

Assessing Integrated Understanding of Science

*George E. DeBoer, Hee-Sun Lee, and
Freda Husic*

QUANTITY ASSESSMENT is an essential part of quality instruction. Assessment can be used to monitor student progress toward stated learning goals and to measure the effectiveness of instruction. In an era of public accountability, assessment instruments have enormous potential to pull curricula and instructional practices toward what is included in those assessments. When the goal is for students to develop an integrated understanding of science, assessment can support this goal through careful structuring of the kinds of assessment tasks students perform and through provision of feedback on how well students are moving toward accomplishing the goal. In this chapter, we focus on (a) ways of designing test items to measure a variety of learning outcomes, including development of particular knowledge and skills, mental models of science processes, and hypothesized mental constructs related to an integrated understanding of science; (b) the use of qualitative and psychometric analyses to obtain information on the quality of assessment instruments; and (c) the use of assessment results to revise the curriculum materials being developed. We use examples from the Center for Curriculum Materials in Science (CCMS) and the Technology Enhanced Learning in Science (TELS) Center to illustrate how assessment can lead to the development of integrated understanding in science.

DECIDING WHAT TO ASSESS: DEFINING THE CONSTRUCT

The first step in developing assessment items and instruments is to clearly define what is to be measured. This might be knowledge of a particular scientific fact,

principle, or interconnected set of ideas; the ability to use scientific knowledge to make predictions and explanations of real-world phenomena; or the ability to engage in scientific practices such as designing experiments. At both CCMS and TELS, the primary focus is on measuring students' ability to use knowledge to solve problems and explain scientific phenomena, and on identifying gaps in knowledge that limit students' understanding of events in the natural world.

In the section that follows, we provide three examples of constructs used to guide the item-development process. In the first example, we describe a 2-year item-development cycle being used by researchers at Project 2061 of the American Association for the Advancement of Science (AAAS) to design test items aligned with ideas in the national content standards, specifically ideas in *Benchmarks for Science Literacy* (AAAS, 1993) and the *National Science Education Standards* (NRC, 1996). In the second example, we illustrate how CCMS researchers have adapted the Project 2061 approach in their development of assessment items that measure students' use of science practices linked to science content, specifically students' ability to generate scientific explanations of phenomena related to the properties of substances. In the third example, we describe how TELS researchers design assessments to measure students' ability to connect multiple science ideas to solve problems in the real world, a major component of the TELS knowledge integration construct.

Aligning Assessment Items with Science Ideas in the National Content Standards

Recognizing the importance of high-quality assessment items that are aligned with the science ideas in state and national content standards and the poor quality of many of the items currently being used (American Federation of Teachers, 2006), Project 2061 has been engaged in a multi-year project to develop assessment items precisely aligned with middle school content standards in science (DeBoer, 2005).

Project 2061 has developed a five-stage process for creation of assessments: (a) clarifying each content standard targeted for assessment; (b) applying a set of alignment criteria in devising assessment items that are aligned with the specific ideas in the content standards; (c) obtaining feedback on the items from students during interviews and pilot testing, and revising the items based on that feedback; (d) having the items formally reviewed by science specialists and experts in science education, and making revisions based on those reviews; and (e) field testing the items on a national sample of students.

Clarify the Content Standards. Both *Benchmarks for Science Literacy* (AAAS, 1989) and the *National Science Education Standards* (NRC, 1996) are organized around ideas and skills that all students should learn by the end of certain grade

bands in order to effectively engage in a world in which science and technology play such an important role. In addition to identifying what should be learned, these standards documents also organize the knowledge and skills into coherent and well-integrated accounts of events and processes in the physical and biological world. These standards provide guidance for developing curriculum, instruction, and assessment at CCMS.

Although the standards provide considerably more detail than a topic list does, they are also acknowledged to be just the first step in defining what students should know. The accounts of natural processes in the standards documents are not intended to be complete, and by themselves the statements do not provide enough information to assessment developers about exactly what students can or should be held accountable for. Therefore, to increase precision of content alignment and precision in diagnosing gaps in students' knowledge that may interfere with their understanding of natural phenomena, Project 2061 researchers subdivide the content standards into finer grained statements of knowledge, or *key ideas*. The key ideas are then clarified and elaborated to specify what the boundaries of that knowledge are for purposes of assessment. Consider the following key idea for a benchmark from the topic of plate tectonics:

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

Clearly there are concepts in this statement about earth's plates that need to be described beyond the simple statement that the outer layer of the earth is made of plates. Exactly what knowledge should students have of what a plate is? Researchers used three questions to guide clarification of the key idea:

1. Is this description of plates what is needed for students at this age to form a mental image of the Earth's plates that will allow them to predict and explain phenomena involving plates?
2. Is this description of plates what 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?

The *clarification statement* answers these questions to provide a more detailed description of what students are expected to know:

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.

In addition to providing detailed guidance for assessment, 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 will use when learning subsequent ideas about plate motion and the consequences of plate motion, which come later in the instructional sequence. In particular, this mental model will help students understand such things as mountain building and where earthquakes and volcanoes form when the students are introduced to those ideas. With respect to terminology, it was decided for assessment purposes not to expect students to know certain technical terms such as lithosphere, because these terms were not likely to contribute significantly to explaining phenomena related to plate motion. Although individual teachers may choose to show students the relationship between lithosphere, upper mantle, and plates during instruction, the assessment items do not include the term.

Expectations for students are also based on what research on student learning tells us regarding the age-appropriateness of the ideas being targeted and the level of complexity of the mental models that students can be expected to develop. Research on student learning also describes many of the misconceptions that students may have, which are then included as distracters in the items so that these nonnormative ideas can be tested alongside the targeted ideas.

After the ideas that are to be assessed have been identified and clarified, the next step is to determine how these ideas relate to other ideas within a topic and

across grade levels. The objective here is to be as clear as possible about the boundaries around the ideas that are being explicitly tested and the prior knowledge students can be assumed to have. For example, if students are being tested on their understanding of digestion at the molecular level, in which molecules from food are broken down to simpler molecules that have a different number and arrangement of atoms, can it be assumed that students already know that molecules are made of atoms? If not, are questions on chemical digestion, written in terms of atoms or molecules, to some extent also testing whether students know the relationship between atoms and molecules?

In making judgments about which ideas precede a targeted idea, Project 2061 researchers make use of the conceptual strand maps published in the *Atlas of Science Literacy* (AAAS, 2001, 2007). The strand maps were developed to visually represent the interconnections among ideas in *Benchmarks for Science Literacy* and the *National Science Education Standards*. The map for the topic of “Diversity of Life” in Figure 7.1, for example, has three strands: classification, similarities and differences, and diversity and survival. 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 and the links between ideas across strands. In the *diversity and survival* strand, a benchmark at the 6–8 grade level says, “In any particular environment, the growth and survival of organisms depend on the physical conditions” (AAAS, 2007, p. 31). This is preceded on the map by a benchmark at the 3–5 grade level that says, “For any particular environment, some kinds of plants and animals thrive, some do not live as well, and some cannot survive at all” (AAAS, 2007, p. 31). In testing whether students know that organisms in ecosystems depend on the physical conditions, it is assumed that they already know that not all organisms are as successful as others in a given ecosystem. As a rule, unless there are good reasons to believe otherwise, it is assumed that students already know the ideas listed at an earlier grade band, and the ideas and language from those earlier ideas are used freely in item development for the grade band that follows. But it is also recognized that these earlier ideas are a good place to look when students do not know a targeted idea. Not knowing an earlier idea, such as the idea that not all organisms are as successful as others in an ecosystem (or the idea that molecules are made of atoms), is often the reason why students have difficulty with the idea being tested. The relationships identified in the *Atlas* maps (see Figure 7.1) can help developers focus on ideas that may be needed for understanding the targeted ideas and encourage them to think about whether or not it is reasonable to assume that students already know that earlier idea.

Design Assessment Items Aligned with Content Standards. Test items should always be written in such a way that teachers and researchers can draw

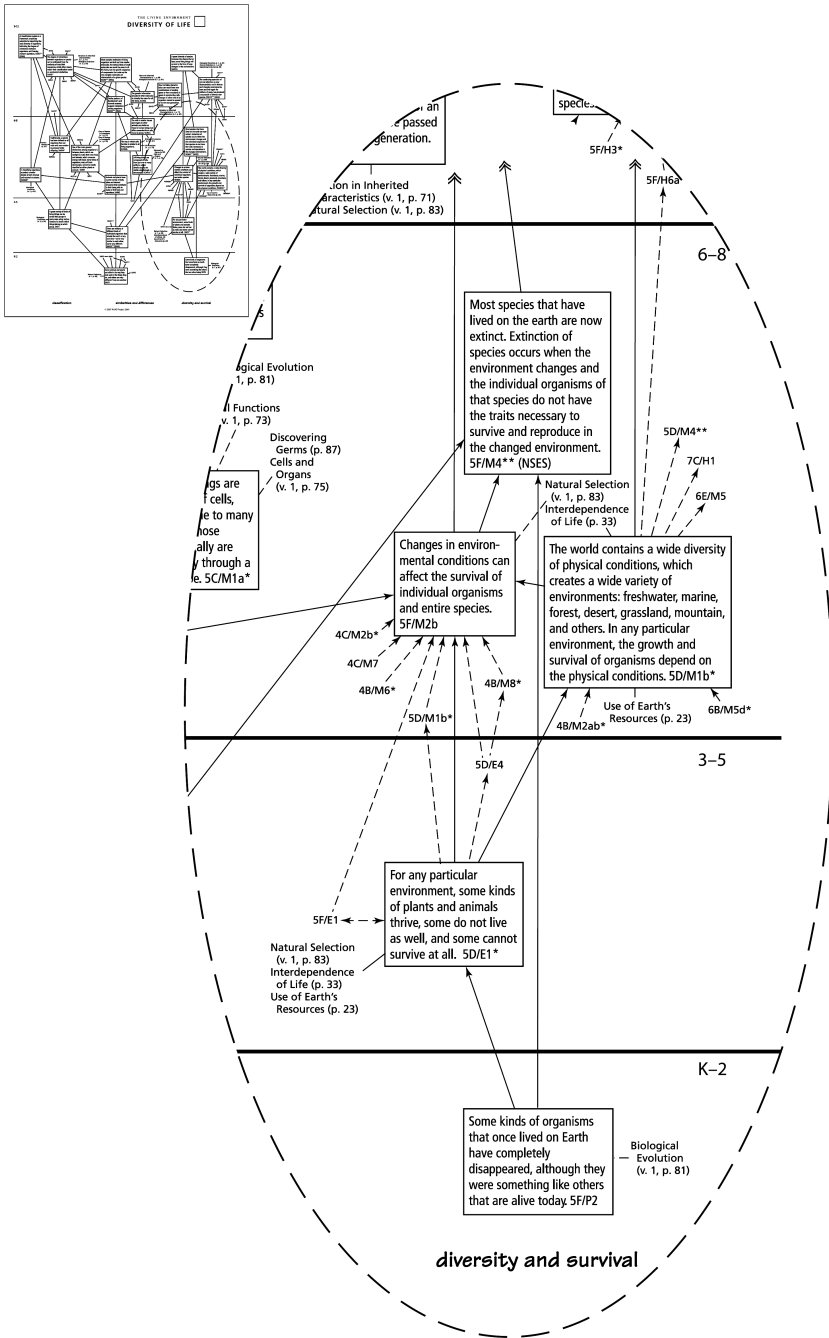


FIGURE 7.1 *Atlas of Science Literacy* strand map for the topic “Diversity of Life” (AAAS, 2007, Map 5A).

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 whether 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 an idea, and they should choose an incorrect answer only when they do not know the idea.

The first thing to consider when designing test items is the alignment of the item with the targeted idea. Project 2061 researchers use two criteria to determine whether the content that is being targeted by an assessment item is aligned with the content specified in a particular key idea. The *necessity* criterion addresses whether the knowledge 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. If the targeted knowledge is not needed to answer the question, then the item is obviously not a good indicator of whether students know the 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 criteria of necessity and sufficiency are used both in the initial design of the items and in subsequent analyses and revisions of the items. (See DeBoer [2005] for a further discussion of the application of the necessity and sufficiency criteria in item development.)

Despite its critical importance in item development, content alignment alone is not enough to determine whether or not an item should be used. There are many other factors related to construct validity that can also affect the usefulness of an assessment item in providing accurate insights into student understanding of the targeted content. For example, students should not be able to answer correctly by using test-taking strategies that do not depend on knowing the idea (a false positive response) or be so confused by what is being asked that they choose an incorrect answer even when they know the idea being tested (a false negative response). To improve the validity of conclusions that can be drawn from assessment results, it is important to identify and eliminate as many problems with comprehensibility and test-wiseness as possible. (The criteria for judging alignment and threats to validity are detailed in Project 2061's Assessment Analysis Procedure, available on the Project 2061 Web site at <http://www.project2061.org/assessment/analysis>.)

Obtain Feedback from Students. Rigorously applying a set of criteria to determine the alignment of test items with learning goals and to identify features that obscure what students really know are both important steps in the item development process. However, findings from research indicate that this analytical approach works much more effectively when used in combination with one-on-one interviews with students or pilot tests of items in which students are asked

to explain why they chose the answer that they did (DeBoer & Ache, 2005). By comparing the answer choices that students select with their oral or written explanations, it is possible to determine if an assessment item is measuring what it is supposed to measure or if students are giving false negative or false positive responses to the item. In the Project 2061 item development work, students are asked the questions shown in Figure 7.2 to get feedback on problems they may have in interpreting the items and to find out what ideas they have about the content being assessed, especially the misconceptions they have. Pilot tests are carried out in urban, suburban, and rural middle schools serving a wide range of students (DeBoer, Herrmann Abell, & Gogos, 2007).

The following examples illustrate the kinds of information that can be obtained from these pilot tests. The examples also show how this information can be used to improve the items' alignment with the key ideas and improve their validity as measures of student learning.

Example A: Atoms, Molecules, and States of Matter. The item shown in Figure 7.3 tests whether students know that molecules get farther apart when they are heated and whether they know that this molecular behavior explains why most substances expand when heated. The item includes answer choices that test

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 to 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? Where? (TV, museum visit, etc.?)	Yes	No	Not sure

FIGURE 7.2 Project 2061 questionnaire for student interview on an assessment item.

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

FIGURE 7.3 Project 2061 assessment item development questionnaire for students: Atoms, molecules, and states of matter (Coffey, Douglas, & Stearns, 2008).

common misconceptions related to thermal expansion and the behavior of molecules, especially the idea that there are “heat molecules.”

Pilot testing showed that 25.9% of the students answered this question correctly. The most common response (44.4%) 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. Based on the results of pilot testing, the following revisions were made: First, because answer choice A is the only one that has the word “heat” in it and students may choose that answer choice because they connect the liquid rising in the thermometer with heat rising, the word “heat” was added to other answer choices. Also, the word “alcohol” was

changed to “liquid” to remove a word that some students find confusing in the context of thermometers.

When students were interviewed about this item, a number of them had difficulty reconciling what they expected to be a very small expansion of the liquid in the bulb of the thermometer 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 chose answer choice A (“heat molecules”). 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 easily engage in proportional reasoning in middle school, the thermometer context may be a difficult context for general testing of middle school students’ understanding of thermal expansion. But it is also possible that focused instruction might help students see that a small change in the volume of a liquid is amplified in a narrow tube. The thermometer could then be used as an example of how measuring devices in general are often designed to amplify the effect being measured. Often a close examination of student reasons for selecting answer choices during assessment leads to insights that can be applied to instruction as well as to assessment.

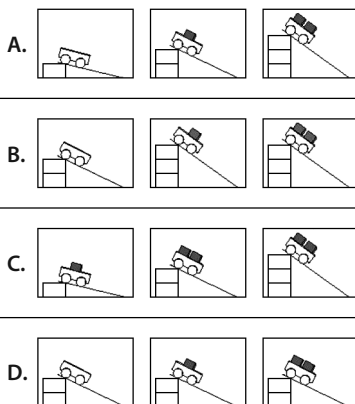
Example B: Control of Variables. The item shown in Figure 7.4 was developed to determine whether students understand that the way to determine if one variable is related to another is to hold all other relevant variables constant. The item also tests a number of common misconceptions that students often have regarding the control of variables, including the idea that all of the variables should be allowed to vary in a controlled experiment.

The results of pilot testing showed that 53.9% of the students answered correctly and that 26.3% chose answer choice A, which targets the misconception that both variables should vary at the same time. Answer choices B and C were less popular distracters. Answer choice B was chosen by only one student. Of the six students who chose C, three 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 (including 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 blocks in the carts were replaced by metal balls and the number of balls in each cart was increased so that there were no empty carts. The stem was changed to read, “Which set of tests should he compare?”

External Review. Following pilot testing, the items are reviewed by panels of experts in science and science education to ensure content accuracy, alignment with the targeted content standards, and construct validity. The reviewers are

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?



Students who chose each answer:

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

FIGURE 7.4 Project 2061 assessment item development questionnaire for students: Control of variables (Coffey et al., 2008).

trained in the use of Project 2061's Assessment Analysis Procedure (see <http://www.project2061.org/assessment/analysis>), which uses the criteria of necessity and sufficiency for content alignment, and focuses reviewers' attention on issues of test-wisness, comprehensibility, accessibility, and appropriateness of task context. The reviewers also make use of the results of pilot testing, including students' written comments, to help them in the analysis of the items. The reviewer ratings and comments, combined with the pilot test data, are then examined by Project 2061 staff, and the items are then revised and prepared for field testing.

Field Testing. Each test item is field tested in topic clusters of 12–25 items per cluster with a national sample of approximately 1,000 students from a wide range of urban, suburban, and rural schools across the country. The results are analyzed

using item response theory (IRT) and classic test item analysis procedures to determine relative difficulty of the items, the factor structure of the cluster of items, the contribution that each item makes to assessing student understanding of the topic, and whether or not the items function similarly for the various subgroups tested. Data are analyzed by gender, race and ethnicity, and whether English is the students' primary language.

For example, on a set of 14 items designed to measure middle school students' understanding of control of variables, it was found that boys and girls performed equivalently on the set of items as a whole. However, although they performed equivalently on most of the individual items, some of the individual items functioned differently for boys and girls. These differences demonstrate how important it is to be aware of item context when writing test questions. Test writers may use their best judgment in predicting how boys and girls (or members of different ethnic groups) will respond to the scenarios they create, but an analysis of actual test results is needed to see if those predictions are accurate. Providing test items that are accessible to a wide range of students and on which subgroups of students will perform similarly is an issue of equity and one of the goals of this assessment work. Equity is also a core principle of CCMS and TELS. (See Chapter 4, Tate, Clark, Gallagher, and McLaughlin, for further discussion of how equity issues are addressed at the two centers.)

Connecting Science Content to Science Practices

In the second example, we illustrate how CCMS researchers have designed assessment items to measure student ability to engage in scientific practices involving real-world phenomena related to a targeted science idea (Harris et al., 2006; McNeill & Krajcik, 2008a). The stated goal is for students to be able to use their knowledge of science content as they engage in a particular science practice. The example is taken from the "Investigating and Questioning Our World Through Science and Technology" (IQWST) middle school curriculum unit focusing on properties of substances and chemical reactions (McNeill et al., 2004). The particular science practice that is targeted is the ability to provide a scientific explanation for a claim the student makes based on data provided.

The assessment task shown in Figure 7.5 requires students to apply two science ideas. First, they must know that different substances have different characteristic properties. Second, they must know that a property such as density, color, or melting point is a characteristic property of a substance that does not change when the amount of the substance changes. The students are judged on the basis of their ability to (a) make accurate claims from the data using their knowledge of properties of substances, (b) justify their claims about these substances using the available evidence, and (c) provide a reason why the evidence

	Density	Color	Mass	Melting Point
Liquid 1	0.93 g/cm ³	No color	38 g	−98 °C
Liquid 2	0.79 g/cm ³	No color	38 g	26 °C
Liquid 3	13.6 g/cm ³	Silver	21 g	−39 °C
Liquid 4	0.93 g/cm ³	No color	16 g	−98 °C

Write a scientific explanation that states whether any of the liquids are the same substance.

FIGURE 7.5 IQWST assessment item: Scientific explanation for properties of substances.

justifies the claim (i.e., link their claim and evidence to the general rule or scientific principle that is being targeted). This claim-evidence-reasoning format is taught explicitly in the IQWST curriculum (see Chapter 3, Krajcik, Slotta, McNeill, & Reiser) and is tested using items such as the one in Figure 7.5.

Identify and Unpack the Content Standard. The first step in writing an IQWST assessment task is to identify the content standard to be assessed. The content standard is then unpacked to identify exactly which science ideas are to be tested, those statements are elaborated, and boundaries are set around the expectations for students. What comes earlier and later in the learning trajectory is also noted so that specific expectations for the age range can be determined. Item developers also consider which nonnormative ideas or misconceptions students may have about the content, and these misconceptions are then used in the design of the assessment task either as distracters in multiple-choice questions or as part of a scoring rubric for open-ended questions.

Unpack the Scientific Inquiry Practice. The next step that IQWST developers take is to consider which scientific inquiry practices (e.g., modeling, designing an investigation, or providing a scientific explanation) they want to measure. In the chemistry example in Figure 7.5, the science practice that was chosen was scientific explanation. This science practice was unpacked into three separate components: claim, evidence, and reasoning. Unpacking the scientific practice specifies what it is that students will be expected to do with their understanding of the science content. Each constructed response assessment task requires a separate scoring rubric for each content area and each scientific inquiry practice, although a base rubric can be developed for a particular science practice and applied to different contents (McNeill et al., 2006).

Create Learning Performances. Learning performances are then developed to make explicit what students should be able to do with the targeted content knowledge (Krajcik et al., 2008). A learning performance combines both a content standard and a science practice. In the case of the chemistry example (Figure 7.5), the learning performance clarifies how science principles related to properties of substances are to be used in reasoning about scientific phenomena involving substances and their properties.

Write the Assessment Task. The next step is to design assessment tasks that ask students to apply both their content knowledge and their ability to engage in scientific explanation.

Review and Revise the Assessment Task. After creating the assessment task, three questions adapted from Project 2061's assessment framework (DeBoer, 2005; Stern & Ahlgren, 2002) are used to review the assessment tasks.

1. Is the knowledge *needed* to correctly respond to the task?
2. Is the knowledge *enough by itself* to correctly respond to the task, or is additional knowledge needed?
3. Are the assessment task and context likely to be *comprehensible* to students?

These questions help to determine whether the assessment task aligns with the desired learning goal and whether or not it will be accessible to the students.

Develop Specific Rubrics. The next step is to create a rubric for each assessment task by determining what counts as appropriate application of the science practice in the context of the question being asked. In this case, the scoring rubric is used to evaluate the appropriate use of scientific explanation (claim, evidence, and reasoning) for a question involving the properties of substances.

The approach described here can be applied to many other combinations of science content and science practice. (See McNeill and Krajcik [2008a] for an example of scientific explanations dealing with predators and prey.) In addition to providing scientific explanations of phenomena, students can also be tested on their ability to use science ideas to predict what will happen given a set of conditions, generate physical models of abstract science ideas, or use their knowledge of science ideas to design investigations.

Measuring Students' Ability to Link Ideas Together

In the third example, we describe how the TELS knowledge integration construct is used to guide item development and scoring processes. TELS researchers have developed assessment items that act as prompts for students to connect multiple

science ideas together to explain scientific phenomena they encounter in the real world, which is a central component of the knowledge integration construct. By examining the scientific relevance of students' ideas and the connections they make between those ideas, different levels of integrated understanding can be identified. (See Chapter 2, Roseman, Linn, and Koppal, for a detailed discussion of the knowledge integration construct and of the notion of integrated understanding.) Student explanations are coded for the following levels of knowledge integration:

- *Level 0: No Information.* If students do not provide an answer to an item, their understanding is scored at the “no information” level.
- *Level 1: Irrelevant.* If students answer an item incorrectly, perhaps because they do not have the knowledge relevant to the item or the motivation to take the item seriously, and if they show no evidence of understanding, their level of understanding is “irrelevant.” These students may provide nonsense explanations that have nothing to do with the science context being described, descriptions that reveal their lack of motivation such as “I do not like science,” or confessional statements such as “I do not know.”
- *Level 2: No Link.* If students appear to have made an attempt to answer an item with scientifically invalid reasons based on nonnormative ideas or links, their understanding is scored at the “no link” level.
- *Level 3: Partial Link.* If students respond to an item with at least some relevant and correct ideas but do not meaningfully connect the ideas to the task, their understanding is scored at the “partial link” level. To illustrate, consider the Spoon Item, which asks students to explain why a metal spoon just taken from a cup of hot water feels hotter than a wooden or plastic spoon. Students might say “metal is a heat conductor.” Although the statement is correct, without explicitly comparing the heat conductivity of metal with that of plastic or wood, the statement does not explain why the metal spoon feels hotter than the other materials.
- *Level 4: Full Link.* If students explicitly link a set of correct and relevant ideas to the item, their understanding is scored at the “full link” level. In the case of the heat conductivity of the spoons, they would indicate first that heat conductivity is a relevant concept to consider and second that metal has the highest conductivity of the three materials.
- *Level 5: Complex Link.* If students can meaningfully link three or more normative and relevant ideas together in answering an item, their understanding is scored at the “complex link” level. For instance, on the Spoon Item a student might indicate that heat conductivity is a relevant concept, that metal is a much better heat conductor than wood or plastic, and that better heat conductors both absorb and release heat faster when they come in contact with other objects, such as a hand.

- *Level 6: Systemic Link.* If students can systematically apply a particular science concept to multiple contexts by recognizing common features relating to the particular science concept, their understanding is scored at the “systemic link” level. For example, students at this level can consistently explain the heat conductivity concept in contexts using a variety of combinations of materials including gases, liquids, and solids. This level of knowledge integration is determined by examining student responses to a number of items that address the same science concept across different contexts.

Figure 7.6 summarizes the relationship between hypothesized levels of the knowledge integration construct and scoring of the items on the knowledge integration scale. The placements of “incorrect” and “correct” responses to multiple-choice items on the scale in Figure 7.6 are estimations. Because there are only two (dichotomous) scoring levels for multiple-choice items, we assume that students with relevant ideas would provide a correct answer to a multiple-choice item, whereas students lacking relevant knowledge (irrelevant) or understanding (no link) would provide incorrect answers. We will discuss whether the TELS assessment data support this assumption in the next section.

To obtain information about knowledge integration levels, TELS researchers use concrete contexts in which key scientific ideas are elicited to solve real-world problems or explain phenomena (Lee & Songer, 2003). Items based on abstract contexts, esoteric terminology, or simple computations are typically not used because they are less effective in eliciting what students know or providing them with opportunities to link ideas together. The TELS items also make use of tables, graphs, pictures, and models because these devices help students visualize the situation presented in the item. These strategies are expected to make items more accessible to students and to provide more complete information about their knowledge and its connection to the problem being presented.

TELS researchers use an argumentation structure to frame many of their items by asking students to choose an answer to a question and then provide a written explanation to justify their choice. To illustrate, consider the item shown in Figure 7.7A. The Spoon Item consists of a multiple-choice question and a follow-up question that asks students to explain their answer to the multiple-choice question. The stem reads, “A metal spoon, a wooden spoon, and a plastic spoon are placed in hot water. After 15 seconds which spoon will feel hottest?” The four options provided are “the metal spoon,” “the plastic spoon,” “the wooden spoon,” and “the three spoons will feel the same.” This multiple-choice question is an item released by the Trends in International Mathematics and Science Study (TIMSS, 1995) for public use. To this multiple-choice question, TELS researchers added the follow-up portion: “Explain your choice.” Providing students with choices such as “metal,” “wooden,” and “plastic” spoons and “the three spoons will feel the same” facilitates their search for their knowledge relevant to the item.

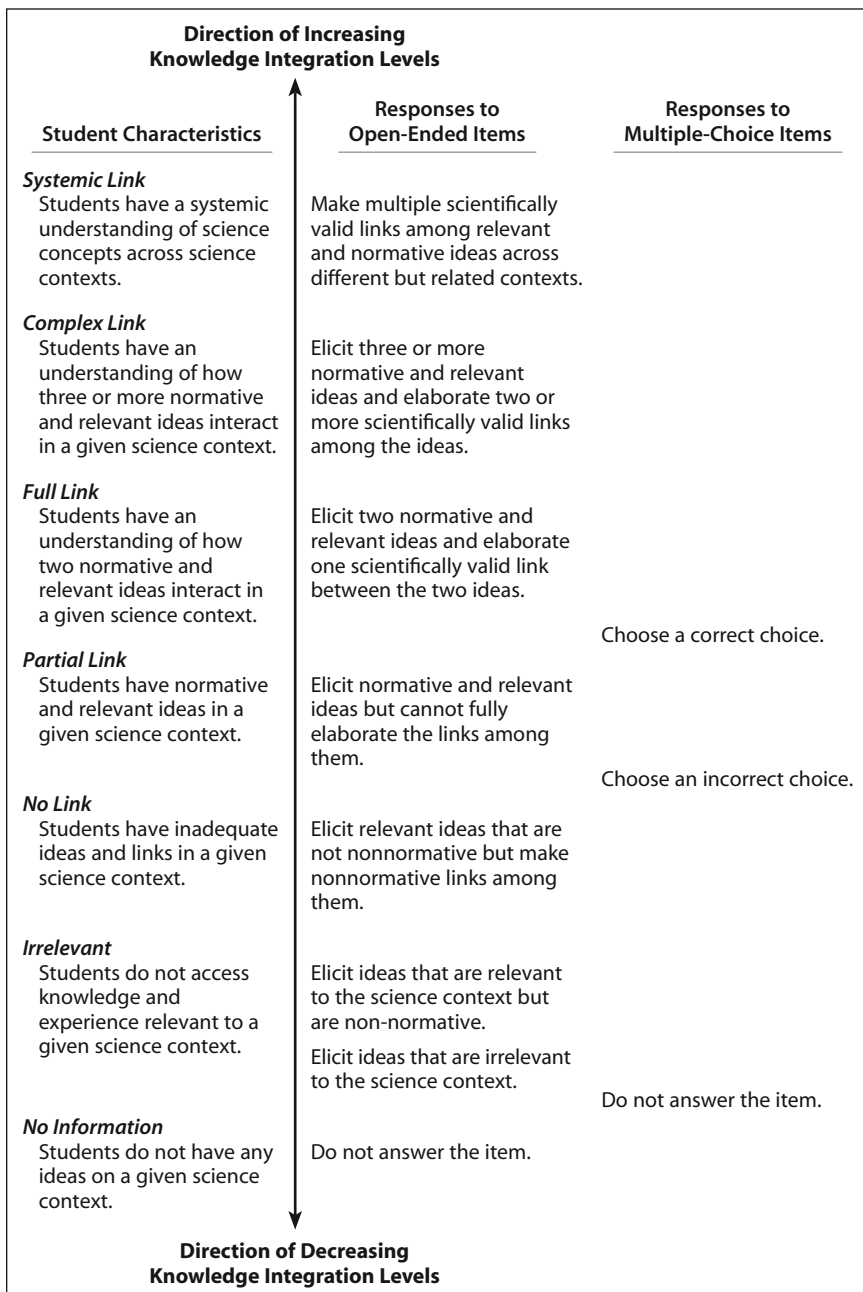


FIGURE 7.6 TELS knowledge integration construct map.

A	A metal spoon, a wooden spoon, and a plastic spoon are placed in hot water. After 15 seconds which spoon will feel hottest?			} Item Stem
	(a)	The metal spoon	} Claim: Multiple-Choice	
(b)	The wooden spoon			
(c)	The plastic spoon			
(d)	The three spoons will feel the same			
	Explain your choice.			} Explanation: Open-Ended
	<i>Metal is a much better conductor than wood or plastic so it would both gather heat quicker and exert that heat faster on to you when you touch it.</i>			
B	Score	Level	Description	Examples
	0	No answer	• No information is gathered about the student	• Blank
	1	Irrelevant	• Mentions experience, opinion, or interest considered not relevant to the science context	• I don't know • I do not like science
	2	No link	• Refers to nonscientific personal experience, opinion, or interpretation • Includes only non-normative ideas • Includes only non-normative links	• Because when a metal spoon gets hot it stays hot for a little while • The metal spoon traps heat the best and will stay hot longer • Because the metal attracts the heat • The metal has atoms that transfer heat and wood and plastic don't have many heat transfers
	3	Partial link	• Mentions heat absorption ability of one material AND • Does not compare with the other materials • Does not consider heat absorption over time (heat transfer rate)	• Because metal absorbs heat more than wood or plastic • The metal spoon because metal heats up very much in a small amount of time
	4	Full link	• Compares heat transfer rates (e.g., conductor, heat absorption rate, heat absorption over time) among three materials	• The metal gets hot the fastest • Metal transfers heat faster than plastic or wood
	5	Complex link	• Compares heat transfer rates in absorbing heat from the hot water to the spoon and transferring heat from the spoon to the hand	• Metal is a much better conductor than wood or plastic so it would both gather heat quicker and exert that heat faster on to you when you touch it

FIGURE 7.7 Spoon Item: (A) Multiple-choice question (TIMSS, 1995) and follow-up explanation question; (B) Knowledge integration scoring rubric.

This strategy enhances the “outcome space” (Wilson, 2005, p. 63) of written explanations and reduces the likelihood of responses that are completely unrelated to the scenario described in the item stem. From the explanation portion of the item, it is then possible to determine how many relevant ideas students reveal and how well they connect those ideas.

ANALYZING DATA TO OBTAIN PSYCHOMETRIC PROPERTIES OF TEST ITEMS

Our discussion up to this point has focused on the development and use of individual test items to reveal what students do and do not know about particular science ideas and their ability to apply that knowledge in the solution of real-world problems. We now continue the discussion by describing how psychometric analysis can be used in item and instrument development. Although similar analyses are made in both the TELS and Project 2061 item development work, the examples discussed here are from the TELS work. The examples are used to demonstrate how some items are more useful than others in estimating students’ knowledge integration abilities.

To compare how each item in a test contributes to the measurement of knowledge integration, TELS researchers conducted IRT analyses based on the Rasch partial credit model (Liu et al., 2008). IRT analyses produce various statistical outputs that allow assessment developers to inspect how each item functions compared with the other items in a test, as well as how the test functions as a whole to measure the underlying construct. This section illustrates analysis of a middle school physical science test that included 2 short-answer items, 12 multiple-choice items, and 7 explanation items. Seven of the 12 multiple-choice items were linked to the 7 explanation items. An example of an item with linked multiple-choice and explanation questions is the Spoon Item shown in Figure 7.7A. The full list of items can be found at the TELS Web site (<http://www.telscenter.org>). Half of the test addressed heat and temperature concepts, and the other half addressed kinematics concepts such as distance, velocity, motion, and force. Three multiple-choice items and two short-answer items were selected from released standardized tests such as TIMSS and the National Assessment of Educational Progress (NAEP). The rest of the questions were designed by TELS researchers. The set of these 21 items had a Cronbach’s alpha reliability coefficient of .84.

The discussion that follows focuses on the information provided in the Wright map shown in Figure 7.8. The numbers on the left side of the Wright map represent the knowledge integration scale. The higher the number, the higher the knowledge integration level. For example, students at the -1.0 level have a lower estimated knowledge integration ability than those at the 1.0 level. The left side of the Wright map also shows the knowledge integration ability distribution of all students who took the test ($N = 1,129$). Each “x” on the map

represents about seven students, and their position on the scale is determined by their performance on the set of 21 questions. Most of the students fall between -1.0 and 1.0 .

The distribution of students on the knowledge integration scale also can be used to show the relative difficulty of the items. Because the multiple-choice items are scored as either correct or incorrect, there is just a single difficulty location marked with a “C” for each item. The first multiple-choice item (M1) has a location “C” at 0.82 , which means that students with a knowledge integration ability estimate of 0.82 will have a 50% chance of getting the M1 item correct. From its position at 0.82 , we can see that the item is somewhat more difficult than some of the other items, because students with an ability estimate of 0.82 , which is considerably above the average, have only a 50% chance of getting the item correct. Students with a knowledge integration ability estimate higher than 0.82 will have a greater than 50% chance of getting the M1 item correct, and those with a knowledge integration ability estimate lower than 0.82 will have a less than 50% chance of getting this item correct. In contrast to the multiple-choice items, the explanation items have five score locations on the map because they were scored on a scale of 0 to 5. A location of “1” represents the transition from getting a score of 0 to a score of 1 or higher on the knowledge integration scale; “2” represents the transition from getting a score of 1 to a score of 2 or higher; and so forth.

Figure 7.8 shows the item difficulties of the 19 items used in the TELS physical science test, plus the 2 short-answer items that were used for comparison between multiple-choice and explanation item formats. From their positions on the distribution, we can determine whether the student responses collected from the test match the conceptualization of the knowledge integration construct shown in Figure 7.6. Results show that although the locations “1” and “2” are very close to each other for most explanation items and could probably be represented as a single level, the locations for all of the items on the map match the order of increasing levels of knowledge integration ability.

To further illustrate the knowledge integration spectrum, the bars on the right side of the Wright map diagram represent the range of ability estimates for the entire set of 21 items. See, for example, that the range of ability estimates for students who were scored at level “3” is from -0.03 (KE3) to 1.84 (KE7), and that this is represented as the “partial link” bar in the right column. The diagram also shows that there is a distinct ordering of the levels, which matches the initial conceptualization of the knowledge integration construct shown in Figure 7.6.

From the Wright map analysis, it is also possible to obtain information about how each item contributes to the measure of the knowledge integration construct. First, it can be seen that for the most part, the multiple-choice items do not effectively estimate high levels of knowledge integration at the full and complex link levels. (Note the positions of the “C”s for each multiple-choice item on the Wright map.) For multiple-choice items to be located at the higher levels on

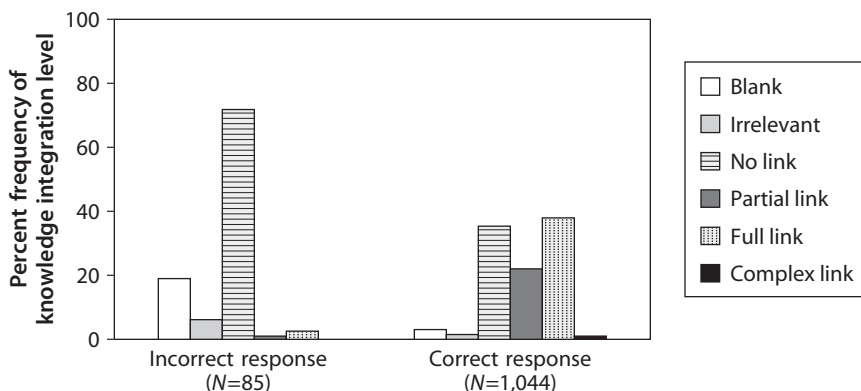
the map, they would have to require students to draw upon a set of ideas in the solution of a problem (or, at least, scores on those items would have to be correlated with the ability to make such connections). Nor do the multiple-choice items match well where students fall on the knowledge integration scale as measured by the corresponding explanation items. For example, although 92.5% of the students answered the multiple-choice Spoon Item *correctly*, suggesting that they understood the science content, when their written explanations for the Spoon Item were analyzed using the scoring rubric shown in Figure 7.7B, 39.9% of those students could not adequately explain their multiple-choice answer and were placed in the no link level or below on the knowledge integration scale (see Figure 7.9A).

Of the 85 students who answered *incorrectly* on the multiple-choice question, the match between student answers on the multiple-choice and explanation questions is much better. Here, 82 of the 85 students (96.5%) who answered the multiple-choice question incorrectly are also placed in the lowest levels of the knowledge integration scale based on their answer to the explanation question (see Figure 7.9A). These students chose the incorrect answer (or left it blank) on the multiple-choice question and also showed a lack of knowledge in their written explanations. This demonstrates that this item is more useful as an indicator of what students do not know than of what they do know.

Although the Spoon Item context is often used in research on heat conductivity (Clough & Driver, 1985; Erickson, 1979; Harrison, Grayson, & Treagust, 1999), the multiple-choice portion of this item did not prove to be useful for differentiating better from worse understanding of heat conductivity. As shown on the Wright map (item KM1, Figure 7.8) and in Figure 7.9B, even students whose ability estimate is at the extreme low end of the knowledge integration scale (as determined by the full set of questions) still have a better than 50% chance of getting this multiple-choice question correct. Although some of the students who answered correctly may have an understanding of the targeted idea, many do not. This suggests that this particular item has limited usefulness on an instrument that measures knowledge integration. The reason so many students got this item correct without understanding the science concept involved is probably because many students know from experience that when a metal spoon is used to stir a liquid it feels hotter to the touch than a spoon made of plastic or wood, without knowing why that is true. When there is a mismatch between the multiple-choice portion of an item and the explanation portion of an item, the item can be examined to find out why the mismatch occurred. As was discussed in the first part of this chapter, there are a variety of reasons why multiple-choice items may yield false negative or false positive responses, but there are also ways to correct many of those problems.

Other multiple-choice items in the TELS instrument produced results that more closely matched what was learned from the explanation items. The multiple-

A What knowledge integration levels are represented in correct and incorrect responses to the multiple-choice question?



B Does high knowledge integration lead to a correct response to the multiple-choice question?

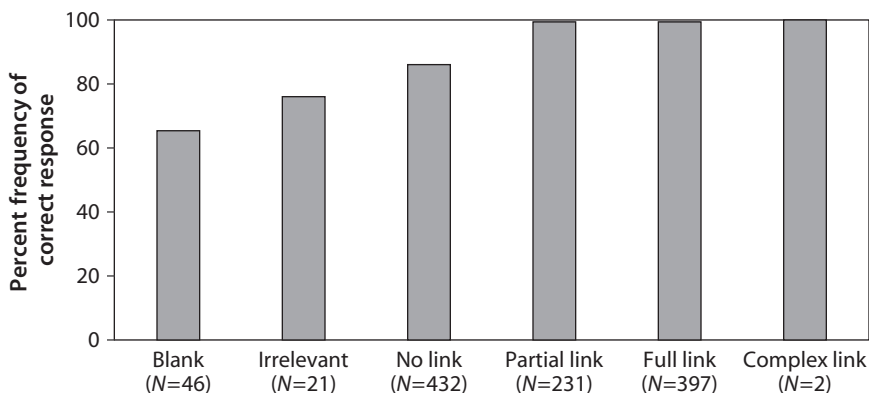


FIGURE 7.9 Spoon Item: Effectiveness of the multiple-choice question as an indicator of knowledge integration.

choice item KM5, for example, was a difficult one for students, and the results closely matched what was learned from the explanation portion of the question, KE5. In the multiple-choice part of the question students were given a choice of three graphical representations of a situation involving rates of motion. To answer correctly, students had to understand the concept of rate and how to

TABLE 7.1 Percent Frequency Distribution of Knowledge Integration Levels Assessed with Explanations*

Level	KE1	KE2	KE3	KE4	KE5	KE6	KE7
Complex Link	0.2		2.8	1.4	0.4	5.5	0.3
Full Link	35.2	2.3	19.7	21.0	7.5	11.3	2.04
Partial Link	20.5	25.8	33.6	16.9	32.3	23.7	10.4
No Link	38.3	54.3	29.9	50.5	27.4	40.7	59.6
Irrelevant	1.9	6.3	2.9	2.6	10.2	4.1	7.3
Blank	4.1	11.3	11.2	7.6	22.1	14.9	20.5

*KE1–KE7 are the explanation items in the TELS physical science test summarized in the Wright map in Figure 7.8. $N = 1,129$ students.

represent it graphically. Students who answered this item correctly scored between “3” and “4” on the knowledge integration scale based on their answer to the explanation portion of the question, indicating that a fairly sophisticated understanding of the idea being tested was needed in order to get the multiple-choice item correct.

Wright maps also can be used to indicate whether individual items provide the opportunity for students to demonstrate their proficiency along the continuum of the construct being measured. For this set of items, estimates of a student’s knowledge integration ability tend to fall in the middle and lower ranges (irrelevant, no link, partial link), with a smaller number of students in the upper range (full link, complex link). For item KE2, for example, which tests students on the difference between heat and temperature, a very small percentage of students scored at the full link level and no students at the complex link level of knowledge integration, making this explanation item a difficult one on which students could demonstrate an integrated understanding, compared with some of the other items. (See Table 7.1 for the percentage of students at each knowledge integration level for each explanation item in the TELS physical science test.)

For an item to elicit the higher levels of knowledge integration, the item must describe a situation that is complex enough to make use of multiple ideas. When an item does not elicit the desired level of knowledge integration compared with what other items elicit, the item can be eliminated or modified. Of course, failure of students to score at a high level probably also indicates that the students had not achieved the goal of knowledge integration, which points to how important it is to develop and implement curriculum and instruction materials that can help students develop a more integrated understanding of science.

USING ASSESSMENT RESULTS TO MODIFY CURRICULUM MATERIALS

In this section we describe how assessment is used at TELS and CCMS to inform the revision of curriculum materials. First we show how TELS researchers use assessment results to modify their online modules, and then how researchers at CCMS use assessment results to make modifications in the IQWST units. Descriptions of the design principles that led to the development of these materials and the learning environments in which they are used appear in Chapter 3 and Chapter 8 (Kali, Fortus, & Ronen-Fuhrmann).

TELS Curriculum Modification Examples

TELS modules focus on particular content that teachers had identified as being difficult for students to learn with typical instructional methods (Chapters 2 and 3). Using the design principles described in Chapter 8, instructional modules were developed to support student learning of this science content. The modules went through a number of development, implementation, assessment, and revision cycles. Changes to the TELS modules were made following a review of the assessment data, observations of student use of the modules, and teacher input regarding their first-hand experience with the modules' effectiveness (for more details about the TELS review process, see Chapter 2). The effectiveness of each TELS module was examined based on the results of tests administered before and after implementation of the module. The test items were designed following knowledge integration design patterns (Linn & Eylon, 2006) and design principles (Kali, 2006) and were scored according to the knowledge integration construct map (Lee, Liu, & Linn, 2008) shown in Figure 7.6.

For instance, in the TELS "Airbags" module (see detailed description in Chapter 3), students learn about graphical representations and interpretations of velocity, force, and acceleration concepts in an airbag deployment situation. Based on the assessment results of the first version of the "Airbags" module, two modifications were made (McElhaney, 2007a). First, the pretest and posttest results for the original version of the module indicated that these mechanical concepts were too easy for most grade 11 and 12 students. Based on this observation, the "Airbags" module was revised to increase content complexity, for example, by adding velocity–time graphs to the position–time graphs that were already in the module. These additions were made both to the simulation activities in the module and to the embedded assessments. Second, analyses of pretest and posttest data, and the logging of data collected during students' simulation experiments in the original module, demonstrated a significant positive relationship between students' sophistication with simulation experimentation strategies

and their posttest achievement. Based on this finding, steps were added to scaffold student reflection on the activities.

Another example is the “Recycling” module, in which students learn about different types of chemical bonds (e.g., covalent, ionic, metallic, and van der Waals interactions) present in the materials of real-world objects. After the first run of the “Recycling” module, students demonstrated difficulties distinguishing among different types of bonds. As a result, in the revised “Recycling” module, reflection prompts were modified to elicit more connections and contrasts between materials with these different types of bonds so that students could better link the microscopic-level visualizations to their macroscopic-level observations (McElhaney, 2007b).

The “Asthma” module described in Chapter 4 includes physiological explanations of breathing, asthma attacks, and asthma as an allergic immune response. Results of pretests and posttests indicated that students had improved their understanding of the physiology of asthma related to breathing processes and asthma attack mechanisms after completing the module (Tate, 2007). However, it was also observed that only a few students wrote explanations that meaningfully connected the various components of the allergic immune system. For example, most students described the allergic immune response as an isolated irritant response rather than a physiologically triggered chain reaction. This finding indicated that the original version of the “Asthma” module did not adequately support students in distinguishing between allergens and other irritants in terms of their effects on the body and its immune system. In the next stage of development, the design team modified the module to focus more on that distinction. An ensuing evaluation suggested that the modified module improved students’ ability to develop criteria for distinguishing the physiological effects of allergens to a greater degree than did the original module.

CCMS Curriculum Modification Examples

At CCMS, a similar approach is taken in the modification of curriculum units. CCMS researchers working on the development of the IQWST materials use a multi-stage process of enactment and revision (Krajcik et al., 2008). In the example that follows, we illustrate the modifications that were made between the first and second enactments of an IQWST chemistry unit dealing with chemical reactions (see Chapter 3). After the first enactment of the unit, it was observed that students did not make the same pretest to posttest gains on the chemical reactions learning goal that they had made on an earlier learning goal on substances and their properties. It was also observed that students did not gain as much as expected in their use of the claim-evidence-reasoning form of scientific explanation. For example, although students knew that a chemical reaction had occurred when new substances had different properties relative to

the old substances, their reasoning lacked precision and rarely used an underlying scientific principle to explain why the evidence supported their claims. For example, a typical student's reasoning stated, "This evidence supports that a chemical reaction occurred because you can follow the evidence and determine that it changed" (Harris et al., 2006).

The learning goals–driven design model described in Chapter 3 relies on unpacking learning goals into their constituent aspects so that instruction can support each one of them. Unpacking the learning goals also allows assessment to be more precise. The feedback obtained by the researchers revealed that important aspects of the targeted claim-evidence-reasoning science practice had not been made clear to teachers or to students. To address this issue, IQWST researchers revised the learning performances to state more explicitly what they wanted students to include in their scientific explanations. They also added a lesson in which teachers introduced the idea of scientific explanations to the students and provided a rationale for using that form of explanation (McNeill & Krajcik, 2008b). In addition, the researchers developed written materials to support students' learning of each of the components (McNeill et al., 2006).

IQWST researchers also found that students were not clear about the difference between chemical reactions and dissolving. When asked, "Which change will produce a new substance?" the majority of students responded that dissolving lemonade powder in water would produce a new substance. Another item presented a situation in which a clear liquid, a white powder, and a red powder were mixed together, and asked students to "describe three pieces of evidence you would look for to determine if a chemical reaction occurred." Here again students wrote that "powder dissolving" would be evidence for a chemical reaction. This idea was actually held by more students after instruction than before (67.4% vs. 55.1%).

The first step the IQWST researchers took in addressing this misconception was to clarify the learning goal by including an expectation that students would know both what did and what did not count as a chemical reaction (by providing both examples and counterexamples), especially the idea that phase change and dissolving were not chemical reactions. They also added a lesson to the unit specifically focused on creating mixtures and comparing the properties before and after to determine if a new substance had been made, and they provided opportunities for students to analyze particle models of chemical reactions, phase changes, and mixtures and to discuss the similarities and differences between these processes. The lessons the researchers learned about the importance of providing counterexamples were then applied more generally to the clarification of other content standards.

After the second enactment of the chemistry unit, the learning gains were considerably larger than after the first enactment, both for the chemistry content

and for the use of the claim-evidence-reasoning model of scientific explanation. Regarding the specific problem students had with dissolving as an example of a chemical reaction, in the second enactment the percentage of students who thought that dissolving lemonade was an example of a chemical reaction dropped from 47.4% to 39.0%. This was an improvement over the first enactment, and probably attributable in part to adding the lesson on mixtures, but it demonstrated that the concept continued to be difficult for students to grasp. Thus, in the next round of revision, this section of the curriculum materials will be revised again to further address the area of concern. During each cycle of revision, assessment results inform the researchers' understanding of the strengths and challenges of the materials and guide them in making the materials more successful in classrooms with teachers and students.

SUMMARY AND CONCLUSIONS

In this chapter we defined three types of learning outcomes that are related to the development of an integrated understanding of science. We also demonstrated how qualitative and quantitative approaches can be used to make judgments about item validity, and we gave examples of how CCMS and TELS use assessment in improving curriculum materials. We also showed the limitations of commonly used multiple-choice items in measuring high levels of knowledge integration and how those items can be improved.

Assessment plays an important role in promoting consistency throughout the science education system by offering clear expectations to all participants and by providing feedback on how well those expectations are being met. For assessment to serve these functions well, the design of assessment instruments, as well as the collection, analysis, and interpretation of assessment data, should be aligned with well-articulated and justifiable learning goals. Furthermore, assessment should be aligned with the same learning goals that the curriculum is organized to teach, so that the entire system can function together to achieve the same goal.

For assessment to contribute to the integrated understanding of science envisioned in this volume, assessment developers need to be aware of the limitations of many of the assessment items currently in use and prepared to construct more effective and informative items and tests. The following design principles emerge from a synthesis of the assessment work of the two centers.

Design Principles

First, assessment development should begin with a clear statement of what students are expected to know and be able to do. These expectations should be

written in terms of specific knowledge and skills students should have, but they should also be consistent with broader learning goals, such as linking multiple ideas together, connecting science content to science practices, or developing mental models of abstract ideas that will allow students to predict and explain events in the world.

Second, the validity of assessments should be examined both qualitatively and quantitatively to take advantage of the strengths of both approaches and to ensure that assessments are measuring what they are supposed to measure. Quantitative analyses are useful for demonstrating what each item adds to a set of items designed to measure a construct, the comparative difficulty of items, how well items discriminate among students along an ability continuum, and whether or not items function similarly for different subpopulations. However, although quantitative data can point to areas of potential concern, qualitative analysis is needed to pinpoint the nature of the problems that need to be resolved. Qualitative review of items is particularly useful for determining alignment of items with the learning goals, threats to validity due to poor item construction, and appropriateness of task contexts for students from different subgroups. Together, qualitative and quantitative analyses provide powerful tools for the development and evaluation of assessment items and instruments.

Finally, it is clear from the work of TELS and CCMS that data from high-quality assessments can be used to improve the quality of curriculum materials and classroom instruction by determining at various times during instruction what students are thinking and then using that information to modify instruction (Casperson & Linn, 2006; Krajcik et al., 2008).

Policy Implications

A number of policy implications emerge from the assessment work of TELS and CCMS. First, for assessment to be a positive force in science education, especially given the extent to which assessments are used as part of high-stakes accountability, learning goals at the local, state, and national levels need to present a coherent account of the natural world, and they need to be well integrated from grade band to grade band and across related topics at the same grade band. Otherwise assessment, and also instruction, will be based on fragmented and disconnected sets of ideas. The learning goals should form a logically consistent whole, and each idea should build one upon another. For these learning goals to be useful as guides to instruction and assessment, they should also be elaborated so that it is clear what the expectations are for students and what boundaries exist at each grade band. This is consistent with the recommendation of the Commission on Instructionally Supportive Assessment (2001): “A state’s high priority content standards must be clearly and thoroughly described so that the knowledge and

skills students need to demonstrate competence are evident.” These descriptions “should result in relatively brief, educator-friendly descriptions of each high priority standard’s meaning” (McColskey & McMunn, 2002, p. 5).

Second, school administrators at all levels, key personnel in funding agencies, education policy makers, and the general public should be aware of the limitations of the assessment instruments currently being used in standardized testing. Most people who make decisions about assessments and use results from the assessments may not be aware of how poorly existing assessments in science measure what students know. This is because most of the users of science assessments do not have the detailed knowledge of science content, the familiarity with research on student learning, or the expertise in item construction needed to make informed judgments about assessments in science.

Finally, given the current lack of high-quality assessment instruments in science, increased funding is necessary for facilitating the development and implementation of more cognitively sensitive and meaningful assessments that can assist teaching and learning at the classroom level and that can be used to conduct high-quality educational research. There is currently a lack of research tools to effectively measure learning outcomes, and without these tools, little progress can be made.