Developing Assessment Items Aligned to Middle School Science Learning Goals

George E. DeBoer
AAAS Project 2061

Center for Curriculum Materials in Science
Knowledge Sharing Institute
Washington, DC
July 22-25, 2007

Assessment Plenary

Introduction

The first step in developing sound assessment items and instruments is to clearly define the construct that is to be measured (National Research Council, 2001). The construct could be knowledge of a particular scientific fact, a principle, or an interconnected set of ideas; the ability to use scientific knowledge to make predications and explanations of phenomena; or the ability to engage in a scientific practice such as judging whether appropriate experimental controls are in place. In the current work being done at AAAS Project 2061, our primary focus is on measuring students' ability to use knowledge to predict and explain phenomena and to identify gaps in knowledge that limit students' ability to make those predictions and explanations.

Recognizing the importance of high quality assessment items that are aligned to content standards and the poor state of many of the items currently in use (American Federation of Teachers, 2006), Project 2061 is engaged in a multi-year, NSF-funded project to develop a bank of assessment items aligned to middle school content standards in science. We are developing items in 16 topic areas from the life, physical, and earth sciences; the nature of science; and mathematics. This work involves the development of distractor-driven, multiple-choice test items (Sadler, 1998) in which common misconceptions are used as distractors.

The procedure that Project 2061 has designed for the development of assessment items involves three stages: (1) clarifying the targeted content standard, (2) designing assessment tasks that are precisely aligned to the specific ideas in that content standard, and (3) using information obtained from students during interviewing and pilot-testing to revise items. The procedure has

been used and adapted by various individuals and groups, including curriculum researchers at CCMS. (Add Krajcik, McNeill, and other references.)

Clarifying a Content Standard (Defining the Construct)

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. The learning goals in these documents and the organization of these learning goals are intended to provide guidance for the development of instructional activities and materials as well as for the development of assessment items and instruments. The expectation is that students will develop mental models of objects, events, and processes in the world that will enable them to explain these events and make predictions about them.

Key ideas. Although content standards provide useful direction to assessment developers regarding what students should know in science, these statements often do not provide enough detail about exactly what students should be held accountable for. To provide additional guidance, 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 from the topic of plate tectonics.

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

Clarification statements. Obviously there are concepts in this statement about earth's plates that need to be elaborated. Exactly what knowledge should students have of what a plate is? The clarification statement that we wrote to describe the mental model we expect students to have 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.

This clarification statement was written in response to three questions that are central to the design of assessments that target key ideas:

1. Does this statement adequately describe the knowledge that is needed for middle school students to form a mental image of earth's plates that allows them to predict and explain phenomena involving plates?
2. Does this statement adequately describe the knowledge that is needed for students to understand *later ideas* and the accompanying phenomena 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 (and, in effect, how large they are with respect to the size of the earth), that the plates are not all the same size and shape, and that the plates fit tightly together. In addition to guiding assessment development,

these elaborations of the term "plate" can also be used to guide instruction that will lead to a mental model of a plate that students can use when learning subsequent ideas about plate motion and the consequences of plate motion, which come later in the instructional 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 decided not to expect students to know the term "lithosphere" because we did not think that knowing that term would contribute significantly to explaining phenomena related to plate motion. We recognize that individual teachers may choose to teach students the relationship between lithosphere, upper mantle, and plates, but our assessment items will not include the term for reasons explained above.

Research on student learning. We also examine the research on student learning (see, for example, Driver et al., 1996) as we are clarifying our expectations for students. This research 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 distractors in the items so that we can test for these ideas with the targeted science ideas (Sadler, 1998).

Connections among ideas. After the ideas that are to be assessed have been identified and clarified, our next step is to think about how those ideas relate to other ideas within a topic and across grade bands. The objective here is to be as clear as possible about the ideas we are explicitly testing and the prior knowledge we can assume students will already have. For example, if we expect students to know that digestion of food involves a process in which the atoms of molecules from food are rearranged to form simpler molecules, can we assume that students already know that molecules are made of atoms? If they do not, test questions on chemical digestion are also testing whether students know the relationship between atoms and molecules. Making judgments about which ideas precede the targeted idea and the knowledge students most likely already have is important in item development if the results are to be used for diagnostic purposes.

In making judgments about which ideas precede a targeted idea, we begin by examining the conceptual strand maps published in the *Atlas of Science Literacy* (AAAS, 2001, 2007). The strand maps are derived largely from what the authors of *Benchmarks for Science Literacy*

(AAAS, 1993) and the *National Science Education Standards* (NRC, 1996) thought were appropriate grade placement of the various ideas. The map for the topic of natural selection in Figure 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 we freely use the ideas and language from those earlier ideas in item development for the grade band that follows. But we also recognize that these earlier ideas also are a good place to look when students do not know the targeted idea. Not knowing an earlier idea is often the reason why students have difficulty with later ideas. The relationships identified in the *Atlas* maps can help us focus on ideas that may be needed for understanding later ideas.

In our assessment work, we also map the finer grained key ideas onto what we call assessment maps. These are analogous to the *Atlas* maps but are written at the idea level rather than at the benchmark level. (See DeBoer, 2005 for further discussion of assessment maps.) We then use the results of student testing of these ideas to validate the hypothesized links on the maps. An example of one way this validation is being done appears later in this paper.

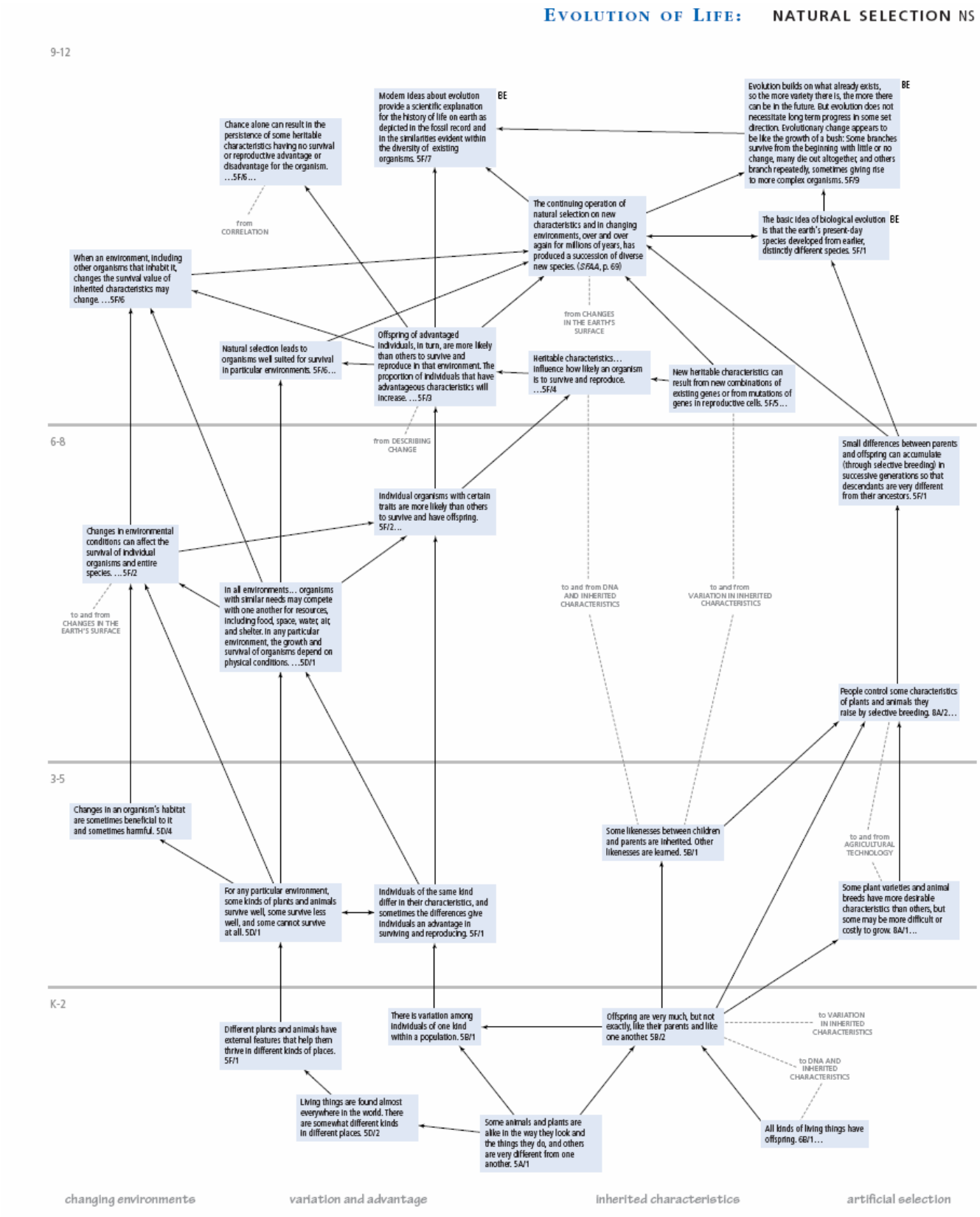


Figure 1: Natural Selection Map from Atlas of Science Literacy (AAAS, 2001)

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 specified 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 in the learning goal is *enough by itself* to successfully complete the task (Stern & Ahlgren, 2002). 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.

When items are well aligned to the ideas in a targeted content standard, student responses can provide accurate insights into their understanding of that content. But as important as content alignment is, content alignment alone is not enough to judge whether an item should be used. There are many other factors that can affect the usefulness of an assessment item. These factors are discussed below.

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 that student knows about an idea. When an item is well designed, students should choose the correct answer only when they know the targeted 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 (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. To do this, we have established a detailed item review protocol that includes specific ways to improve construct validity. (The Project 2061 assessment item review protocol can be accessed at http://www.project2061.org/research/assessment.htm.)

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 analytical approach works much more effectively when used in combination with one-on-one interviews with students and pilot tests of items in which students' answer choices are compared to the reasons they give for their answer choices (DeBoer & Ache, 2005).

By comparing the answers that students select with their oral or written reasons for their answer choices, we determine if an assessment item is measuring what we want it to measure or if students are giving false negative or false positive responses to that item. In the pilot testing, we ask the questions shown in Figure 2 to find out what students are thinking. Note that in addition to saying that an answer choice is correct or incorrect, students also have the option of saying they are "unsure" about an answer choice and why they are unsure. When we see a large number of unsure responses, it may indicate that there are structural problems with the item (perhaps language that is unclear) that need to be addressed. Pilot tests are carried out in urban, rural, and suburban middle schools serving a wide range of students. Approximately 100 to 150 students respond to each item during pilot testing. (See DeBoer, Abell, & Gogos, 2007; Abell & DeBoer, 2007; and Gogos & DeBoer, 2007 for addition discussion of our pilot testing methods.)

Figure 2: Questions Students are asked about each Test Item

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.

(For questions 3 – 6, students are asked to explain why an answer choice is correct or not correct, or why they are "not sure.")

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? Yes No Not Sure
    Where? (TV, museum visit, etc)?

Results of Pilot Testing

The following examples illustrate the information we are able to derive from pilot test results. The examples also show how we use what we learn to improve the items' alignment to the key ideas and their validity as measures of student learning.

Example 1: Atoms, molecules, and states of matter. The item shown in Figure 3 tests whether students know that molecules get farther apart when they are heated and if they know that this molecular behavior explains why most substances expand when heated. The item includes as answer choices common misconceptions related to thermal expansion and the behavior of molecules, especially the idea that there are "heat molecules."

Figure 3: Pilot Test Results for a Test Question on Thermal Expansion

---

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.

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.

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

Pilot testing showed that approximately 26% of the students answered this question correctly. The most common incorrect response (44%) was that heat molecules push the alcohol molecules upward. Pilot testing also revealed that a number 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." 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 some students 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 also 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 answer choice 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 children's ability to engage in proportional reasoning, we need to determine through

additional testing and interviewing if the thermometer context is appropriate for general testing of middle school students' understanding of thermal expansion. The thermometer example also raises questions about middle school students' ability to understand more generally that some measuring devices proportionately amplify small changes so they can be more easily observed.

Example 2: Atoms, molecules and states of matter. Figure 4 shows an item that tests 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, the width of a hair, etc.

Figure 4: Pilot Test Results for a Test Question on the Size of Atoms

---

Key Idea: All atoms are extremely small.

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

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

The results of pilot testing showed that approximately 23% of the students answered this question correctly. In their written comments, 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. In addition, a large percentage of students (45%) indicated they were not sure about the answer choices. 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 was difficult for some students, try to create a simpler context for this item.

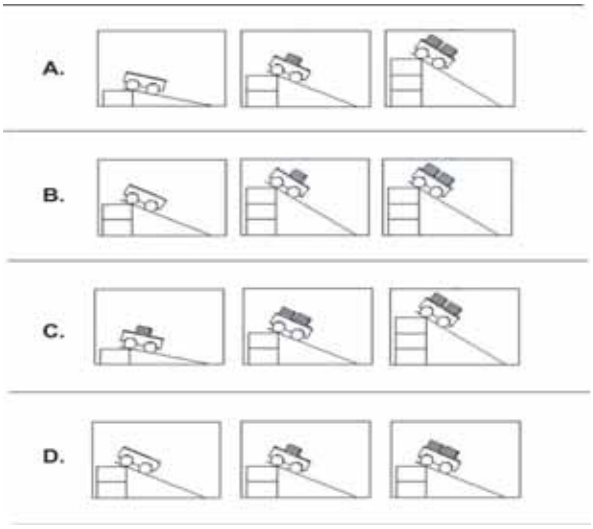
Example 3: Control of variables. The item shown in Figure 5 was developed to determine if students understand that the way to determine if one variable is related to another is to hold all other relevant variables constant and to test a number of common misconceptions that students have regarding the control of variables, including the misconception that all variables may vary in a controlled experiment.

Figure 5: Pilot Test Results for a Test Question on the 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.

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?



	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 showed that approximately 54% of the students answered correctly and that 35% of the students chose either answer choice A, B, or C, all of which are confounded designs in which both the weight and height are allowed to vary. Of the three incorrect answers, the most popular was A, in which there is a progressive increase in both weight and the height. Answer choices B and C were less successful distractors. 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 one so that there are no empty carts.
2. Replace the question in the stem with: "Which three sets of experiments should the student do?" Label the three images for each answer choice "Experiment 1," "Experiment 2," and "Experiment 3."

Using Field Testing to Determine Psychometric Properties of the Items and to Model Student Understanding at the Topic and Key Idea Levels

After pilot testing is completed, scientists and science education experts review the items using a set of criteria to ensure content alignment and construct validity. (The complete Project 2061 assessment item review protocol can be accessed at http://www.project2061.org/research/assessment.htm.) Reviewers also use the pilot test data, which include the number of students who selected each answer choice, student comments about why they think each answer choice is correct and incorrect, and lists of words students said they did not understand. After revisions are made, the items are field tested on a national sample to determine the psychometric properties of the items and clusters of items. The field test data are used to construct difficulty scores, item discrimination indices, and differential item functioning scores. Field test data are also used to model student understanding of the targeted science ideas

at both the key idea level and topic level using item response theory (IRT) modeling (Wilson, 2005).

[Preliminary analysis is currently being made of the results of field testing of two topics—1) Control of Variables and 2) Atoms, Molecules, and States of Matter. Some of these data are presented in the remainder of this paper.]

For each topic, 3000-4000 students were field tested. Students from a wide range of urban, suburban, and rural school districts across the country responded to the items. For the two topics combined, approximately 42% of the students were students of color, and 13% of the students indicated that English was not their primary language. Because we were testing more items than students could reasonably complete in a typical class period, different test forms were created that contained subsets of the available questions. Also, because we were interested in describing both the students' understanding of the topic as a whole and their understanding of the ideas within each topic, some of the test forms included a random selection of all of the items and some of the test forms included items from selected clusters of the ideas being tested. For both the Control of Variables and Atoms, Molecules, and States of Matter topics, we created four different test forms. Additionally, for each form of the test, half of the students took the items in reverse order so that the last items on the test would not be disproportionately omitted when students ran out of time.

Some Field Test Results for the Control of Variables Items:

For the control of variables field test, we presented students with four types of questions.

1. Given an idea to be tested (hypothesis) and an experimental setup, know why certain variables are (or should be) kept constant.
2. Select an experimental setup to test the effect of a variable on the experimental outcome.
3. Identify the variable(s) being tested in a given controlled experimental setup.
4. Given an experiment with two variables changing at the same time, determine that no (unconfounded) conclusion can be drawn regarding the effect of each variable.

Type 1 and Type 4 questions were on test Form 1, and Type 2 and Type 3 questions were on Form 2. Forms 3 and 4 included a randomly selected subset of the total set of items. Each test form included 14 or 15 questions. The four forms of the Control of Variables tests had Cronbach's alpha reliability coefficients of .821, .718, .768, and .764 respectively. Four items (one of each type) appeared on all four forms and were used as linking items for various

statistical tests. Students were tested during the spring of 2007. The field testing results for these four items appears in the table below. (The four items referenced in the table appear in Appendix A.)

Table 1: Percent of Students Answering Correctly on Four Linking Items by Grade (Four Forms Combined)

Item	Type	6th grade	7th grade	8th grade	Total (N=2812)
15-5	1	51%	50%	51%	51%
1-3	2	44%	41%	41%	42%
17-2	3	36%	39%	37%	37%
22-2	4	20%	20%	22%	21%
Four Items		38%	38%	38%	38%

Two observations stand out. First is that Type 4 questions are considerably more difficult for middle school students than are the other three types. We found this to be true during pilot testing as well (Gogos & DeBoer, 2007). Type 4 items present students with the results of a confounded study and ask them what they can learn from it. Most students think you can learn about the effect of both variables. The second observation is that there is no observable growth in understanding of ideas about control of variables between 6th and 8th grades on any of the item types. Students do as well at the end of 6th grade on all item types as they do at the end of 8th grade.

Additional analyses that are being conducted include comparison of means on each item and item type by grade, by gender, by race and ethnicity, and by English as the students' primary language. We will also use differential item functioning analysis to determine if the items perform similarly for students who do and do not have English as their primary language and for students from varying racial and ethnic groups. We will also examine item characteristic curves to determine the contribution that each item in a set makes to measuring students' understanding of controlling variables (Reference).

Field Test Results for the Carts and Ramps Item

The carts and ramps item that was used in pilot testing and described earlier in this paper was revised following the suggestions of various reviewers and the results of pilot testing of students. Balls were placed in the carts so that the blocks the ramp rested on would not be

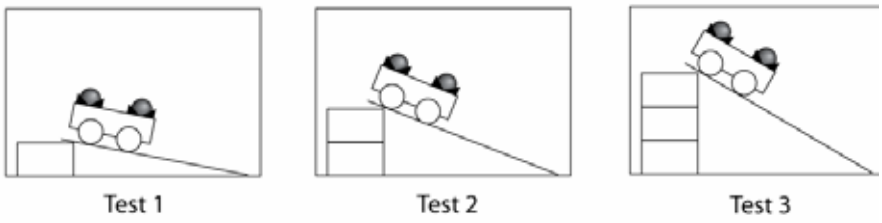
confused with the blocks in the carts. Balls were placed in all the carts so that students would not reject an answer choice because one of the carts was empty. In addition, the misconception that the thing to be tested (weight) should be held constant was added as a distractor.

Figure 6. Revised Carts and Ramps Items for 2007 Field Testing

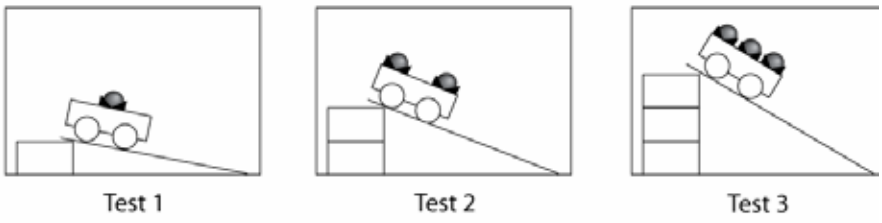
A student wants to know if the weight of a cart affects its speed at the bottom of a ramp. He can change the weight of the cart by adding different numbers of balls, and he can change the height of the ramp by using different numbers of blocks.

Which set of tests should he compare (set A, B, C, or D)?

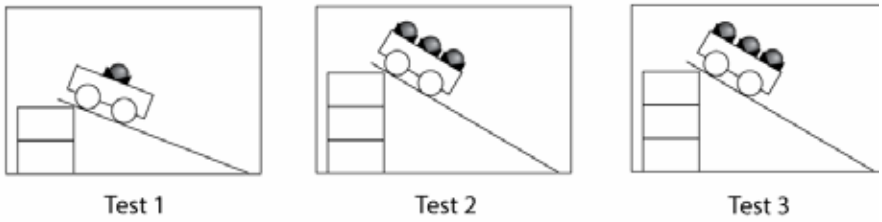
○ A.



○ B.



○ C.



○ D.

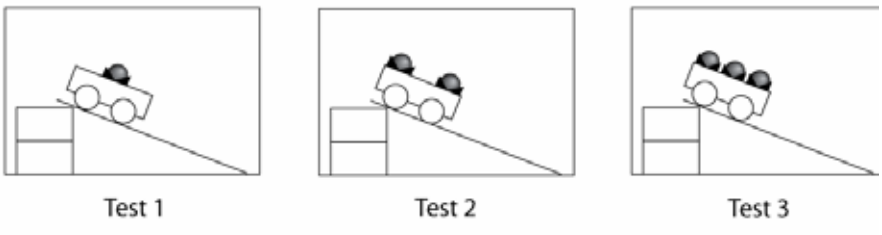


Table 3: Percent of Students Selecting each Answer Choice for Carts and Ramps Item (2007 Field Test Data)

Answer Choice	A	B	C	**D**
	Hold constant the thing to be tested (weight)	Let more than one thing vary	Let more than one thing vary	**Correct Answer**
Form 1 (N=755)	18%	31%	6%	**44%**
Form 2 (N=601)	22%	30%	4%	**43%**
Form 3 (N=779)	21%	31%	7%	**40%**
Form 4 (N=677)	17%	31%	8%	**42%**
Combined (N= 2812)	20%	31%	6%	**42%**

Similar to the results for the carts and ramps item that was pilot tested, 37% of the students selected answer choices B or C, in which both the height and the weight are allowed to vary (confounded studies). Most chose answer choice B, and a much smaller number chose answer choice C, in which Test 3 is simply a duplication of Test 2. Answer choice A, in which the height varies and the thing to be tested (weight) is held constant, was chosen by 20% of the students.

A smaller number of students were successful on the field test item than on the pilot test item (42% vs. 54%), perhaps because the field test item included the added misconception that the thing that you want to test should be held constant. Based on these data of 2812 middle school students, the stronger of the two misconceptions is the idea that more than one thing can vary in an experiment (31%). The idea that the thing to be tested should be held constant was chosen by 20% of the students.

Some Field Test Results for the Atoms, Molecules, and States of Matter Items:
During field testing, students answered questions related to seven key ideas. The seven key ideas are:

Idea A: All matter is made of atoms.

Idea B: All atoms are extremely small.

Idea E: All atoms and molecules are in constant motion.

Idea F: There are differences in the spacing, motion, and interaction of atoms and molecules that make up solids, liquids, and gases.

Idea G: For any single state of matter, changes in temperature typically change the average distance between atoms or molecules. Most substances or mixtures of substances expand when heated and contract when cooled.

Idea H: Changes of state can be explained in terms of changes in the arrangement, motion, and interaction of atoms and molecules.

Idea I: For any single state of matter, the average speed of the atoms or molecules increases as the temperature of a substance increases and decreases as the temperature of a substance decreases.

As with the Control of Variables field test, four forms of tests were created. The results that are presented here are based on the number of students who answered an item for a particular idea. The results shown in Table 4 indicate that students had the most difficulty with ideas related to thermal expansion and changes in state, and to the idea that atoms are always in motion. They were most successful with questions testing the idea that all matter is made of atoms and that atoms are extremely small. We will conduct a distractor analysis to determine which misconceptions are the most common.

Table 4: Field Test Results for Atoms, Molecules, and States of Matter Topic (By Idea)

	Idea A	Idea B	Idea E	Idea F	Idea G	Idea H	Idea I
# correct	5439	4684	3699	9256	7787	4729	6738
# incorrect	4329	3343	6900	9424	13540	9513	6577
% correct	56%	58%	35%	50%	37%	33%	51%

Unlike for the Control of Variables items, there is a change in student knowledge of ideas about atoms, molecules, and states of matter from grade-to-grade. As can be seen in Table 5, the greatest change is from 6th grade to 7th grade (41% correct to 48% correct). At a later date, we will examine this trend for each idea (controlling for certain demographic variables) to get a more complete picture of these grade-to-grade changes.

Table 5: Field Test Results for the Atoms, Molecules, and States of Matter Topic (Overall Percent Correct by Grade)

Grade	Grade 6	Grade 7	Grade 8
% correct	41%	48%	46%

Field Test Results for the Thermometer Item

The thermometer item that was used in pilot testing and described earlier in this paper was revised following the suggestions of various reviewers and the results of pilot testing of students. The word "alcohol" was changed to "colored liquid" and the word "heat" was added to answer choice C so that B would not be the only answer choice with the word heat in it. Reviewers were concerned that if B was the only answer choice with the word heat in it, students would choose it because they have learned that "heat *rises*" and the question asks why the liquid *rises* when heated.

Figure 7: Revised Thermometer Item for 2007 Field Testing

A glass thermometer has a colored liquid inside it. The level of colored liquid rises when the thermometer is placed in hot water. Why does the level of liquid rise?

A. Water molecules are pushed into the thermometer.
B. Heat molecules push the molecules of the liquid upward.
C. Heat causes the molecules of the liquid to get farther apart.
D. The molecules of the liquid break down into atoms and take up more space.

As seen in Table 6, 44% of students thought that heat molecules pushed the molecules of the liquid upward in the thermometer. This is the same percentage that selected this answer choice on the pilot test. On the pilot test, approximately 26% of the students chose the correct answer compared to 39% on the field test. As with the pilot test, very small numbers of students selected the answer choices that tested the idea that, when heated, water molecules are pushed

into the thermometer (4.6%) or the idea that the molecules of the liquid break down into atoms that take up more space (6.5%). The difference between the number correct on the pilot and field test can be partly explained by the fact that on the pilot test students could indicate that they were not sure about their answer. By changing "alcohol" to "colored liquid," some of the students who would have been "unsure" now answered correctly.

Table 6: Field Test Results for Thermometer Item (N=2687)
(All forms combined)

Answer choice	A	B	**C***	D
# of students	216	1169	**1061**	238
% of students	8%	44%	**39%**	9%

Field Test Results for the Size of Atoms Item

The size of atoms item that was used in pilot testing and described earlier in this paper was revised following the suggestions of various reviewers and the results of pilot testing of students. The phrase "approximately how many" was changed to "about how many" and "carbon atoms" was changed to "atoms."

Figure 8: Revised Size-of-Atoms Item for 2007 Field Testing

About how many atoms would it take to make a line across this dot: • ?

A. 6
B. 600
C. 6000
D. 6,000,000

As seen in Table 7, 38% of students knew that it would take millions of atoms to cross the dot on the paper. Although the percentage correct was lower on the pilot test item (22.8%), that lower number was due mainly to the fact that students could indicate that they were not sure on the pilot test but not on the field test. On the pilot test, a large number of students (45%) said

they were not sure. On both the pilot and field tests there was a progression in the answer choices that matched the size of atoms. Only 8.5% of students on the pilot test and 11% on the field test thought that it would take only 6 molecules to cross the dot. Somewhat more thought it would take 600; somewhat more than that thought it would take 6000; and in both pilot and field tests, the largest percentage thought it would take 6 million atoms.

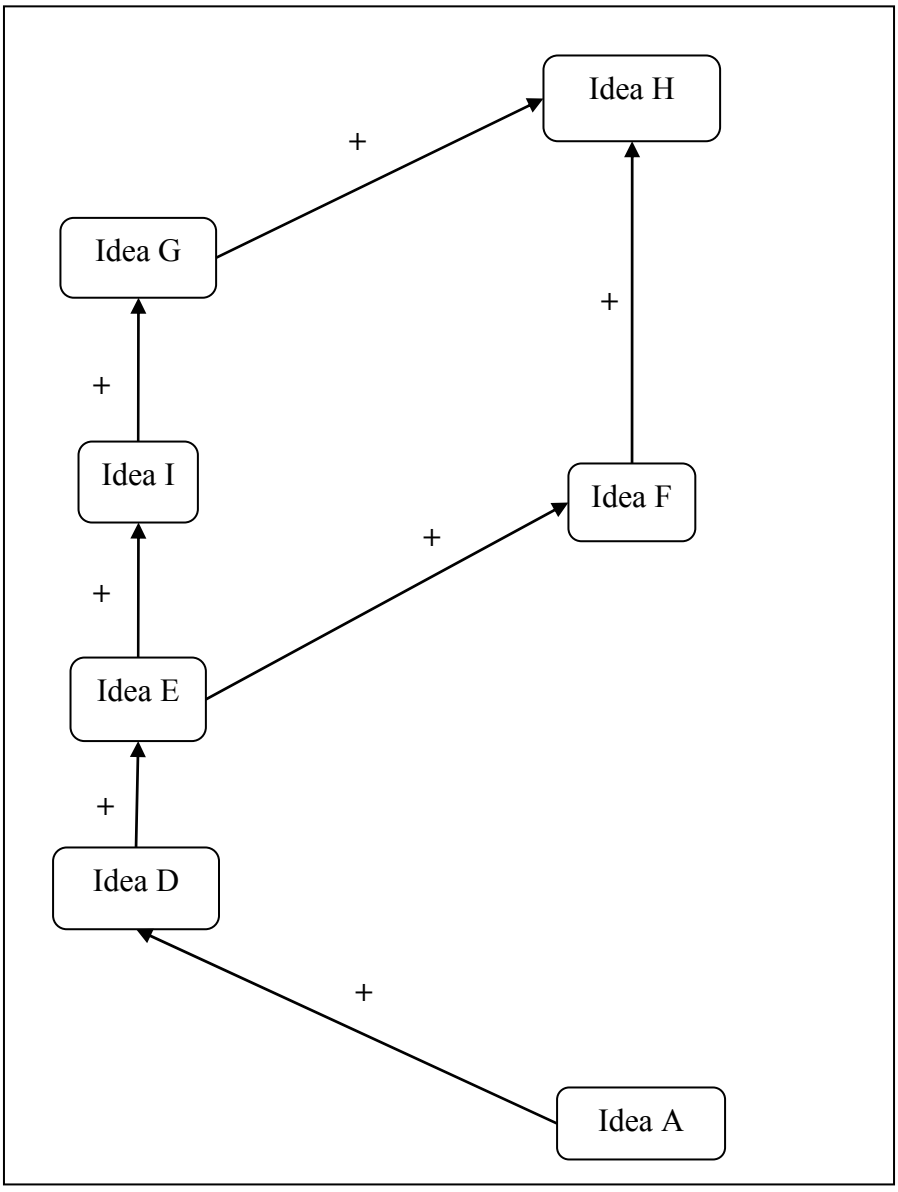
Table 7: Field Test Results for Size of Atoms Item (N=2687)
(All forms combined)

Answer choice	A	B	C	**D***
# of students	199	380	527	**681**
% of students	11%	21%	29%	**38%**

Testing Hypothesized Connections among Ideas Represented on Assessment Maps Using Path Analysis

On our assessment maps, we hypothesize a set of relationships among the targeted ideas. In our hypothesized model, ideas A, D, E, I, F, and G are precursors (either directly or indirectly) to idea H; ideas A, D, E, and I are precursors to idea G; ideas A, D, and E are precursors to idea F; ideas A, D, and E are precursors to idea I; A and D to idea E; and A to D. (See page 17-18 of this paper for a description of each idea.) We did not have items to test student understanding of idea D, so the model we tested had A feeding directly into E instead of through idea D. The model was tested twice, using two forms of a test covering the same ideas. For the first form (Form 3ab), a set of items was randomly selected from available items. If random selection did not yield at least two items on an idea, additional items were randomly chosen from that idea set to yield additional items on that form. The remaining items were used on the second form (Form 4ab).

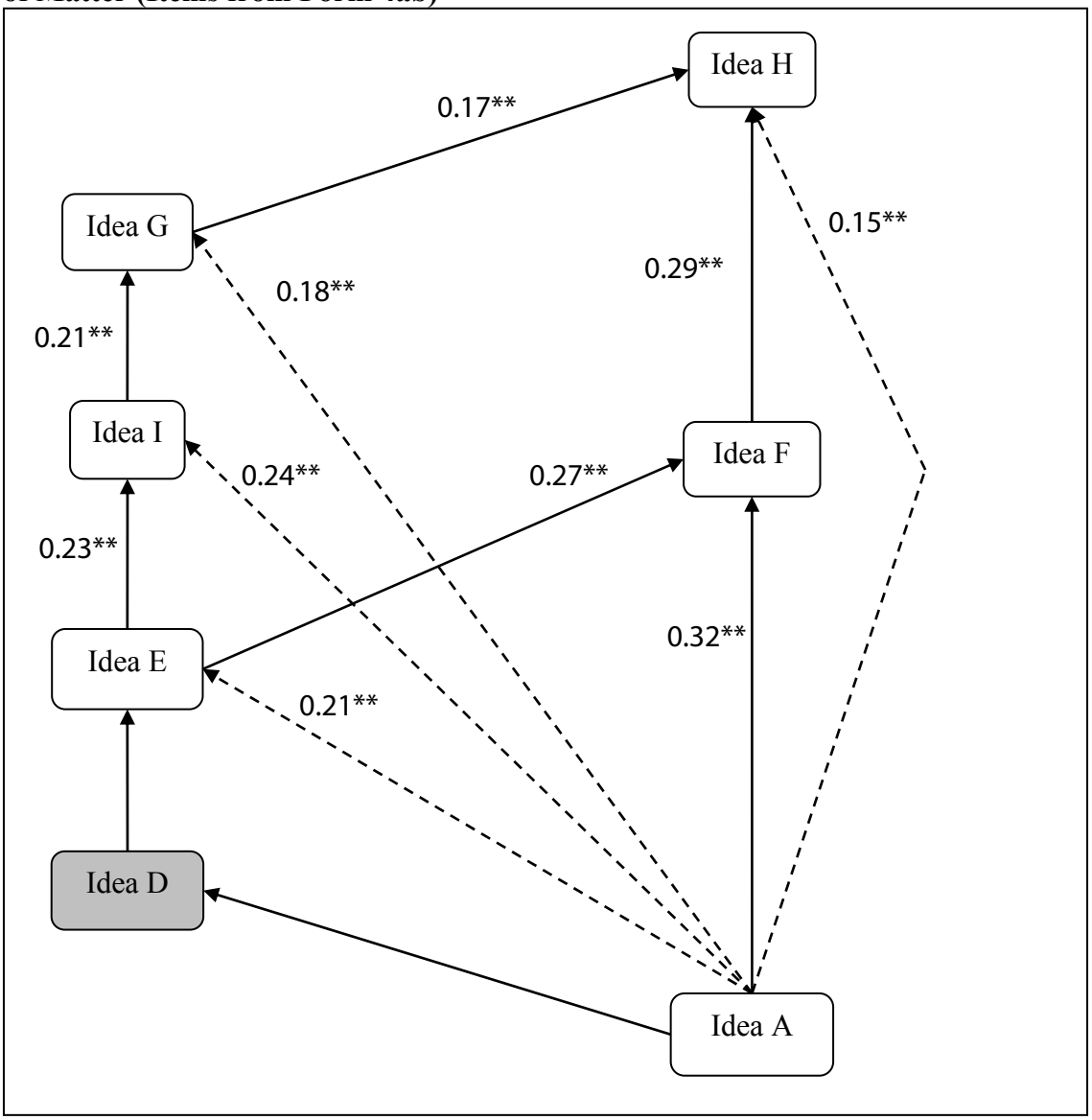
Figure 9: Hypothesized Relationships among Ideas for Atoms, Molecules, and States of Matter



Standardized regression coefficients (beta weights) were calculated for a series of structural equations with Idea H, Idea G, Idea I, Idea F, and Idea E successively being used as the dependent variable in the equation. Significant beta weights (at the .01 level of significance) were observed for all hypothesized links in the model. Beta weights ranged from .15 to .32 using Form 4ab of the test, and ranged from .14 to .37 using Form 3ab. Besides the hypothesized effects, additional effects were observed that had not been

hypothesized in the model. These were direct effects of idea A (all matter is made of atoms) on later ideas in the sequence. Direct effects of Idea A were observed on idea I, F, G, and H using items from Form 4ab. Direct effects of A were observed on ideas I, F, and G using items from Form 3ab. R square values for the full model were .12 and .24 for the two forms of the tests.

Figure 10: Observed Relationships among Ideas for Atoms, Molecules, and States of Matter (Items from Form 4ab)



Summary

In our work, we place a great deal of emphasis on the qualitative alignment of assessment items to learning goals. Using the criteria of necessity and sufficiency ensures a high degree of accuracy in this alignment. And we are rigorous in applying these alignment criteria, having the items reviewed both internally and externally by teams of reviewers who are carefully trained in the use of the alignment criteria. During review, we pay close attention to possible construct-irrelevant features of the test items that could produce false negative or false positive responses on the part of students. Test features such as low comprehensibility of the items and easy application of test-wiseness strategies by students can increase the number of students whose answers are not a valid indicator of their actual knowledge. In pilot testing and interviewing, students are asked to explain why answer choices are correct or incorrect, and what they say is then matched with the answer choice they select to further validate student responses or to revise the item when structural problems are found.

In addition to our emphasis on making qualitative judgments about validity, we also pay attention to the psychometric properties of test items. Items are field tested on a national sample to determine the psychometric properties of items and clusters of items. The field test data are used to construct difficulty scores, item discrimination indices, and differential item functioning scores. Field test data are also used to model student understanding of the targeted science ideas at both the key idea level and topic level using item response theory (IRT) modeling.

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. 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. In addition, multiple choice items have the advantage of being able to carefully define the space in which students think about the problem or task and, in that way, reduce the irrelevant responses that often come with open-ended test questions. In particular, multiple choice items can deliberately focus on particular misconceptions that students may have and in that way assess the relative strength of those misconceptions on large numbers of students.

References

Abell, C. F. & DeBoer, G. E., Probing Middle School Students' Knowledge of Thermal Expansion and Contraction through Content-Aligned Assessment. Paper presented at the Annual Conference of the National Association for Research in Science Teaching. New Orleans, LA, April, 15-18, 2007.

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

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

American Association for the Advancement of Science. (2001, 2007). *Atlas of science literacy.* Washington, DC: Author.

American Association for the Advancement of Science. (2006). Project 2061 procedure for analyzing science assessment items. [Online utility]. Retrieved July 18, 2007, from http://www.project2061.org/research/assessment.htm.

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

DeBoer, G. E., Abell, C. F., & Gogos, A., 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, April, 15-18, 2007.

DeBoer, G.E. & Ache, P. (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. Retrieved September 29, 2006, from http://www.project2061.org/research/assessment/aera2005.htm.

DeBoer, G. E. (2005). Standard-izing test items. Science Scope, January, 2005.

Driver, R., Leach, J., Millar, R., & Scott, P. (1996). *Young people's images of science.* Bristol, PA: Open University Press.

Gogos, A. & DeBoer, G. E. (2007). Assessing students' understanding of controlling variables. Paper presented at the Annual Conference of the National Association for Research in Science Teaching. New Orleans, LA, April, 15-18, 2007.

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

National Research Council (2001). *Knowing what students know.* Washington, DC: National Academic Press.

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.

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.

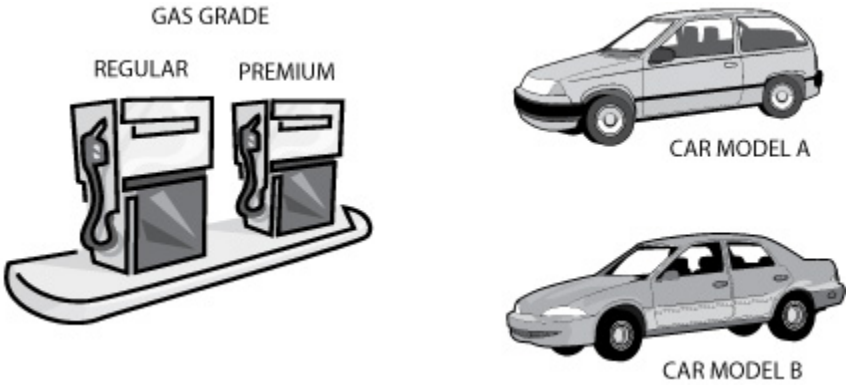
Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Associates: Mahwah, NJ.

Appendix A

Item CR15-5 (Type 1)

A consumer group wants to find out which of two new car models gets the best gas mileage. A car's gas mileage is the number of miles a car can go for each gallon of gas it uses.

They decide to fill the gas tanks of each car with the same amount of gas and compare how far each car goes. They use "regular" grade gas in both cars. Neither car gets the "premium" grade gas.



Why is it important that the two cars get the same grade of gas?

A. By using the same grade of gas, the consumer group can learn both which car model gets the best mileage and which grade of gas gives the best mileage.
B. By using the same grade of gas, the consumer group can learn which grade of gas gives the best mileage.
C. If the cars do not get the same grade of gas, the consumer group cannot find out which car model has the best mileage.
D. It is NOT important for both cars to have the same grade of gas because they are not testing which grade of gas gives the best mileage.

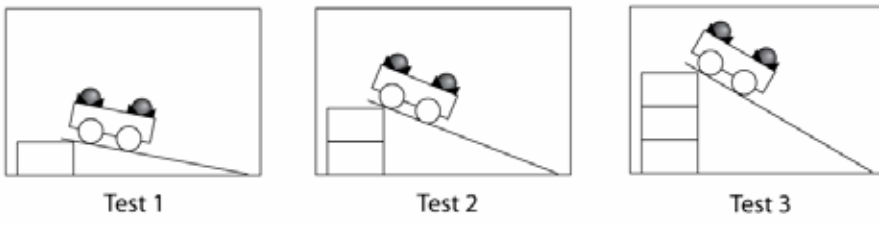
Item CR1-3 (Type 2)

1. A student wants to know if the weight of a cart affects its speed at the bottom of a ramp.

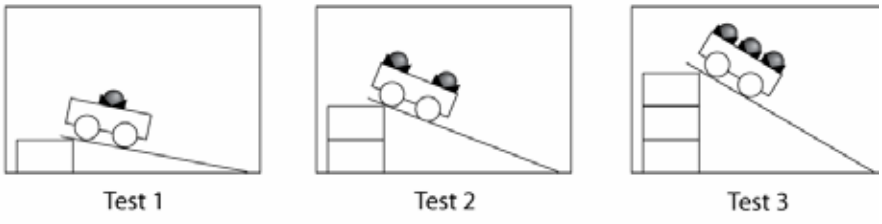
   He can change the weight of the cart by adding different numbers of balls, and he can change the height of the ramp by using different numbers of blocks.

   Which set of tests should he compare (set A, B, C, or D)?

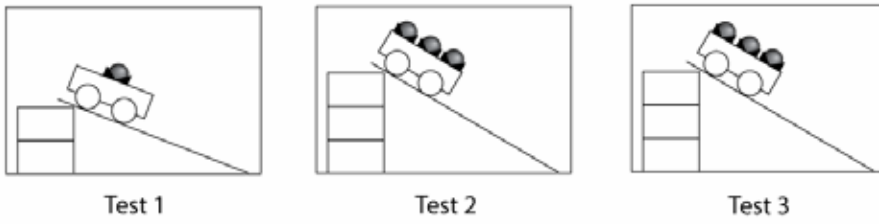
○ A.



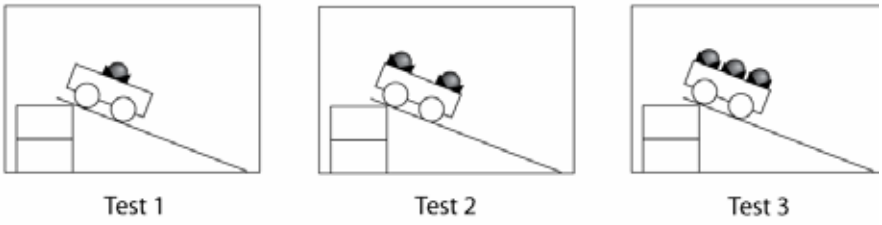
○ B.



○ C.

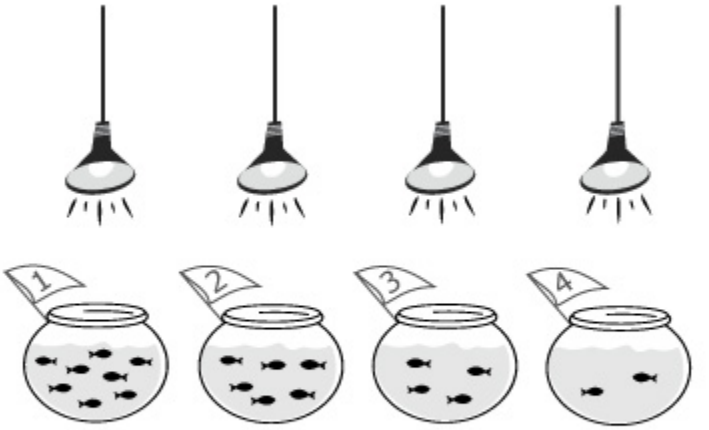


○ D.



Item CR17-2 (Type 3)

A student is interested in the behavior of fish. He has four fish bowls and 20 goldfish. He puts eight fish in the first bowl, six fish in the second bowl, four fish in the third bowl and two fish in the fourth bowl. He places each fish bowl under light, he keeps the temperature at 75°F for all four bowls, and he observes the behavior of the fish.



Number of fish	8 fish	6 fish	4 fish	2 fish
Temperature	75°F	75°F	75°F	75°F

What can the students find out from doing just this experiment?

A. If the number of fish in the fish bowl affects the behavior of the fish

B. If the temperature of the fish bowl affects the behavior of the fish

C. If the temperature of the fish bowl and the amount of light affect the behavior of the fish

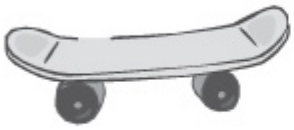
D. If the number of fish, the temperature, and the amount of light affect the behavior of the fish

Item CR22-2 (Type 4)

A student wants to buy a new skateboard. He wants to find out if the size of the wheels affects how far he can coast on the skateboard. He also wants to find out if the type of material the board is made of affects how far he can coast on the skateboard.

He decides to compare two skateboard models that are the same size but are made of different materials and have different size wheels:

Model **1**



Plastic skateboard
Large wheels

Model **2**

Wooden skateboard
Small wheels

He pushes off as hard as he can and stands on the skateboard until the skateboard comes to a stop. He tries each skateboard five times to see how far he can go. He uses the same pavement and the same starting point for all the trials.

He finds out that he can coast farther with Model 1.

What can he conclude from this test?

A. He can conclude that the size of the wheels affects how far he can coast on the skateboard.

B. He can conclude that the material of the board affects how far he can coast on the skateboard.

C. He can conclude that the size of the wheels affects how far he can coast on the skateboard and that the material of the board affects how far he can coast on the skateboard.

D. It is not possible to conclude from this test if the size of the wheels affects how far he can coast on the skateboard or if the material of the board affects how far he can coast on the skateboard.