JOURNAL OF RESEARCH IN SCIENCE TEACHING VOL. 39, NO. 9, PP. 889–910 (2002)

Analysis of Students' Assessments in Middle School Curriculum Materials: Aiming Precisely at Benchmarks and Standards

Luli Stern,[1] Andrew Ahlgren[2]

[1]*Department of Education in Technology and Science, Technion–IIT, Technion City, Haifa 32000, Israel*

[2]*Project 2061, American Association for the Advancement of Science, 1333 H Street, NW, Washington, DC 20005*

Received 9 October 2000; Accepted 30 April 2002

Abstract: Assessment influences every level of the education system and is one of the most crucial catalysts for reform in science curriculum and instruction. Teachers, administrators, and others who choose, assemble, or develop assessments face the difficulty of judging whether tasks are truly aligned with national or state standards and whether they are effective in revealing what students actually know. Project 2061 of the American Association for the Advancement of Science has developed and field-tested a procedure for analyzing curriculum materials, including their assessments, in terms of how well they are likely to contribute to the attainment of benchmarks and standards. With respect to assessment in curriculum materials, this procedure evaluates whether this assessment has the potential to reveal whether students have attained specific ideas in benchmarks and standards and whether information gained from students' responses can be used to inform subsequent instruction. Using this procedure, Project 2061 had produced a database of analytical reports on nine widely used science middle school curriculum materials. The analysis of assessments included in these materials shows that whereas currently available materials devote significant sections in their instruction to ideas included in national standards documents, students are typically not assessed on these ideas. The analysis results described in the report point to strengths and limitations of these widely used assessments and identify a range of good and poor assessment tasks that can shed light on important characteristics of good assessment. © 2002 John Wiley & Sons, Inc. J Res Sci Teach 39: 889–910, 2002

Assessment of student performance exerts extraordinary influence on the lives of children and their families and on every level of the education system. If used properly, good assessment can be a powerful catalyst for improving both curriculum and instruction. Poor assessment practices, on the other hand, can impoverish our expectations for learning science, focusing teachers' and students' efforts on less important concepts and skills or on test taking as an end in itself.

Correspondence to: L. Stern; E-mail: lstern@tx.technion.ac.il
DOI 10.1002/tea.10050
Published online in Wiley InterScience (www.interscience.wiley.com).

Standards-based reform is founded on the premise that fundamental improvement of Grades Kindergarten through 12 (K–12) education in the United States begins with a coherent, well-articulated set of specific learning goals. Project 2061 of the American Association for the Advancement of Science (AAAS), through *Science for All Americans* (1989) and *Benchmarks for Science Literacy* (1993), and the National Research Council, through the *National Science Education Standards* (1996), have provided such a vision for reform of education in science. In general terms, these documents emphasize understanding ideas and skills rather than memorizing vocabulary and procedures. To achieve such meaningful understanding, the science curriculum does not need to include more and more content, but to focus on a relatively small number of important ideas in order to teach them better. The convergence of the two sets of recommendations—*Benchmarks for Science Literacy* and the *National Science Education Standards*—demonstrates a substantial consensus on what all students should know and be able to do at specific K–12 grade levels and what topics can be postponed or excluded so that there is time to teach core topics well. Although many states have developed their own standards or frameworks to guide their reform efforts, most have used the national documents in crafting their frameworks. Many of them organize their own recommendations according to topics in *Science for All Americans* and the *National Science Education Standards*, and some even adopt statements from these documents and AAAS' *Benchmarks* verbatim (Zucker, Young, & Luczak, 1996).

With a reasonably good consensus on what all students should know and be able to do in science, researchers and materials' developers have attempted to take the next step to link instruction and assessment to benchmarks and standards. For the past 4 years, with input from hundreds of K–12 teachers, teacher educators, materials developers, scientists, and cognitive researchers nationwide, AAAS' Project 2061 has developed and field-tested a rigorous procedure for evaluating science and mathematics curriculum materials, including their assessments (Kesidou & Roseman, 2002; Roseman, Kesidou, & Stern, 1996; Kulm, 1999; www.project2061.org). What makes this procedure unique is that it examines the content and instructional strategies used in curriculum materials for how well they appear to support students achieving specific learning goals, such as the AAAS' benchmarks and the National Research Council's science standards. Rather than matching the curriculum to general topic headings that students are supposed to learn about—such as ''cells'' or ''structure of matter'' (under which it is possible to align almost any curriculum material)—this analysis tool matches ideas in the curriculum to fine-grained learning goals. The criteria for making judgments about the instructional design of curriculum materials are derived from research on learning and teaching and refer to numerous aspects in the curricula (Kesidou & Roseman, 2002). For example, because the positive effects of providing students with a clear understanding of the purpose of an activity have been repeatedly confirmed by research (Wise & Okey, 1983; Hart, Mulhall, Berry, Loughran, & Gunstone, 2000), materials are examined for whether they attempt to make their purpose clear and meaningful to students. Fostering understanding in students also requires taking time to attend to the ideas they already have, both ideas that are incorrect and ideas that can serve as a foundation for subsequent learning (Eaton, Anderson, & Smith, 1984; McDermott, 1991; Osborne & Freyberg, 1985). In addition, students need to have a sense of the range of phenomena that science can explain through hands-on activities, demonstrations, audiovisual aids, or discussions of familiar phenomena (Anderson & Smith, 1987). Students then need opportunities to apply the scientific ideas in a variety of contexts to describe and explain phenomena (including their own experiences), solve practical problems, or consider alternative positions on issues (Ericsson, Krampe, & Tesche-Romer, 1993).

Teams of educators and scientists trained in the use of the Project 2061 procedure have examined how well curriculum materials treat mathematics and science benchmarks and standards. In science, their analysis focused on middle school materials and on some fundamental

ideas in physical science, life science, and earth science (Kesidou & Roseman, 2002; Stern & Roseman, 2000; Caldwell & Stern, 2000). Detailed reports on the treatment of these ideas are now available on the Project 2061 Website (www.project2061.org) and contain information intended to help educators make better decisions about which middle grade textbooks can support their students in learning important science ideas and what needs to be done to improve them.

This article focuses on one of the instructional features of the commercially available middle school curricula analyzed: namely, the assessments found in these materials. These assessments are widely used and thus play an essential role in the classroom practice. The underlying premise behind Project 2061's criteria to evaluate assessment is that assessment should be used mainly to promote student learning and improve instruction. For this reason, classroom assessment should focus on important ideas and skills and should provide opportunities for students to apply the ideas they learned. Curriculum materials need to provide guidance for teachers on how to interpret and act on student responses. Related to these characteristics, judgments about assessments in curriculum materials are made on the basis of (a) whether assessment questions and tasks appear to aim at specific benchmarks and standards, (b) whether these questions and tasks are likely to reveal what students actually know about the content specified in benchmarks and standards (as opposed to rote memorization of these goals), and (c) whether assessment embedded in curriculum materials throughout instruction can be used for making modifications in instruction. This report describes common types of assessment tasks found in the comprehensive middle school curriculum materials analyzed and points to major strengths and weaknesses of these assessments. The objectives of this study are to (a) report on the alignment of assessments in widely used curriculum materials to national standards,[1] and (b) identify examples of tasks that can illustrate characteristics of good and poor assessments included in currently available materials.

Design and Procedures

Analysis of Content and Instructional Quality

The evaluation procedure takes account of both the content and instructional design of the materials. The procedure carefully examines how well a material's content aligns with each benchmark (or standard) and how well the instructional strategies in the student text and the teacher guide can support students' learning of this content. Criteria for judging instructional quality examine many aspects of instruction, such as whether the material conveys a sense of purpose, whether it takes account of students' prior ideas, and whether it includes relevant phenomena that make the scientific ideas plausible [for more details about the procedure and the specific criteria, see Roseman et al. (1996), Kulm (1999), Kesidou and Roseman (2002), *Resources for Science Literacy: Curriculum Materials Evaluation* (in preparation), and the Project 2061 web site at www.project2061.org].

Selection of Key Science Ideas

The time and rigor required for evaluating curriculum materials using the Project 2061's procedure make it impractical for use on every topic included in a yearlong or multiyear curricula. Therefore, ideas from three important topics—the kinetic molecular theory (physical science), flow of matter and energy in ecosystems (life science), and processes that shape the earth (earth science)—were used as the basis for the analysis. These ideas are examples of the core science content likely to appear in any middle grade material and are common to benchmarks, national standards, and most state frameworks. Furthermore, the topics selected are the basis for learning other, more complex ideas in middle school as well as in high school. Yet research on student

learning indicates that students have difficulties learning these topics (Bell & Brook, 1984; Brook, Briggs, & Driver, 1984; Freyberg, 1985; Nussbaum, 1985; Smith & Anderson, 1986; Roth & Anderson, 1987; Johnston & Driver, 1989; Lee, Eichinger, Anderson, Berkheimer, & Blakeslee, 1993; Anderson, Sheldon, & Dubay, 1990). [Summaries of the research findings can be found in *Benchmarks for Science Literacy* (1993) and Driver, Squires, Rushworth, & Wood-Robinson (1994).] The specific ideas that relate to the kinetic molecular theory are included below; Appendix A includes the ideas in life and earth science.

Design of Analysis Process

Nine comprehensive middle school materials, both newly developed and widely used, were examined for their treatment of the selected ideas in physical, life, and earth science (the complete list of the materials appears in Table 1). Only comprehensive middle school science programs—that typically cover 3 years of instruction—were selected. These materials are now being widely used, or considered for use, in school districts and states.

The analysts had the appropriate expertise in physics, biology, and earth science and included experienced classroom teachers and university faculty. In each of these subjects, two 2-member teams independently analyzed the instruction and assessment in each curriculum material. To prepare for the evaluation, analysts had been extensively trained in the use of the Project 2061's evaluation procedure: Initially, they attended a 5-day workshop discussing and applying the evaluation criteria to a variety of examples. Working in independent teams, they then completed the analysis of their first curriculum material (off-site), and then met again to discuss and reconcile discrepancies in the reports, while getting feedback from Project 2061's staff. For each criterion,

Table 1
Assessment scores in middle school science materials (Topic: kinetic molecular theory)

Textbook Series (in Alphabetical Order)	Criterion 1: Aligning to Goals	Criterion 2: Testing for Understanding	Criterion 3: Informing Instruction
Glencoe: Physical Science (Glencoe/McGraw-Hill, 1997)	■	■	■
Macmillan/McGraw-Hill Science Series (Macmillan/ McGraw-Hill, 1995)	■	■	■
Middle School Science & Technology (Kendall/Hunt Publishing, 1999)	■	■	■
Prentice Hall Exploring Science (Prentice Hall, 1997)	■	■	■
PRIME Science (Kendall/Hunt, 1997)	■	■	■
Science 2000 (Decision Development Corporation, 1995)	■	■	■
Science Insights (Addison Wesley, 1997)	■	■	■
Science Interactions (Glencoe/McGraw-Hill, 1997)	▬	▬	■
Science Plus: Technology & Society (Holt Rinehart and Winston, 1997)	▬	□	■

■ = poor (0–1); ▬ = fair (1.5); □ = Satisfactory (2); ▬ = very good (2.5).

reviewers cited relevant evidence from the student text or the teacher guide, and scored the material in each of the three topics examined as ''excellent,'' ''very good,'' ''satisfactory,'' ''fair,'' or ''poor.'' The specific evidence analysts used to justify their judgments—such as text segments, activities, teacher notes, or assessment questions—served as the basis for the reconciliation process. After each team completed the analysis of subsequent materials, reconciliation between teams was coordinated by Project 2061.

The consistency of inter analyst scores was studied before this evaluation. In brief, when seven highly trained 2-member teams independently evaluated a middle school science material for its treatment of ideas related to the topic ''flow of matter and energy in ecosystems,'' 87% of their scores were in agreement (Kesidou & Roseman, 2000). Similar consistency was obtained when interanalyst reliability was tested on middle school mathematics materials (Kulm & Grier, 1998).

Assessment Criteria

The assessment tasks, including both end-of-instruction assessment and assessment included throughout instruction, were evaluated using the criteria below. Developers' claims were carefully checked to ensure the inclusion not only of tasks that were explicitly labeled as assessment, but also of any task that developers suggested teachers might use as such. By ''assessment task'' we mean a variety of item types ranging from true/false and multiple choice questions up through essays and hands-on performance.

The criteria to analyze assessment are explained below and illustrated in the findings section. They address three important aspects of assessment: Clearly, high-quality assessment tasks should be aimed at assessing important ideas and skills (Treagust, 1988; Shavelson, Carey, & Webb, 1990; Gallagher, 1996), so the materials are first examined for their inclusion of sufficient goal-relevant assessments. Alignment of an assessment task to a set of learning goals is often judged by topic coverage—that is, whether the assessment task fits into one of the goal topics. Topic coverage, however, does not guarantee that the task targets effectively specific learning goals. Assessment tasks might fall under the same topic heading (e.g., cells) or the same subtopic heading (e.g., cell replication) and still miss the point. Judgments about alignment are therefore made on the basis of whether assessment tasks aim at specific ideas in benchmarks and standards (see Criterion 1: Aligning to Goals, below). Second, assessment tasks in curriculum materials that appear to aim at specific benchmarks or standards are examined for whether they merely require rote memorization or really attempt to probe students' understanding. Rather than memorizing bits of information, science-literate adults should be able to use knowledge to describe, explain, and predict real-world phenomena, consider alternative positions on issues, and even solve practical problems. Meaningful assessment tasks should reflect these expectations (White & Gunstone, 1992). In the ideal, some of the assessment should consist of familiar tasks to judge student comprehension of what was taught; other assessment items should pose novel tasks to judge transfer of what has been learned (see Criterion 2: Testing for Understanding, below). Finally, including quality tasks and collecting student responses have little point unless the information can be acted upon (Black, 1998; Black & William, 1998; Treagust, Jacobowitz, Gallagher, & Parker, 2001; Bell & Cowie, 2001). Assessments should be used not only as instruments for grading students at the end of a unit or a course, but also as diagnostic instruments that help determine learners' needs. Materials are examined for whether they provide assessment tasks along the way to gauge student progress and whether these tasks can be used to diagnose students' remaining difficulties and inform instruction accordingly (see Criterion 3: Informing Instruction, below).

In summary, the following three criteria and indicators were used to examine and judge how well the materials assess science literacy ideas.

CRITERION 1. ALIGNING TO GOALS. Assuming a content match between the curriculum material and the benchmark, are assessment items included that match the same benchmark?[2]

Indicators of meeting the criterion.

1. The specific ideas in the benchmark are **necessary** to respond to the assessment items.
2. The specific ideas in the benchmark are **sufficient** to respond to the assessment items (or, if other ideas are needed, they are not more sophisticated than 6–8 benchmarks and have been taught earlier).

CRITERION 2. TESTING FOR UNDERSTANDING. Does the material assess understanding of benchmark ideas and avoid allowing students a trivial way out, such as repeating a memorized phrase from the text without understanding?

Indicators of meeting the criterion.

1. Assessment items focus on **understanding** (as opposed to recall) of benchmark ideas.
2. Suggested assessment include both **familiar and novel** tasks.

CRITERION 3. INFORMING INSTRUCTION. Are some assessments embedded in the curriculum along the way, with advice to teachers as to how they might use the results to choose or modify activities?

Indicators of meeting the criterion.

1. The material uses embedded assessment as a **routine strategy** (rather than just including occasional questions).
2. The material assists teachers in **interpreting** student responses (e.g., by providing annotated samples of student work pointing to typical difficulties).
3. The material provides **specific suggestions** to teachers about how to use the information from the embedded assessments **to make instructional decisions** about what ideas need to be addressed by further activities.

Scoring schemes for Criteria 1 and 2 were based on the number of assessment tasks that met the indicators. Although it is undoubtedly desirable to use more than one assessment task to probe the understanding of any idea—especially when assessing knowledge of broad generalizations—the exact number of tasks that can be considered adequate is empirical. Still, to judge adequacy of assessment tasks, reviewers considered whether all benchmark ideas are assessed and how many tasks are included for each. The Findings section below illustrates how reviewers arrived at particular judgments. Curriculum materials were not penalized for including assessment tasks that do not meet the criteria. Rather, they could get credit only for the inclusion of tasks that do. For Criterion 3, scoring scheme was based on the number of indicators met by the material. Project 2061's staff compared and reconciled the scores for all curriculum materials. The results of this assessment analysis are discussed below.

Findings

This section is divided into three parts: The first presents the assessment scores of the curriculum materials examined, the second illustrates the characteristics of quality tasks, and the

third includes typical examples of assessment tasks from curriculum materials. Most examples are for the assessment of the kinetic molecular theory ideas included below.

Physical Science Ideas: Kinetic Molecular Theory

a. All matter is made up of **particles** called atoms and molecules (as opposed to being continuous or just *including* particles.)
b. These particles are extremely **small**—far too small to see directly through a microscope.
c. Atoms and molecules are perpetually **in motion**.
d. Increased **temperature** means greater molecular motion, so most materials expand when heated.
e. Differences in **arrangement and motion** of atoms/molecules in solids, liquids, and gases:

   In solids, particles (i) are closely packed, (ii) are [often] regularly arranged, (iii) vibrate in all directions, (iv) attract and ''stick to'' one another

   In liquids, particles (i) are closely packed, (ii) are not arranged regularly, (iii) can slide past one another, (iv) attract and are loossly connected to one another.

   In gases, particles (i) are far apart, (ii) are randomly arranged, (iii) spread through the spaces they occupy, (iv) move in all directions, (v) are free of one another, except during collisions

f. Explanation of **changes of state**—melting, freezing, evaporation, condensation—and perhaps dissolving in terms of changes in arrangement, interaction, and motion of atoms/molecules.

Part 1: Assessment Scores in Middle School Materials

Clearly, one cannot expect materials to assess ideas they were not intended to teach. Therefore, analysis of assessment in a curriculum material was carried out only if the material's content aligns with the benchmarks' ideas examined in this study. Detailed analyses of the content in the curriculum materials indicated that, with one exception, nearly all the specific ideas are addressed in each of the curriculum materials (Kesidou & Roseman, 2002; www.project2061.org). Moreover, the specific benchmarks' ideas are often explicitly listed as the learning objectives in the materials and significant sections are devoted to teaching these ideas.

Whereas the materials do not differ greatly in their inclusion of content related to the specific ideas in physical, life, and earth science, scores for the quality of instructional support of the physical science ideas were typically slightly higher than the scores for life and earth science ideas. The detailed findings suggest that the physical science ideas that served as the basis of this study (kinetic molecular theory) were the focus of instruction more than the key ideas that served as the basis of analysis for the life and earth science topics (Kesidou & Roseman, 2002). Because the materials treat best the physical science ideas, the analysis scores of the assessment aimed at these ideas will be presented and discussed first, and then compared with findings related to the assessments aimed at the life and earth science ideas.

Table 1 shows how well each of the curriculum materials examined scored on the assessments for the topic ''kinetic molecular theory'' (technical reports including the assessment scores for the topic ''flow of matter and energy in ecosystems'' and the topic ''processes that shape the earth''

can be found at www.project2061.org). However, as is evident, most materials examined did not receive high scores on the embedded and end-of-instruction assessments they include.

End-of-Instruction Assessment Scores in Physical Science. In most materials analyzed, the tests do not assess students on the core ideas; and if they do, they often do not require application of ideas, relying instead on students' memorization of statements in the text. On the criteria ''aligning to goals'' and ''testing for understanding,'' which examine the quality of end-of-instruction assessment, only two materials (*SciencePlus* and *Science Interactions*) received a score higher than ''poor,'' and only one material (*SciencePlus*) received a ''very good'' score. Considering that all the materials examined state the kinetic molecular theory in their learning objectives and include considerable amounts of relevant content—both text and activities—these scores are surprisingly poor. In the absence of empirical data, judgments about the number of assessment tasks that can be considered sufficient may seem somewhat arbitrary. However, our assumption was that multiple probes are required to assess any idea, let alone broad generalizations such as the ones examined in this study. Judgments were therefore based on whether all the kinetic molecular theory ideas are assessed at least once and how many tasks are included for each. In all the curriculum materials examined, the numbers of tasks provided are inconsistent across the set of key physical science ideas. Even the top-rated material (*SciencePlus*) provides many tasks that focus on particular ideas (e.g., the idea that increased temperature means greater molecular motion), a few for other ideas (e.g., the idea that particles are perpetually in motion), and none for another idea in the set (the idea that particles are extremely small). Moreover, in many of the materials most of the key physical science ideas are barely assessed. To illustrate the differences between the materials, consider the assessment tasks *SciencePlus* includes for the idea that all matter is made up of particles (Idea a): Students are asked to explain why many objects with the same volume have different masses (p. 138), describe the evidence showing that matter is made up of particles (p. 138), name three observations that they have made while studying the unit that support the particle model of matter (Assessment, p. 68), and apply the particle model of matter to explain the following situation. ''In five trials, 50 g of copper is allowed to react with oxygen to form copper oxide. Each time, the copper reacts with the same amount of oxygen'' (Assessment, p. 55). In addition, students are asked to list ''everyday observations'' that can be explained by the particle model of matter (p. 138) and explain some observations in a story they read using the particle model (p. 139). In contrast, only a single task that targets this important idea is included in some materials (*Science Insights*; *Science Interactions*; *Glencoe Physical Science*; *Middle School Science & Technology*; *PRIME Science*) and none in the rest (*Macmillan/McGraw-Hill Science Series*; *Prentice Hall Science*; *Science 2000*). The detailed technical reports (www.project2061.org) include evidence-based arguments for all scores.

A poor score may represent different findings in different materials: National Science Foundation (NSF)-funded materials (*PRIME Science*, *Science 2000*, and *Middle School Science & Technology*) typically provide a small number of aligned assessment tasks that tend to be of high quality—that is, require application of the core ideas (and hence meet Criterion 2). For example, relevant to the idea that matter is made up of particles, *PRIME Science* includes a task that asks students to explain (in terms of small particles) how milk mixes with coffee (Level C, p. 153) and *Middle School Science & Technology* includes a task that asks students to describe the evidence scientists have for the particle model (Level B, p. 241). In contrast, most other materials may provide more items that seem related to the kinetic molecular theory, but most of these items are either not at the same level of sophistication as the ideas examined or of poor quality (that is, either unclear or relying on rote memorization). For example, relevant to the idea above, the following tasks are included:

- According to the particle model of matter, all matter is: a. too small to see; b. made up of tiny particles that are in constant motion; c. made up of one type of particle; d. made up of particles that are the same size (*Science Insights*, p. 25, Test 6A).
- The idea that matter is composed of small——called atoms is explained by the atomic theory of matter (*Science Interactions*, Review and Assessment, Course 2, p. 83).
- The particles that make up all matter are called——(*Glencoe: Physical Science*, Chapter 9 Test, p. 59).

In both cases the kinetic molecular theory ideas are not adequately assessed and therefore a poor score was given.

Assessment throughout Instruction (Embedded Assessment) (Criterion 3). Beside grading students, assessment should have other purposes, such as to diagnose students' remaining difficulties to inform and direct subsequent teaching (Black, 1998; Treagust et al., 2001). In evaluating whether materials include assessment throughout instruction and guidance for teachers on how to use students' responses to modify instruction, all materials received poor scores. Many materials include a discussion of the importance of embedded assessment when they describe their pedagogical features, include questions labeled as assessment throughout chapters or units, and may even encourage teachers to use certain assessment components in the text to inform instruction.

Unfortunately, most of the embedded assessment tasks are not aligned with the kinetic molecular theory ideas examined here, and those that are often focus on terms and definitions or can be answered successfully by simply copying answers from the text. Other questions included are not particularly helpful for monitoring students' progress. For example, after introducing the idea that matter is made up of tiny particles, *SciencePlus* has students construct models of molecules and hang them on mobiles (Level Blue, p. 98). Students having difficulty with this key idea can successfully construct the required molecular models while still believing that molecules are in materials, as opposed to believing that molecules constitute materials.

Occasionally, some tasks are included during instruction that may be used to enlighten instruction. For example, students are asked to explain (at the molecular level) how a gasoline can filled to the brim on a cool morning leaks in the afternoon (*Science Interactions*, Course 3, p. 221), why liquids and gases take the shape of their containers (*Macmillan/McGraw-Hill*, Changes in Matter, p. 25S), how a coin placed on the rim of a bottle moves when the bottle is placed in hot water (but not when placed in cold water) (*Middle School Science & Technology*, Level B, p. 210), or why copper shrinks when it cools (*Glencoe: Physical Science*, p. 220). However, even when such questions are included, materials fail to provide guidance to teachers on how to diagnose what students' remaining difficulties are or offer ways to change instruction accordingly. Considering the many difficulties and misconceptions students have with this topic, this lack of interpretive help is a serious flaw.

Scores in Life and Earth Science. Assessment scores of life and earth science topics are almost uniformly poor.

With respect to Criterion 3 (informing instruction), all materials received poor scores in both life and earth science. Whereas in physical science good questions are occasionally included during instruction, in life and earth science quality questions that focus on benchmarks' ideas are scarce. As noted above, even in physical science questions are not used for making instructional decisions.

With respect to end-of-instruction assessment (Criteria 1 and 2), all but one material (*SciencePlus*) received poor scores. *SciencePlus*, the top-rated material, which received very good scores in physical science, received satisfactory scores in earth science and fair scores in life science. This material includes some quality tasks aimed at key earth science ideas. For example, relevant to the idea that the earth's surface is continually changing (Idea a), students consider the changes that may have occurred while Rip van Winkle slept for 20 years (Level Green, p. 492) and list evidence to convince Jerome (who believes that the earth was always the same as it appears today) that the earth has changed over time (Level Red, Assessment, p. 274). However, even in *SciencePlus* the number of tasks that assess each key earth science idea varies, and some key ideas are not assessed at all. In most materials, the majority of relevant tasks target the idea that several processes contribute to building up and wearing down the earth's surface (Idea b); however, these tasks typically target the details of individual earth-shaping processes and students hardly consider how several processes—that constantly work at different rates and locations—cause the earth's surface to look the way it does. For example, students explain the difference between weathering and erosion (*PRIME Science*, Level B, p. 103); contrast weathering, erosion, and deposition (*Prentice Hall*, Teaching Resources, p. 69); identify and explain ways in which rocks are changed into tiny particles (*Macmillan*, Earth's Solid Crust, p. 12), or explain how sand dunes migrate (*Science Interactions*, Course 1, p. 92).

In life science, most materials do not even include questions that target the key life science ideas examined here. Most relevant tasks that are included in *SciencePlus* focus on reactants and products of photosynthesis and respiration but not on the big ideas of how matter and energy are transformed in living things.

Part 2: Characteristics of Good Assessment Tasks

According to the standards-reform movement, science literate adults should know the ideas specified in the standards documents and draw upon them in a variety of contexts. Therefore, students should be assessed on their ability to apply—as opposed to memorize—general propositions. The following list, assembled from examples found in curriculum materials and the cognitive research literature, together with a few examples suggested by the authors, attempts to illustrate possible kinds of tasks that require students to apply benchmarks ideas. These examples do not dictate an assessment format. They could be set up as open-ended or selected response tasks (such as multiple choice), they could be administered as written or oral assignments, and they could be used to track individual or group progress. Tasks that probe understanding and require application of science ideas may include but are not restricted to tasks that require students to[3]:

- Rephrase general propositions in one's own words.

For example,

> Could you explain, in your own words, what is meant by the "particulate" nature of matter?

- Decide whether naive explanations of phenomena are correct and explain why.

For example, relevant to the idea that increased temperature means greater molecular motion, the following task anticipates students' naive attribution of observable properties to the invisible molecules.

> My friend says that when water freezes, the molecules get cold and turn hard. Do you agree? Explain. (Berkheimer, Anderson, Lee, & Blakeslee, 1988)

Relevant to the idea that the seemingly solid earth's surface is continually changing, this task challenges students to think about the commonly held idea that the earth is still.

> After a 20-year absence, Rip returns to his hometown. He is dumbfounded. "Why, it's just as I left it. It hasn't changed a bit. The old sycamore tree we used to have the swing on is still there by the old stone church. The creek is still crystal clear. It hasn't changed.. . ." Were things truly just as Rip left them, or should he take a closer look? What would Rip see if he looked more closely? (*SciencePlus*, Level Green, p. 492)

- Explain phenomena.

For example, relevant to the idea that increased temperature means increased molecular motion:

> Explain [in terms of molecules] why you can smell apple pie just taken out of the oven but can't smell apple pie just taken out of the refrigerator. (*SciencePlus*, Assessment, p. 58)

> Explain why a piece of candy dissolves faster in hot water than in cold water. Talk about substances and molecules (Berkheimer et al., 1988, Test 2)

Or, relevant to the idea that plants assemble some of the sugars they have synthesized from carbon dioxide and water into the plants' body structures:

> A maple tree weighs so much more than a maple seed. Where does this added mass come from? (Schneps & Sadler, 1988)

- Predict new phenomena.

For example, relevant to the idea that increased temperature means greater molecular motion:

> What would happen to a solid chunk of steel that sits outside on a very hot summer day? Explain in terms of particles. (Modified from Berkheimer et al., 1988, Test 2)

Relevant to the idea that matter is made up of particles:

> Consider a piece of copper wire. Divide it into two equal parts. Divide one half into two equal parts. Continue dividing in the same way. Will this process come to an end? Explain your answer. (Stavy & Tirosh, 1993)

Relevant to the idea that plants assemble some of the sugars they have synthesized into the plants' body structures:

> Assuming some animal eats carrots' leaves (but not carrots), do you think it will affect the carrots' size this year? Explain your answer. (Project 2061, unpublished data)

- Decide whether certain phenomena are instances of a generalization (or identify phenomena that could be explained by the generalization).

For example:

> Which of the following are explained by the idea that the solid crust of the Earth consists of separate plates that move constantly? (Circle all possibilities)
>
> a. The Atlantic Ocean is getting wider each year
> b. The center of the Great Rift Valley in Africa is spreading
> c. Sand dunes in African deserts migrate
> d. Most volcanoes are found in certain locations

- Derive the generalization based on relevant instances.

For example, with respect to the idea that particles are in constant motion:

> Could you explain all the following observations using **one** principle? Show how.
>
> a. Maria placed a drop of food coloring in water and watched it spreading out.
> b. Maria smelled the soup that was cooking on the stove.
> c. Maria's mother put the garlic in a sealed jar so that it won't smell.

- Describe the evidence scientists have for a particular idea, consider whether different observations support this idea, or explain how certain pieces of evidence support a theory[4].

For example:

> What is some evidence that supports the theory of continental drift? (*SciencePlus*, Level Green, p. 522)

- Consider the appropriateness of a representation for an idea or compare a representation with the real thing.

For example, relevant to the idea that plants make their own food:

> Some people say that plants are like food factories. Do you agree? Why or why not?

- Represent the benchmark's idea.

For example, relevant to the idea that matter is made up of particles:

> Imagine that you could see everything magnified by many fold, until you could see molecules. Draw what you could see in this flask before and after half of the air in it is removed. (Modified from Novick & Nussbaum, 1978)

- Consider what would happen if the generalization is violated or modified.

For example, related to the idea that several processes contribute to building up and wearing down the earth's surface:

What would the earth's surface be like if erosion ceases?

What would the Earth's surface be like if there were 20 smaller crustal plates instead of the existing major ones? (*SciencePlus*, Level Red, p. 359).

Part 3: Typical Assessment Tasks in Middle School Curriculum Materials

In analyzing assessment tasks in curriculum materials in light of the key ideas, we have found assessment tasks that confound attempts to determine what students do or do not understand about benchmarks and standards ideas. Among the examples are:

- Many tasks that can be answered successfully by general intelligence alone or some ''test wiseness.'' (Successful responses could result without knowing the benchmark.)

For example, in the following task,

> Which of the following does *not* happen as the temperature of a gas increases?
>
> a. Molecules move faster.
> b. Molecules have more collisions.
> c. Kinetic energy increases.
> d. Kinetic energy decreases. (*Macmillan/McGraw-Hill*, Using Energy, p. 17)

Given that Options c and d are relevant opposite statements, one must be the correct answer (because kinetic energy probably does not increase *and* decrease as a result of increasing the temperature), students will be able to eliminate the first two options based on this simple reasoning, so that they will have a 50% chance to get the right answer without knowing anything about the kinetic molecular theory.

In other assessment tasks that we found, students were asked to unscramble relevant terms (e.g., *qateekuhar* for *earthquake*), solve crossword puzzles when definitions are given (in which often relevant terms can be deciphered simply based on letters of irrelevant terms solved in the puzzle), or crack the code of terms written in a ''different language.'' Many of these tasks can be answered successfully just by general intelligence, without the scientific ideas.

- Tasks that require knowledge and skills of experimental design but not of the ideas examined in this study. (Successful responses do not require knowledge of the benchmark idea).

For example, related to the topic ''flow of matter and energy in ecosystems'':

> Noa wants to know if plants grow better if they get more light. Describe an experiment she could set up to answer that question. Be sure to include what information, data, and measurements she should collect. You may draw a picture to help explain your answer. (Modified from *PRIME Science*, Level C, p. 5)

Students' answers might demonstrate various experimental skills, such as designing experiments and explaining procedures. However, this task does not assess students on the idea chosen for this study—that plants use light energy to make their food—because they can respond successfully without knowing this idea. (Although this task may be a good way to assess students' abilities to do some aspects of scientific inquiry, it was not counted in the number of aligned tasks that was used for scoring.)

- Many tasks that can be answered successfully using ideas less sophisticated than those in benchmarks or standards. Again, successful responses could result without knowing the benchmark.

For example, *Benchmarks for Science Literacy* expects middle school students to know that the particles in solids are closely packed and attract (or ''stick to'') one another (Idea e). In the example below, a student familiar with the macroscopic phenomena only will be able to respond successfully.

> Solids have
>
> a. a definite shape but no definite volume
> b. a definite shape and a definite volume
> c. a definite volume but no definite shape
> d. neither a definite shape nor a definite volume (*Prentice Hall*, Chapter 3 test, p. 37)

It is sometimes difficult to determine which idea or skill the developers intended to assess without information about the expected responses and the credit to be assigned to them. Still, students may legitimately respond using other ideas. For example, in the following example, students do not need to know anything about increased molecular motion to answer correctly; it would be enough to recognize the phenomenon that substances expand when they are heated:

> (Question): A metal door will work easily during the winter but stick during the summer. Explain why.
>
> (Expected response): During the summer, high temperatures will cause the metal molecules in the door to move faster, pushing each other further apart, and therefore the door will expand. (*Science Interactions*, Course 3, p. 213)

The expected response provides molecular explanation, and therefore this task was judged by the analysts to be ''aligned.'' However, because the task does not specify that students talk about molecules in their answers, they can give a reasonably correct (but much simpler) answer: ''The metal door expands in the summer when it's hot.''

Some tasks that require ideas less sophisticated than those in benchmarks might be easily fixed. For example, the following task:

> Would you expect a solid iron ball heated on the stove to:
>
> a. be a little smaller than before
> b. be a little larger than before
> c. stay exactly the same size as before
> d. I don't know

does not require knowledge on molecular motion; however, to assess students' understanding at the level of sophistication intended by benchmarks, students could be asked to explain their answers in molecular terms. Alternatively, the choices in the original question could be modified to reflect the molecular level and the cognitive literature:

> When you heat a solid iron ball on the stove:
>
> a. the number of molecules increases
> b. molecules expand or get larger
> c. molecules stay the same size but move farther apart
> d. molecules contract or get smaller

''Fixability'' of assessment tasks was not evaluated in this study, and therefore all tasks that do not require the specific ideas examined were not considered aligned, regardless of whether they could be fixed.

- Open ended tasks whose alignment with benchmarks and standards depends on what students' particular responses are. (Different acceptable responses may nonetheless demonstrate knowledge of quite different ideas.)

For example, consider the following task:

> Imagine that you are a water molecule who is going through a series of phase changes. Describe your experiences as you change from a molecule of ice to a molecule of liquid water. (*Prentice Hall*, Chapter 3, Teaching Resources, p. 20–22)

Several different kinds of response might be satisfactory. Students may show their understanding of the perpetual motion of particles, the effect of temperature on molecular motion, phase changes, water anomaly, or other ideas. Although students would likely use one or more of the ideas chosen for this study (and therefore the task is considered to be aligned), the alignment to any *specific* idea is not guaranteed.

- Tasks for which ideas in benchmarks and standards are necessary but not sufficient, for an adequate response. (Unsuccessful responses could therefore result from not knowing the benchmark or from not knowing something else.)

For example, consider the task:

> Microwave oven works by heating water molecules inside of a substance. Use what you have learned about the particle model of matter to answer the following questions:
>
> Why do potatoes sometimes explode in microwave ovens?
>
> What can a cook do to avoid having a potato explode in a microwave oven?
>
> Why would this work?
>
> [Answer provided in the Answer Key: As the water molecules inside a potato are heated, they move faster and eventually become water vapor (a gas). Because the gas expands quickly, it can cause the potato to explode. By poking a few holes in the potato before heating, a cook can provide an opening for the water vapor to escape through, preventing the gas from building up inside the potato. (Item found in an older edition of Prentice Hall)]

To respond successfully, students must know that substances expand when heated owing to increased molecular motion, an idea that is included in a middle school benchmark, but they must also know that cooks poke holes in potatoes to avoid their explosion (which goes beyond benchmarks). Because it would probably be unreasonable to expect students who had never seen someone poking a potato to come up with the correct answer, perhaps a better way to phrase the second part of the question would be, ''Some cooks poke a few holes in the potato before heating. How would this help to avoid having a potato explode in a microwave oven?''

- Relevant tasks that are not likely to be comprehensible to students. Unsuccessful responses could result from not knowing the benchmark or from lack of clarity about what is required. For example:

> The dots on the balloon represent particles of air. Use what you know about pressure, temperature, volume, and the kinetic theory of matter to write a hypothesis explaining what will happen to the volume of the balloon if pressure is kept constant and the temperature is lowered. (*Glencoe: Physical Science*, Assessment, p. 55)

Whereas this question targets an idea chosen for this study—that increased temperature means greater molecular motion so substances expand when heated—the diagram is likely to be confusing to students. It is not clear whether the air particles are inside or on the balloon. As a result, unsuccessful responses could result from not knowing the benchmark idea or from not understanding the task. Although this task may be easy to fix—for example, by not mentioning the pressure (that will likely introduce the confusion of additional variables) and referring clearly to tiny particles inside the balloon—it was hard to assume that, if used as is, the task would be readily understood.

- Many tasks for which science literacy ideas might be necessary and sufficient (and are therefore aligned) but require no more than rote memorization, rather than application of benchmark ideas—in ways such as those illustrated in Part 2 above.

For example, the task:

(i). The particles in a (*liquid*) are close together but are free to move around.
(ii). In (*solids*), the particles are closely locked in position and can only vibrate. (*Prentice Hall*, Chapter 3 Test)

However, as illustrated before, we did find a few good tasks that assess whether students can apply ideas from benchmarks and standards to describe and explain phenomena or consider alternative positions on issues.

Discussion

This review study finds assessments in curriculum materials severely lacking in value. End-of-unit tests and embedded questions in current middle school materials provide little or no help for finding out what students actually know about important science literacy ideas. With the exception of one material (*SciencePlus*) that received very good and satisfactory scores, all other materials received below satisfactory scores, and most scored poor on alignment, understanding, and informing instruction.

It may be argued that most curriculum materials reviewed in this study were developed before the publication of benchmarks and national standards; however, the topics that served as the basis of the analysis—kinetic molecular theory, flow of matter and energy in ecosystems, and processes

that shape the earth—are fundamental to the scientific disciplines and have been included in textbooks for many years. Eight of the nine curriculum materials that were examined in this study have attempted to target these three topics and included significant coverage in terms of text and activities for the key ideas.[5] In fact, the amount of textbook space devoted to these topics is enormous. In life and physical science, no other topic areas received more attention than the flow of matter and energy and the kinetic molecular theory. In earth science, sometimes as much as half of the textbook dealt with processes that shape the earth (Kesidou & Roseman, 2002; Caldwell & Stern, 2000). Thus, the overall poor scores on the criterion ''aligning to goals'' are unexpectedly low and it is hard to believe that better assessments are included in these materials for other, less treated topics.

In many of the materials that were analyzed, even when assessment tasks are aligned with core ideas, they often do not require application of ideas, relying instead on recall of definitions of terms and statements in the text. Some assessment tasks are incomprehensible and most tasks are not likely to reveal students' difficulties. Furthermore, despite claiming that they use assessment to inform decisions about instruction, curriculum developers do not provide sufficient questions throughout instruction to probe students' understanding, nor do they assist teachers in interpreting students' responses or in using these responses to change the instruction. The end-of-instruction assessments might be useful for teachers to grade their students but not to monitor what students actually know about core ideas as a feedback to instruction, whether for students currently enrolled in the class or for next years' students.

The fact that assessments of most textbooks received poor scores when examined by our procedure may raise some concern regarding the resolution of our scoring schemes; that is, that they do not discriminate between textbooks. For example, in physical science, both NSF-funded materials—that typically include a small number of relatively good tasks—and other materials—that often include tasks that require no more than rote memorization—received poor scores. As illustrated in the previous section, in both cases the benchmarks' ideas are inadequately assessed, and this justifies the similarity among the scores. The differences between the textbooks, however, indicate the way by which these assessments might be improved. Whereas improvement in some books would require increasing the number and variety of standards-based tasks, in other textbooks a more radical change would be required.

Although other studies have argued that classroom assessment often encourages superficial learning (Rudman, 1987; Black, 1998), this study has refined and characterized curriculum assessments by describing and illustrating shortcomings, and desirable qualifications, of standards-based assessments. The underlying premise behind the assessment criteria used in this study is that the primary purpose of assessment is to improve student learning (Neill, 1997; Black, 1998; Black & William, 1998). Admittedly, classroom assessment has other purposes, mainly the grading of students. However, there is ample evidence that the grading function is overemphasized (Rudman, 1987; Black, 1998). This point is also reflected and illustrated in our findings. Classroom assessment can and should indeed serve these two seemingly contradictory purposes. The widely used multiple choice tasks, for example, are easily scored and their results are useable immediately. Unfortunately, they are used in texts mainly for assessing students' rote memorization of science facts. If, however, they were created on the basis of students' ideas that had been documented during exhaustive clinical interviews or careful open-ended tasks, they could be used by teachers as powerful diagnostic tools to ascertain alternative conceptions in their classrooms (Sadler, 1998; Mintzes, Wandersee, & Novak, 2000). Our findings from the larger study, that contained the analysis of mathematics curriculum materials as well as a few research-based science modules, indicates that better assessment tasks are possible to design (Kulm, 1999; www.project2061.org).

The potential of assessment to improve teaching and learning has not been sufficiently explored and empirical evidence that documents the use of assessment to inform subsequent instruction is difficult to come by (Black, 1998). The difficulty behind fully exploiting this strategy might be explained by the unique pedagogical demands it requires from teachers and the fact that it is relatively time-consuming. However, this is a promising approach for monitoring progress and promoting learning (Black, 1998; Bell & Cowie, 2001). Therefore, it is hoped that the Project 2061's evaluation procedure and the publication of these results will stimulate empirical tests of students' learning. For example, could assessment tasks such as those included in Part 2 of the Findings section indeed be used to promote student learning? And what kind of remedies are effective to challenge students' remaining difficulties?

Given that the current assessment tasks in materials do not probe for understanding, high student scores on such tests would likely mislead teachers into thinking that their students understand more than they actually do (and might discourage teachers from seeking better instructional materials). Until teachers, parents, and scientists see the results of probing student understanding in depth, their reluctance to change is understandable. As the research shows us repeatedly, when understanding is probed in effective ways (such as those included in Part 2 of the previous section), students do not perform as well.

The results reported in this report are part of a larger study designed to characterize the instructional support provided in curriculum materials for the attainment of science literacy goals. Analysis of instructional strategies used in materials, such as taking account of students' ideas, engaging students with relevant and vivid phenomena, developing and using scientific ideas, and guiding students' interpretation and reasoning about phenomena and ideas, indicate that currently available textbooks do not support the attainment of national science standards (Kesidou & Roseman, 2002; Stern & Roseman, 2000; Caldwell & Stern, 2000; and relevant reports on project 2061 Web site). Most science materials include inadequate support for learning and inadequate ways to probe students' understanding. The full reports contain information that is intended not only to help educators improve decision choices, but also to show teachers what they should be asking for and developers what to provide.

Development of effective assessment tasks is similar to the development of effective instructional materials and requires iterative research and development cycles (including student interviews) in which students' responses are used as the basis for revision and refinement of assessment questions (Treagust, 1988; Gallagher, 1996). Development cycles of commercial curriculum materials in science typically involve a single field-testing phase (Kesidou & Roseman, 2002). Revisions are often made on the basis of teacher feedback and student responses to written questionnaires that are not necessarily effective in probing learning (Tyson, 1997). Given that the research and development efforts for developing instruction are fairly limited, it is not surprising that materials developers do not invest the costly efforts in developing their assessments.

In the ideal, curriculum and assessment should be aligned both with each other and with specific, worthwhile learning goals (such as those in *Benchmarks for Science Literacy* and the *National Science Education Standards*). However, in most curriculum materials it seems that the curriculum development drives the assessment development, and that assessment is designed to align to the actual content included in the material. This would explain why so many assessment tasks appear tailored to fit incidental details of the curriculum rather than important generalizations to be remembered. For instance, regarding the topic ''processes that shape the earth,'' students are asked many questions that require the knowledge of excruciating details related to the mechanisms of individual processes that shape the earth, but rarely are they asked about the impact of these processes on the continually changing surface of the earth.

Today, classroom teachers, administrators, and test developers who are required to choose, assemble, or develop assessments have little to guide them. As shown in this article, it is not trivial to decide whether assessment tasks are truly aligned with benchmarks and standards and effective in revealing what students actually know. Questions used in students' tests and interviews in high-quality research projects can serve as a source of good examples of assessment questions (White & Gunstone, 1992; Sadler, 1998; Black & William, 1998; Black, 1998; Mintzes et al., 2000). The assessment criteria used in this study have been helpful in identifying a range of examples of assessment tasks (both poor and good) that illustrate important aspects of assessment. The examples, in turn, have further characterized strengths and limitations of assessments. These findings have been sufficiently promising to warrant full-scale development of the criteria, procedures, and experience necessary to produce a reliable tool for analyzing and developing assessments—including state and standardized tests—and Project 2061 is expanding its current work on assessment to do so. Part of this work includes the production of case studies that illustrate the creation of standards-based assessments through research and development cycles. It is hoped that criteria and considerations that were used in this study will be useful to commercial developers and publishers of instructional and assessment materials, to districts and states that select and administer testing programs, and to classroom teachers who create or assemble their own tests.

The authors thank George (Pinky) Nelson and three anonymous reviewers for their helpful comments on earlier drafts of this article.

Notes

[1] *Benchmarks* and *standards* are used interchangeably in this article.

[2] Analysis of assessment in curriculum materials is carried out only if the material's content aligns with benchmarks and standards.

[3] Note that the possibilities listed are not necessarily applicable to all benchmarks ideas.

[4] These expectations are applicable only when the evidence is explicitly stated in *Benchmarks*.

[5] The ninth material (*Middle School Science & Technology*) includes extensive coverage for two of the three topics.

Appendix A: Ideas That Served as the Basis for Project 2061's Textbook Analysis

Ideas used for the evaluation were drawn from statements in *Benchmarks for Science Literacy and National Science Education Standards*.

Life Science Topic: Flow of Matter and Energy in Ecosystems

a. **Food** (for example, sugars) serves as **fuel and building material** for all organisms.
b. Plants make their own food, **whereas** animals obtain food by eating other organisms.

Matter is transformed in living systems (Ideas c1 – c4):

c1. Plants make sugars from carbon dioxide in the air and water.
c2. Plants break down the sugars they have synthesized back into simpler substance—carbon dioxide and water; assemble sugars into the plants' body structures (including some energy stores).
c3. Other organisms break down the stored sugars or the body structures of the plants they eat (or in the animals they eat) into simpler substances; reassemble them into their own body structures (including some energy stores).

c4. Decomposers transform dead organisms into simpler substances, which other organisms can reuse.

Energy is transformed in living systems (Ideas d1 – d3):

d1. Plants use the energy from light to make ''energy-rich'' sugars.
d2. Plants get energy by breaking down the sugars, releasing some of the energy as heat.
d3. Other organisms get energy to grow and function by breaking down the consumed body structures to sugars and then breaking down the sugars, releasing some of the energy into the environment as heat.

e. **Matter and energy are transferred** from one organism to another repeatedly and between organisms and their physical environment.

Earth Science Topic: Processes That Shape the Earth

a. The (seemingly solid) earth is **continually changing** (not only has it changed in the past but it is still changing).
b. **Several processes** contribute to building up and wearing down the earth's surface.
c. The processes that shape the earth today are **similar** to the processes that shaped the earth in the past (not comparing rates).
d. Some of the processes are **abrupt**, such as earthquakes and volcanoes, while some are **slow**, such as continental drift and erosion.
e. Slow but continual processes can, **over very long times**, cause significant changes on earth's surface (e.g., wearing down of mountains and building up of sediment by the motion of water).
f. Matching coastlines and similarities in rocks and fossils suggest that today's **continents** are separated parts of what was long ago a single vast continent.
g. The solid crust of the earth consists of **separate plates that move** very slowly, pressing against one another in some places, pulling apart in other places.
h. Major geological events, such as earthquakes, volcanic eruptions, and mountain building, **result from** these **plate motions**.

References

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. (manuscript in preparation). Resources for science literacy: Curriculum materials evaluation. New York: Oxford University Press.

Anderson, C.W. & Smith, E . (1987). Teaching science: In Richardson-Koehler V. (Ed.), The educator's handbook: A research perspective (pp. 84–111). New York: Longman.

Anderson, C., Sheldon, T., & Dubay, J. (1990). The effects of instruction on college nonmajors' conceptions of respiration and photosynthesis. Journal of Research in Science Teaching, 27, 761–776.

Bell, B. & Brook, A. (1984). Aspects of secondary students understanding of plant nutrition. Leeds, UK: University of Leeds, Centre for Studies in Science and Mathematics Education.

Bell, B. & Cowie, B. (2001). Formative assessment and science education. Science & Technology Education Library, Vol. 12. Boston: Kluwer Academic.

Berkheimer, G.D., Anderson, C.W., Lee, O., & Blakeslee, T.D. (1988). Matter and molecules. Teacher's guide. Activity book (Occasional Paper No. 122). East Lansing: Michigan State University, Institute for Research on Teaching. (ERIC document Reproduction Service No. ED 300 275).

Black, P. (1998). Assessment by teachers and the improvement of students' learning. In Fraser B.J., and Tobin K.G. (Eds.), International handbook of science education (pp. 811–822). UK: Kluwer Academic.

Black, P. & William, D. (1998). Assessment and classroom learning. Assessment in Education, 5, 7–74.

Brook, A., Briggs, H., & Driver, R. (1984). Aspects of secondary students' understanding of the particulate nature of matter. Leeds, UK: University of Leeds, Centre for Studies in Science and Mathematics Education.

Caldwell, A. & Stern L. (2000, April). Can middle-school science textbooks help students learn important ideas? Findings from Project 2061's Curriculum Evaluation Study: Earth science. Paper presented at the annual meeting of the National Association for Research in Science Teaching, New Orleans, LA.

Driver, R., Squires, A., Rushworth, P., & Wood-Robinson, V. (1994). Making sense of secondary science: Research into children's ideas. London: Routledge.

Eaton, J.F., Anderson, C.W., & Smith, E.L. (1984). Student preconceptions interfere with learning: Case studies of fifth-grade students. Elementary School Journal, 64, 365–379.

Ericsson, K.A., Krampe, R.T., & Tesche-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. Psychological Review, 100, 363–406.

Freyberg, P. (1985). Implications across curriculum. In: Osborne R. & Freyberg P. (Eds.), Learning in science (pp. 125–135). Auckland, NZ: Heinemann.

Gallagher, J.J. (1996). Structure of matter. East Lansing, MI: Michigan State University Press.

Hart, C., Mulhall, P., Berry, A., Loughran, J., & Gunstone, R. (2000). What is the purpose of this experiment? Or can students learn something from doing experiments? Journal of Research in Science Teaching, 37, 655–675.

Johnston, K. & Driver, R. (1989). A case study of teaching and learning about particle theory. Leeds, UK: University of Leeds, Centre for Studies in Science and Mathematics Education.

Kesidou, S. & Roseman, J. (2002). How well do middle school science programs measure up? Findings from Project 2061's curriculum review. Journal of Research in Science Teaching, 39, 522–549.

Kulm, G. & Grier, L. (1998). Mathematics curriculum materials reliability study. Washington, DC: Project 2061, American Association for the Advancement of Science.

Kulm G. (1999). Evaluating mathematics textbooks. Basic Education, 43, 6–8.

Lee, O., Eichinger, D.C., Anderson, C.W., Berkheimer, G.D., & Blakeslee, T.D. (1993). Changing middle school students' conceptions of matter and molecules. Journal of Research in Science Teaching, 30, 249–270.

McDermott, L. (1991). Millican lecture 1990. What we teach and what is learned-closing the gap. American Journal of Physics, 59, 301–315.

Mintzes, J.J., Wandersee J.H., & Novak, J.D. (Eds.). (2000). Assessing science understanding: A human constructivist view. Boston: Academic Press.

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

Neill, D.M. (1997). Transforming student assessment. Phi Delta Kappan, 79, 34–40.

Novick, S. & Nussbaum, J. (1978). Junior high school pupils' understanding of the particulate nature of matter: An interview study. Science Education, 62, 273–281.

Nussbaum, J. (1985). The particulate nature of matter in the gaseous phase. In Driver, R, Guesne, E., & Tiberghien A. (Eds.), Children's ideas in science (pp. 124–144). Milton Keynes, UK: Open University Press.

Osborne, R. & Freyberg, P. (1985). Learning in science: The implications of children's science. Auckland: Heinemann.

Roseman, J., Kesidou, S., & Stern L. (1996, November). Identifying curriculum materials for science literacy: A Project 2061 evaluation tool. Paper distributed at the National Research Council colloquium, "Using Standards to Guide the Evaluation, Selection, and Adaptation of Instructional Materials," Washington, DC.

Roth, K. & Anderson, C. (1987). The power plant: Teacher's guide to photosynthesis. Occasional paper no. 112. Institute for Research on Teaching. East Lansing, MI: Michigan State University Press.

Rudman, H.C. (1987). Testing and teaching: Two sides of the same coin? Studies in Educational Evaluation, 13, 73–90.

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, 265–296.

Schneps, M.H. & Sadler, P.M. (1988). A private universe. Program 2. Biology: Lessons pulled from thin air. In: New York: Annenberg/CPB.

Shavelson, R.J., Carey, N.B., & Webb, N.M. (1990). Indicators of science achievement: Options for a powerful policy instrument. Phi Delta Kappan, 71, 692–697.

Smith, E. & Anderson, C. (1986, April). Alternative conceptions of matter cycling in ecosystems. Paper presented at the annual meeting of the National Association for Research in Science Teaching, San Francisco, CA.

Stavy, R. & Tirosh, D. (1993). When analogy is perceived as such. Journal of Research in Science Teaching, 30, 1229–1239.

Stern, L. & Roseman, J. (2000, April). Can middle-school science textbooks help students learn important ideas? Findings from Project 2061's Curriculum Evaluation Study: Life science. Paper presented at the annual meeting of the National Association for Research in Science Teaching, New Orleans, LA.

Treagust, D.F. (1988). Development and use of diagnostic tests to evaluate students' misconceptions in science. International Journal of Science Education, 10, 159–169.

Treagust, D.F., Jacobowitz, R., Gallagher J.L., & Parker J. (2001). Using assessment as a guide in teaching for understanding: A case study of a middle school science class learning about sound. Science Education, 85, 137–157.

Tyson-Bernstein, H. (1988). America's textbook fiasco: A conspiracy of good intentions. Washington, DC: Council for Basic Education.

Wise, K.C. & Okey, J.R. (1983). A meta-analysis of the effects of various science teaching strategies on achievement. Journal of Research in Science Teaching, 20, 419–435.

White, R. & Gunstone R. (1992). Probing understanding. London: Falmer Press.

Zucker, A.A., Young, V.M., & Luczak, J. (1996, October). Evaluation of the American Association for the Advancement of Science Project 2061 (SRI Project 7838). Menlo Park, CA: SRI International.