of the mountains and cliffs is part of a plate.
C. No. The plates are deep within the earth and can never be seen.
D. No. The plates are sometimes visible but they cannot be seen in this picture.

AAAS Project 2061 Copyright © 2006

Reasons for revisions: In the original item, students were asked about the material that earth's plates are *made of*. Students' comments suggested that they may have been thinking about the materials that make up the plates, i.e., their raw materials. A number of them thought that all of what they see in the picture eventually ends up being part of plates. The item was changed to focus more on the idea that plates can be seen where they are not covered by water, sand, and soil.

Example 7: Plate Tectonics

Original Item:

How thick are earth's plates?

A. Several inches thick
B. Several feet thick
C. Many miles thick
D. Thousands of miles thick

AAAS Project 2061 Copyright © 2006

Is Answer Choice Correct?	A	B	C	D	% Correct (n=63)
Yes	9	5	26	10	39.7%
No	40	39	17	33	
Not Sure	12	17	19	17	
No Response	1	1	0	2	
Multiple selections for single answer choice	1	1	1	1	

Revised Item:

Which of the following is a reasonable thickness of one of earth's plates?

A. Six inches
B. Six feet
C. Sixty feet
D. Sixty miles

AAAS Project 2061 Copyright © 2006

Reasons for revisions: Although most students apparently had no problem differentiating between "many miles thick" and "thousands of miles thick" in the original item, those answer choices are too similar and, in a sense, are saying the same thing. We changed the item to be more specific and to provide four clear choices.

What we Learn from Pilot Testing: An example from Controlling Variables:

Of four types of questions we use in assessing student understanding of control of variables, one is more difficult than the others. For Type IV items (see below), only about 19% of students overall answer correctly, compared to 46.5% for Type I items, 62.6% for Type II items, and 59.9% for Type II items. These percentages are based on pilot testing results of approximately 450 students.

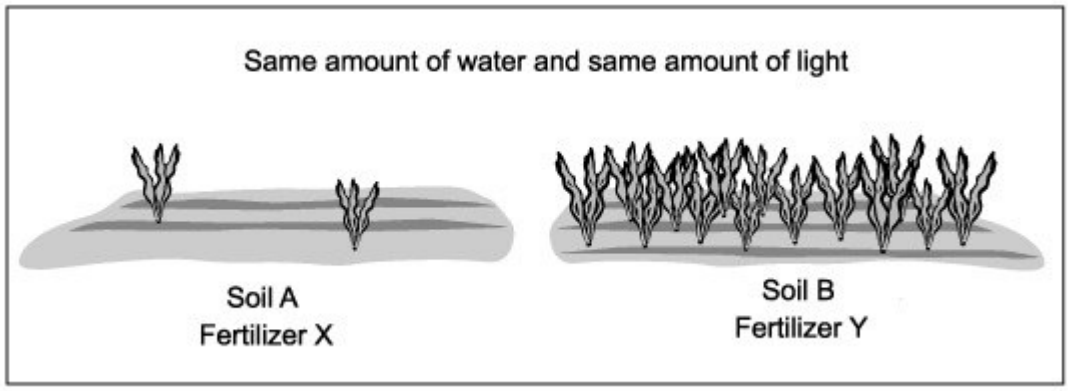
I. Given an idea to be tested (hypothesis) and an experimental setup explain *why* certain variables are (or should be) kept constant. All relevant variables are provided.
II. Select an experimental setup to test the effect of a variable on the experimental outcome, when all relevant variables are provided (i.e. which variables from a defined set should be allowed to vary and which should be kept constant).
III. Select an idea (hypothesis) that could be tested by using a certain controlled experimental setup (one independent variable is allowed to change and two are held constant).
IV. Given the results of an experiment that involves two variables changing at the same time, determine what conclusion can be drawn regarding the effect of each variable on the experimental outcome. Students are expected to know that under these conditions no conclusion can be drawn about the effect of each variable.

To find out why fewer students answered Type IV items correctly, we looked closely at the comments they made and the answer choices they selected.

Original Item:

A gardener wants to find out which type of soil and which brand of fertilizer are best for his vegetables.

He does the following experiment:



What did the gardener learn from this experiment about the quality of the different soils and fertilizers?

A. Soil B is the best soil for his vegetables.
B. Fertilizer Y is the best fertilizer for his vegetables.
C. Soil B is the best soil and Fertilizer Y is the best fertilizer for his vegetables.
D. This experiment does not tell the gardener which soil is best or which fertilizer is best for his vegetables.

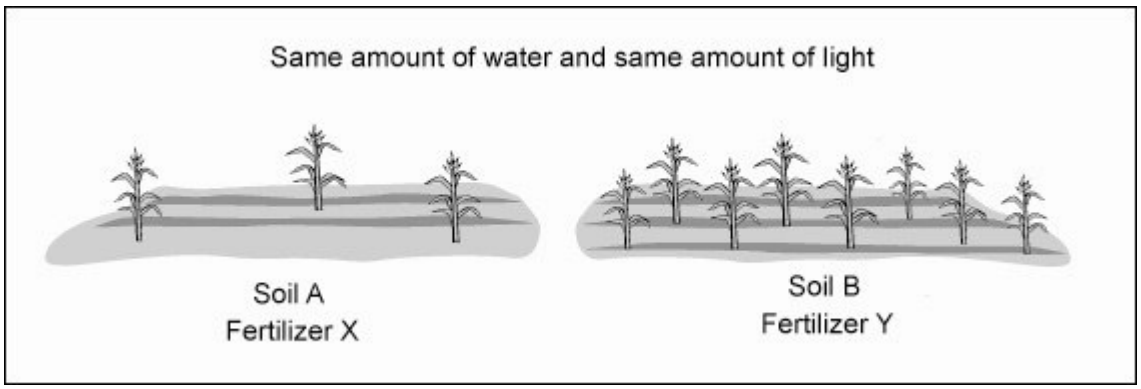
AAAS Project 2061 Copyright © 2006

Is Answer Choice Correct?	A	B	C	D	% Correct (n=109)
Yes	19	20	75	15	11.0%
No	81	80	26	84	
Not Sure	8	6	6	6	
No Response	1	3	2	4	
Multiple selections for single answer choice	0	0	0	0	

Revised Item:

A farmer wants to find out which type of soil is best for growing his corn. He also wants to find out which type of fertilizer is best for growing his corn.

He does the following experiment using two different types of soil and two different types of fertilizer:



What can the farmer conclude from this experiment?

A. He can conclude that Soil B is the best soil for growing his corn.
B. He can conclude that Fertilizer Y is the best fertilizer for growing his corn.
C. He can conclude that Soil B is the best soil for growing his corn and that Fertilizer Y is the best fertilizer for growing his corn.
D. It is not possible to conclude from this experiment which soil is best for growing his corn or which fertilizer is best for growing his corn.

AAAS Project 2061 Copyright © 2006

Reasons for Revisions: We learned from pilot testing that answer choice C was by far the most popular answer choice (75/109 students chose this incorrect answer). Students apparently had difficulty thinking about soil and fertilizer separately and combined the two variables into a single variable instead. In written comments, students said that Soil B and Fertilizer Y produced the best combination. In the revised item, we were much more explicit about the distinct nature of the two variables and that the farmer was interested in learning about both of them.

Summary of the AAAS Project 2061 Item Development Process

This paper describes various aspects of a multi-stage item development process that is intended to produce assessment items that clearly target the ideas specified in national standards documents. The steps in the process are summarized in the box below. Following this process enables us to create test items that are precisely aligned with content standards and which have a high degree of construct validity.

Steps in the AAAS Project 2061 Item Development Process

1. Select a set of benchmarks and standards to define the boundaries of a topic (e.g. chemical reactions, interdependence of life, etc.).
2. Tease apart the benchmarks and standards into a set of key ideas (e.g. *Substances may react chemically in characteristic ways with other substances to form new substances with different characteristic properties*).
3. Create an assessment map showing how the key ideas build on each other conceptually.
4. Review the research on student learning to identify ideas students may have about the topic.
5. Design items:
    a. using student misconceptions as distractors
    b. using our assessment analysis criteria
    c. following a list of design specifications
6. Use open-ended interviewing to supplement published research on student learning
7. Use "item camps" to get feedback on items from staff
8. Revise items
9. Pilot test items and conduct think aloud interviews
10. Analyze pilot test data
11. Revise items and expectations for students (clarification statements)
12. Conduct formal reviews of items using the assessment analysis criteria
13. Revise items
14. Conduct national field test of items

Discussion and Future Directions

The goal of this assessment work is to develop multiple choice test items that are closely aligned to the ideas in national standards documents. In addition, all of the items include common misconceptions as distractors. The item development process involves precisely specifying each idea being targeted, elaborating that idea in a clarification statement that draws upon existing research on student learning to determine the appropriateness of our expectations for middle school students, writing items to predetermined item development specifications to maximize construct validity, obtaining feedback on the items from students through pilot testing and think-aloud interviews, having items formally reviewed by content and science education experts, and field testing items on a national sample to determine various psychometric properties of items and of clusters of items. The field test data provide us with difficulty scores, item discrimination indices, and differential item functioning scores. It also allows us to model student understanding of science ideas at both the key idea level and topic level using item response modeling (IRM).

These items and clusters of items will serve a number of purposes. First, they will provide item developers and users of test items with models of how items can be precisely aligned to learning goals at the idea level, not just at the topic level. Many items that are claimed to be aligned to a content standard use less rigorous criteria to assess alignment. The items can be used diagnostically by individual teachers and school or district level personnel to measure student understanding of specific ideas and to determine gaps in student understanding of those ideas. Many of the assessment items that are being developed expect students to use knowledge to explain and predict phenomena that they may not have encountered before in school. By embedding the items in real-world contexts that are accessible to students but different from those commonly used in textbooks or classroom lessons, these items enable teachers to gauge more precisely their students' knowledge of the targeted science *ideas*. Because these assessment items are aligned to the ideas in the content standards and not to any particular curriculum material, they require students to demonstrate their understanding of those important science ideas rather than merely repeating words they heard in class. The items also provide diagnostic information to help teachers determine what misconceptions or other problems may be impeding their students' learning. Assessment items aligned to content standards allow 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. Used formatively, this feedback allows teachers to modify their instruction to focus on the specific ideas that the content standards target. The items also provide curriculum developers and researchers with high-quality tools for comparing the effectiveness of various instructional materials objectively. Most 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 those ideas and skills.

But perhaps the most important use will be as research tools to assess student understanding of science at various points in time. The recent interest in identifying learning progressions in science will require that researchers have available to them instruments that allow them to draw valid conclusions about student understanding along a learning trajectory. Although learning progressions are often described as broad categories of understandings (substance level vs. molecular level vs. subatomic level understanding of matter), it is also important to identify exactly what knowledge enables students to analyze, explain, and predict phenomena within those broad categories.

Without effective tools to measure student knowledge, research in science education will not produce results that we can have confidence in and on which we can base sound instructional decisions. The work described here is a start toward developing a set of instruments that can be used to precisely assess student understanding in science in various content areas and at various points along a learning trajectory. The work builds on the work done by Project 2061 to analyze content knowledge in the disciplines and to combine it with research on student learning to map the conceptual terrain for the ideas that are in *Benchmarks for Science Literacy*. (See *Atlas of Science Literacy*, AAAS, 2001, 2007 for maps that Project 2061 has developed.)

At the middle school level, our initial efforts to identify progressions of understanding will come from searching for patterns in how students in 6th vs. 7th vs. 8th grade answer various kinds of questions. We will try to learn when does their knowledge of controlling variables become secure? When in middle school are students comfortable with the language of atoms and molecules? When do they begin to think of each element being made of distinctly different atoms versus thinking that all matter is made of a single generic particle called an atom? Although our assessment work is currently focused at the middle school level, in the future it will move into both the 9-12 and the 3-5 grade bands. The work is also a first step toward identifying other factors, especially reading level and cognitive load that make test items more or less difficult for students. All of our test items are categorized by whether students are being asked to recognize the truth of a scientific generalization, analyze a situation, predict a phenomenon, or explain a phenomenon. We are also developing strategies to assess and understand the effect that various aspects of cognitive load place on students at different grade levels. When can they correctly explain the indirect effects that a change in an ecosystem has on other organisms in the ecosystem? When can they successfully analyze food web diagrams when the effect is several steps removed? How does the number of inferences that have to be made when answering a question affect success? How does the number of distinctly different pieces of information in a question affect success? Is it easier for students to explain phenomena or predict what will happen given certain conditions? These are the kinds of questions that our work is addressing and the kinds of questions that need to be answered if we are to have success in identifying both what students currently know and what they are capable of learning if provided with the highest quality instruction.

One thing we have learned in our work is that as hard as we try, we cannot create a perfect assessment item that always measures exactly what we are hoping to test. There is always error associated with measurement, and this is certainly true when measuring student understanding of ideas in science. We have also learned that although there is value in precisely aligning test items to specific target ideas, there is also virture in aligning test items to more than one idea. Aligning items to a single idea allows us to diagnose specific gaps in understanding, but aligning to multiple ideas allows us to determine how well students can bring ideas together in the solution of more complex problems. Both are important, and we intentionally develop different kinds of items to accomplish those two goals.

We have also learned that when we look closely at what students know and think, they reveal things that are almost impossible to think of in advance. We always ask how much can be taken for granted for middle school students—with respect to vocabulary, with respect to their understanding of prior ideas from earlier grade bands, and with respect to logical reasoning. But unless we get feedback from the students themselves, we never know for sure. We learned, for example, that not all middle school students know that the energy from the sun is captured in the leaves of a plant. A test item that we wrote had a tree with leaves and oranges hanging from it. Some students thought that the light from the sun was captured directly by the oranges.

Multiple choice tests are often criticized for assessing student knowledge of only 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. 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. They also have the advantage of being able to focus student attention on particular aspects of the ideas that are being targeted and on the misconceptions that students are likely to hold. This focus is particularly useful for diagnosing gaps in student knowledge.

References

1. American Association for the Advancement of Science. (1993). *Benchmarks for science literacy*. New York: Oxford University Press.
2. American Association for the Advancement of Science. (2001). *Atlas of science literacy.* Washington, DC: Author.
3. DeBoer, G.E. and Ache, P. (2005). Aligning Assessment to Content Standards: Applying the Project 2061 Analysis Procedure to Assessment Items in School Mathematics. Presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada. Retrieved April 11, 2007 from http://www.project2061.org/research/assessment/aera2005.htm.
4. National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.