CBE—Life Sciences Education
Vol. 9, 441–452, Winter 2010

Article

The Development of a Conceptual Framework and Tools to Assess Undergraduates' Principled Use of Models in Cellular Biology

Gail Richmond,* Brett Merritt,* Mark Urban-Lurain,† and Joyce Parker*†

*Department of Teacher Education and †Division of Science and Mathematics Education, Michigan State University, East Lansing, MI 48824

Submitted November 16, 2009; Revised May 25, 2010; Accepted May 25, 2010
Monitoring Editor: Marshall Sundberg

Recent science education reform has been marked by a shift away from a focus on facts toward deep, rich, conceptual understanding. This requires assessment that also focuses on conceptual understanding rather than recall of facts. This study outlines our development of a new assessment framework and tool—a taxonomy— which, unlike existing frameworks and tools, is grounded firmly in a framework that considers the critical role that *models* play in science. It also provides instructors a resource for assessing students' ability to reason about models that are central to the organization of key scientific concepts. We describe preliminary data arising from the application of our tool to exam questions used by instructors of a large-enrollment cell and molecular biology course over a 5-yr period during which time our framework and the assessment tool were increasingly used. Students were increasingly able to describe and manipulate models of the processes and systems being studied in this course as measured by assessment items. However, their ability to apply these models in new contexts did not improve. Finally, we discuss the implications of our results and the future directions for our research.

INTRODUCTION

Over the past 30 yr U.S. science education has repeatedly been cited as being in need of repair (National Science Board Commission on Precollege Education in Mathematics, Science, and Technology, 1983; National Commission on Excellence in Education, 1983; Rutherford, 1990). Such characterizations have focused primarily on K–12 science education. However, criticisms of postsecondary science education in the mid- to late-1980s and throughout the 1990s have strengthened the opinion that the entire K–18+ continuum of science education is in need of serious reform (National Science Board, 1986; National Research Council [NRC], 1996a; National Science Foundation, 1996; Boyer, 1998). It has been widely suggested that one solution to the present problems is to recalibrate science teaching so as to emphasize deep understanding of scientific concepts rather than acquisition of facts (e.g., Tanner and Allen, 2005).

However, when teaching and learning are more intimately wedded to understanding rather than fact acquisition, a particular problem arises—what counts as understanding? Many efforts to improve scientific understanding have focused on finding ways to bring students' everyday experiences to bear on scientific problems (NRC, 1999; Nemirovsky *et al*., 2005). The difficulty with such efforts is that most scientific ideas are not closely tied to learners' everyday experiences (e.g., those associated with extremely large or small spatial and/or temporal scales). As a result, students rely on one of two strategies (or some combination of the two) for learning. They may try to fit their experiences to the situation, which results in a narrative explanation of how a system operates or a process unfolds. Alternatively, they may make use of disconnected bits of knowledge or facts they glean from their school-based science experiences to make some sense of these systems or processes (Bransford *et al*., 1999). Neither of these strategies involves

DOI: 10.1187/cbe.09–11–0082
Address correspondence to: Gail Richmond (gailr@msu.edu).

students in making use of scientific principles in rigorous ways or in seeing the system in a way that is similar to how scientists might conceptualize them—as a complicated set of interacting and dynamic processes governed by fundamental principles. Nor does this kind of reasoning provide predictive power. Several researchers have argued that student knowledge might appear fragmented and inconsistent because it consists of weakly organized resources (McDermott, 1991; e.g., DiSessa, 1993)—what Redish (2003) calls modular reasoning. In contrast, we want students to engage in what we call "the principled use of models," that is, we want them to use the models presented to them in class along with fundamental principles such as conservation of energy or matter to make sense of complex processes they are studying.

In addition to the problem of deciding what counts as understanding, two other problems arise when teaching, learning, and understanding are set within the context of large-enrollment undergraduate science courses undergoing reform. First, how does one assess large numbers of students to determine the extent to which they have developed deep, rich, and/or conceptual understanding of the course material in a useful and timely manner? Second, what is a reasonable taxonomy for sorting assessment items so that items that demand similar levels of principled use of models can be compared in longitudinal studies of courses undergoing change?

The work of our research group has been directed at these problems; the study presented here is part of a larger research program investigating teaching and learning in undergraduate science courses. In this article, we describe the development of a unique assessment taxonomy that allows biology instructors to sort assessment items according to the level of the principled use of models they demand of students for the purpose of generating descriptions of student performance and transfer of learning. We also present preliminary data arising from the application of our assessment taxonomy to the exam questions used in a large-enrollment foundational cell and molecular biology course over a 5-yr period for the purpose of studying the effects of instructional changes. More specifically we asked two research questions:

1. From 2002–2007 (the years that immediately preceded and spanned the first three years of our project work), what level of performance with respect to models of photosynthesis and cellular respiration did instructors of a large-enrollment cells and molecules course (hereafter referred to as "Bio101") demand of their students on multiple-choice exam questions?
2. From 2002–2007, how did students' performance in this course change with respect to assessment items about models of photosynthesis and cellular respiration?

Finally, we discuss the implications of our findings and future directions for research.

BACKGROUND

Efforts to Define Understanding in Science Education

At the K–12 level, prominent national documents such as the NRC's National Science Education Standards (NRC, 1996b), the American Association for the Advancement of Science's (AAAS) Science for All Americans (Rutherford, 1990), and AAAS's Benchmarks for Science Literacy (AAAS, 1993) have created frameworks for understanding. In all of these reform documents, understanding requires both the "integration" and the "use" of knowledge; understanding is thus conceptualized in terms of both knowledge and performances or actions.

What counts as understanding at the undergraduate level, however, is more opaque than in K–12 documents. Although some college and university instructors currently look to Standards and/or Benchmarks for guidance in aspects of their planning, assessment, and instruction, in most cases, undergraduate biology instructors are likely to seek more local ideas regarding what counts as understanding (Tanner and Allen, 2002). For example, college and university biology instructors might look to statements of goals and/or objectives in textbooks, to their previous experiences as instructors (or as students themselves), or to curricular materials from previous course instructors.

One effort to address understanding specifically at the undergraduate level is the NRC report, Improving Undergraduate Instruction in Science, Technology, Engineering, and Mathematics: Report of a Workshop (2003a). The focus of this document is the critical importance of "conceptual" and "functional" understanding; thus, as in K–12 documents, understanding is conceptualized in terms of both knowledge and performances—though in the university context, application of concepts or principles rather than action is used. Application, as Bloom (1956) defined it, is a way of describing a learning objective. "Given a problem new to the student, he will apply the appropriate abstraction without having to be prompted as to which abstraction is correct or without having to be shown how to use it in that situation" (p. 120). This ability or practice is also referred to as *transfer* (cf Committee on Developments in the Science of Learning, 2000).

In another NRC publication, BIO2010: Transforming Undergraduate Education for Future Research Biologists (2003b), "conceptual understanding" is also addressed. However, in this document, while the concepts that undergraduate science students should understand are listed, little attention is paid to the specific types of performances that would reveal this understanding. A similar treatment of key concepts without explicit links to performances appears in the Massachusetts Institute of Technology-developed Biology Concept Framework (BCF; Khodor *et al*., 2004). Thus, the performances that would reveal understanding of key concepts have yet to be articulated. To assess these performances, what is required is not only a tool that measures them (e.g., Smith and Tanner, 2010), but also a framework designed to reflect student understanding that identifies and describes these key performances. What follows is a brief description of three recently developed tools and frameworks.

Existing Tools and Frameworks for Assessing Student Knowledge/Performance

Concept Inventory Tests. Concept Inventory Tests—for example, those currently used widely in physics (Hestenes *et al*., 1992), biology (Anderson *et al*., 2002), and genetics (Smith *et al*., 2008)—are a collection of discipline-specific assess-

ment tools designed to allow instructors to acquire information regarding students' conceptual understanding. In most cases, the tool itself consists mainly of multiple-choice questions with each of the distractors or foils—i.e., the incorrect answers—based on well-documented misconceptions. The framework underlying these tools is based on assumptions about students, teaching, and learning consistent with conceptual change theory (Posner *et al*., 1982). When students select certain distractors in individual or small groups of questions, instructors have access to specific information regarding students' misconceptions. Once identified, these misconceptions can be used to inform instruction via specific types of interventions (Posner *et al*., 1982; Strike and Posner, 1992). Understanding is conceptualized in terms of both knowledge and practices. In the Biology Concept Inventory (BCI), for example, knowledge mainly takes the form of knowing "concepts," and practices primarily take the form of actions like conclude, compare, identify, interpret or translate, predict, and explain (Garvin-Doxas and Klymkowsky, 2008; Klymkowsky and Garvin-Doxas, 2008).

Diagnostic Question Clusters. This approach is designed to allow instructors to look specifically at students' understanding of matter, energy, systems, and scale using Diagnostic Question Clusters (DQCs). Like the BCI, the DQC group has designed an assessment tool that consists of multiple-choice questions with distractors based on patterns in student thinking (Wilson *et al*., 2006). The DQCs also resemble the BCI, in that cluster questions ask students to take actions like predict and explain, but the main emphasis of their assessment tools—whether for photosynthesis or for cellular respiration—is to gauge students' ability to trace and conserve matter and energy in and through dynamic systems at various scales. In other words, the DQCs look at students' ability to use broader principles as opposed to the more specific concepts addressed in the BCI. While tools like the BCI and the DQCs share certain similarities in terms of how they conceptualize understanding, by foregrounding different knowledge and performances they afford instructors different views of student understanding in biology.

Blooming Biology Tool. The two approaches to assessing student understanding outlined above are based on new types of multiple-choice items, in both cases developed through rigorous and recursive processes. An entirely different approach to helping instructors assess students' understanding is to offer instructors a tool for analyzing already existing assessment items. The Blooming Biology Tool (BBT; Crowe *et al*., 2008) uses the familiar language of Bloom's taxonomy of educational objectives (Bloom, 1956) to classify already existing exam questions. Its framework for understanding is defined primarily by familiar Bloom categories such as knowledge, comprehension, application, analysis, synthesis, and evaluation. Via the BBT's interpretation of Bloom's taxonomy, understanding assumes various "levels" and "skills." Memorization and recall, for example, are "lower-order cognitive skills that require a minimum level of understanding," whereas things like the application of knowledge and critical thinking are "higher-order cognitive skills that require deep conceptual understanding" (Crowe *et al*., 2008, p. 368). As with Concept Inventory Tests and the DQCs, understanding involves both knowledge and performances. The tool is not tied to any particular concept or topic in biology. However, because higher levels of understanding in the BBT (i.e., "deep conceptual understanding") are tied to performances such as the "application" of knowledge, understanding is considered the product of both knowledge and performances. The BBT affords instructors a view of student performance in biology that, like the Concept Inventories and the DQCs, is sensitive to assessment challenges faced in large-enrollment courses.

Assessing Understanding in Large-Enrollment Undergraduate Courses

Assessment in large-enrollment courses often aims for a workable balance between comprehensiveness and efficiency which, in most cases, results in the use of exams with items that can be computer-scored. One of the advantages of such assessments is their expediency: they can be scored and returned to students in a short period of time. In addition, the use of scoring offices or testing services found at many large institutions offers instructors additional types of feedback. Instructors, for example, can make use of basic statistical summaries provided by such services to determine the least/most difficult questions on an exam. Instructors can also conveniently do things like view data describing the least/most frequently chosen answers to a multiple-choice question.

One response to the problem of assessing large numbers of students in courses has been to encourage instructors to include additional forms of assessment, such as the use of open-ended questions, project-based assessment, and/or the use of student portfolios (Angelo and Cross, 1993; NRC, 1996b, 2001). Another has been to increase the frequency of different types of assessment, e.g., formative assessments, during instruction (NRC, 1996b; Black and Wiliam, 1998; Yorke, 2003). Technology has also played a significant role in assessment conversations. For example, much has been reported on the use of handheld electronic devices or "clickers" as a tool for assessing student understanding (Mazur, 1996; Brewer, 2004; Fitch, 2004; Zhang *et al*., 2005; Campbell and Mayer, 2006; Caldwell, 2007). Each of these approaches has its affordances and constraints. In large-enrollment science courses, time and resource constraints make the use of some modes of assessment such as open-ended questions a significant challenge.

In our view, the increasing number of frameworks and tools for assessing student understanding provides undergraduate biology instructors with multiple ways of describing student understanding that mobilize a wide range of knowledge and performances; however, these existing tools/frameworks are by no means comprehensive. Below we discuss our development and use of a new framework for assessing student understanding. Like the BBT, our framework is a taxonomic scheme for classifying multiple-choice questions. Unlike what has already been described, our tool is grounded in an underlying framework that considers the critical role that models play in science, as well as the critical role for the use of scientific principles. As such, it offers instructors a resource for assessing understanding

that is grounded in the work of scientists and in the structure of the discipline.

Development of a Taxonomy to Classify Assessments Based on Principled Use of Models

Various case studies of scientific practice have identified the construction and extension of models as one of the central tasks of scientific research (Rouse, 1987; Giere, 1988; Pickering, 1995). Model construction and extension is, as Pickering explains, "constitutive of scientific practice" (1995, p. 55). Such models, in turn, provide frameworks for organizing and directing subsequent research (Downes, 1992). Similar statements about the importance of models and modeling in science have been made by many scientists and scholars of science (Giere, 1988, 1991, 2004; Gilbert, 1991; Magnani *et al*., 1999). Many K–12 science educators and researchers have taken up this general recognition of the importance of models in science, and one can now find an increasing number of studies that examine the use of models in K–12 classrooms (Grosslight *et al*., 1991; Cartier, 2000a, 2000b; Treagust *et al*., 2002; Lehrer and Schauble, 2005).

What counts as a model in science and in science education can be difficult to pin down (e.g., Giere, 2004). Coll and Treagust (2003) recognized this problem in their discussion of the wealth of models including mental models, expressed models, public models, consensus models (including consensus science and consensus teaching models), teaching models, and scientific models. The kind of model that undergirds the assessment tool that we present here is most similar to what Coll and Treagust have called teaching models, which they define as "mental models as presented by teachers." (See also Gilbert *et al*., 2000.) They distinguish teaching models from both "consensus models" (i.e., "public or expressed models made available to the scientific and teaching community") and from "scientific models" (i.e., consensus models that are "subject to and survive rigorous experimental testing, published in scientific literature, and widely accepted by the scientific community"; Coll and Treagust, 2003, p. 465). Teaching models can be viewed as pedagogically tuned simplifications or adaptations of consensus and/or scientific models. Although they are meant to retain many of the intellectual properties of consensus or scientific models, teaching models have been recalibrated for a different audience—science students rather than scientists. An often implicit feature of teaching models is that they adhere to basic scientific principles of consistency.

There is good reason to believe that such principles should be made explicit in science classrooms. The tracing and conserving of matter and energy within and between systems of differing size or scale is fundamental to the different scientific disciplines and has been emphasized by both scientists (c.f. Prigogine, 1961; Schrödinger, 1967; Prigogine and Stengers, 1984; Salthe, 1985) and science educators (c.f. Lemke, 1990, 1995). Perhaps nowhere in the scope of topics of introductory cell biology courses do the importance of these two principles become more apparent than in models of photosynthesis and cellular respiration. For example, tracing a variety of carbon- and oxygen-containing compounds through cells, as well as tracing energy as it is transformed from light to chemical potential energy, are crucial to developing a deep understanding of these two keystone concepts in biology. Therefore the principled use of teaching models is an important goal.

It was with this notion of teaching models that we developed our assessment taxonomy. We used it to evaluate the effects of instructional changes on students' principled use of models of photosynthesis and cellular respiration during a 5-yr period. We present here both the taxonomy and the results of this study.

MATERIALS AND METHODS

Course Context

Bio101 is a large-enrollment, three-credit, introductory cell and molecular biology course offered each semester at our university. From 2003 on the course was cotaught by the same two instructors. Before that, each of the instructors taught the course on his own—one instructor since 1991 and the other since 1996. In many ways, Bio101 is typical of introductory-level courses at other large universities. Each section of the course typically enrolls between 350 and 500 students and meets either thrice weekly for 50-min class periods or twice weekly for 75-min class periods over the course of a 15-wk semester. It has a general chemistry prerequisite, and enrollment is approximately 1600–1800 students per year. Each section is typically taught by one or two instructors (in our case two) with one graduate student in the role of a teaching assistant (TA) to manage homework assignments and assist instructors with the use of technology during lecture. There is no accompanying recitation section, and the associated two-credit laboratory course, which meets once a week for 3 hr during the semester, is not required to be taken concurrently. The course is required for all students majoring in science, including a large percentage of allied health majors. Although homework is a requirement and counted in the overall grading scheme, approximately 90% of a student's course grade comes from performance on three midterm examinations and one final exam.

Development of a Classification Scheme to Delineate Question "Types"

We needed a method to categorize the photosynthesis and cellular respiration multiple-choice questions used on Bio101 exams that would address both knowledge and performance dimensions. We combined a nondisciplinary-specific taxonomic scheme with the disciplinary-specific notion of teaching models. At first, we turned to existing taxonomies such as Bloom's taxonomy of educational objectives (Bloom, 1956) as well as to more recent taxonomies (e.g., Marzano, 2001; Anderson *et al*., 2001). In the end, we decided against the use of these taxonomies because of the disciplinary-specific demands created by our emphasis on teaching models. We set each of the four categories for classifying existing exam questions in relation to specific aspects of such models for photosynthesis and cellular respiration. An example of a common teaching model for photosynthesis, designed specifically for a student audience, is displayed in Figure 1.

Based on our examination of this model alongside others that exist for these topics, we developed a basic four-category taxonomic classification scheme (Table 1).

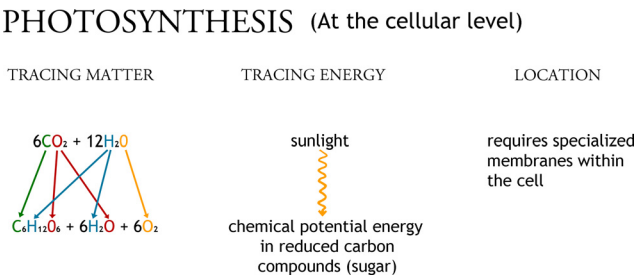

Figure 1. A teaching model for photosynthesis at the cellular level. This teaching model draws particular attention to three key concepts in photosynthesis: tracing matter (keeping track of specific elements in photosynthetic reactions), tracing energy (keeping track of energy transformations in photosynthetic reactions), and location (keeping track of where photosynthetic reactions occur in the cell).

Category 1 represents those questions that may be related to photosynthesis or cellular respiration but do not directly address a core feature of a teaching model as used by the instructors. Category 2 is reserved for questions that require students to describe or reproduce a teaching model in a form similar to that in which it was presented during instruction. In other words, this type of question makes direct use of the same set of standard representations (and/or situations, scenarios, examples, etc.) that were used, for instance, in the lectures and/or in the homework assignments. In Category 3, students are required to infer the logical conclusions of variations (modifications) of a teaching model presented in the question. In other words, they manipulate a model. The model in question, however, is still primarily in a form similar to that in which it was presented during instruction. Finally, Category 4 is reserved for those questions that require students to apply a teaching model (and may also include a manipulation) to a problem beyond the original contextual or instructional boundaries.

RESULTS

Applying the Taxonomy to the 2002–2007 Exam Questions

We applied the taxonomy to all photosynthesis and cellular respiration multiple-choice questions used on midterm and final exams between 2002 and 2007. The questions from years 2006 and 2007 had each been used during the 2002–2005 period, resulting in 149 exam questions to be rated. Five raters (the two course instructors, two science education faculty, and one science education graduate student) rated the questions independently according to the four categories of the taxonomy.

Data were analyzed using SPSS Version 17 for Windows. The ratings from the five raters had a Cronbach's alpha of 0.88. Cronbach's alpha measures how well a set of items (or variables) measures a single unidimensional latent construct and varies from 0 to 1.0. Values above 0.7 or 0.8 are considered very good in the sense that they indicate that raters are attending to the same underlying construct. In this case, 0.88 suggests that the five raters had general agreement on the underlying four-category structure of the taxonomy when applying it to the exam questions. With respect to the intraclass correlations coefficients, there was a Single measure score of 0.60, $p < 0.000$ and an Average measure score of 0.88, $p < 0.000$. Intraclass correlation coefficients measure the degree of reliability of the ratings. They can range from 0.0–1.0. The single measures are the reliability if using just one of the judges. The average measure is the reliability using all the judges. Because our average measures were higher than our single measures, we decided to use the ratings of the 2002–2007 exam questions from all five raters rather than relying on the ratings of any one individual.

To further understand the underlying conceptual structure represented by this taxonomy, we used SPSS to perform a hierarchical cluster analysis on the five individual rater classifications of each of the 149 photosynthesis and cellular respiration exam questions (Figure 2). Cluster analysis groups objects based on their similarity on one or more attributes (in this case exam question ratings from five raters.) The resulting taxonomic tree shows which objects are most and least similar. There are several methods of producing the clusters, but the most commonly used is the unweighted pair-group method using arithmetic averages (UPGMA.) UPGMA begins with a standardized data matrix from which a resemblance matrix is computed that measures the distances between each pair of objects. The objects are then combined into clusters based on their similarities. The clusters are combined until there is a single cluster containing all of the objects (Romesburg, 1984) . The results may be displayed using a dendrogram or tree diagram, as shown in Figure 2. The individual "leaves" at the bottom of the figure are individual questions. The "leaves" have been color-coded based on the average rating each question received. The horizontal lines are the scaled euclidean distances between individuals or groups. Groups are defined by "cutting the tree"—drawing a horizontal line that crosses the vertical lines, or "branches." The branches define the groups. In Figure 2, any horizontal line greater than 11 on the y axis cuts two branches, or groups. The cluster of questions on the right side of the figure (questions with an average rating $\geq$3.0) and the remaining questions on the left side (questions with an average rating $<$3.0) are least similar, with an average distance of 25 between those groups. The maximum difference between any pair of questions in the cluster on the right side of the figure is 7, and the maximum distance between the questions on the left cluster is 11. The only horizontal line that will leave four clusters would be between 10 and 11 on the y axis. However, these four groups do not cluster questions with the most similar interrater agreement, except for the branch on the far right side. The interrater agreement is highest for questions in the

Table 1. Model-based taxonomy for the classification of photosynthesis assessment items

Category	Criteria
1	Not directly associated with features of the specific photosynthesis teaching model as presented
2	Describe or reproduce the specific model
3	Manipulate the photosynthesis model in context
4	Apply the model in situations beyond the original context

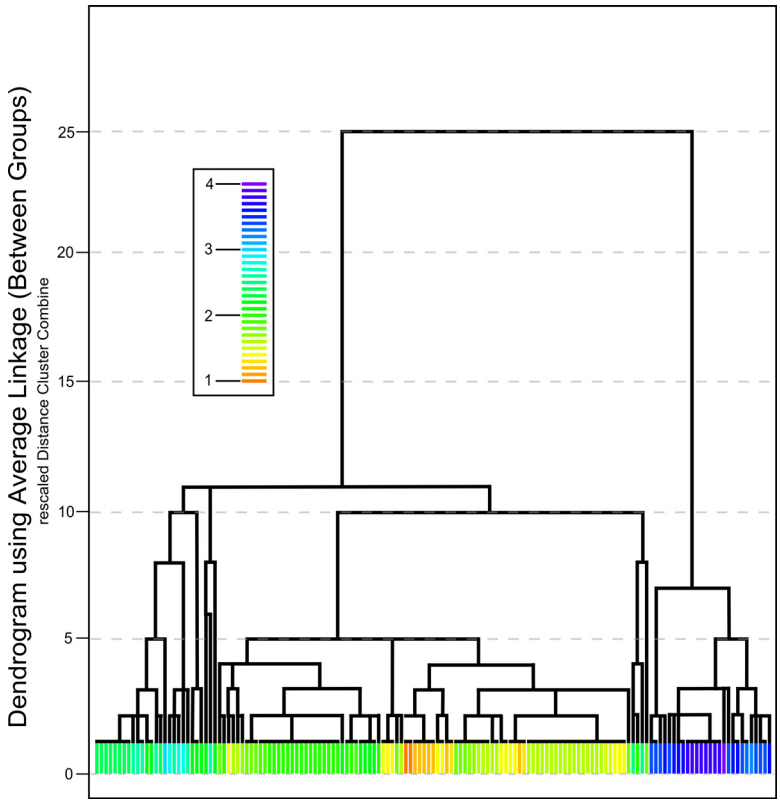

Figure 2. Dendrogram of a hierarchical cluster analysis of 149 photosynthesis and cellular respiration exam questions from 2002 through 2005. Each question is represented by the "leaf" at the bottom of the figure that is color-coded to show the average rating of the question by the five raters. The inset figure indicates the scale for the average of the five reviewers' ratings of categories described in Table 1. The lower the linkage distance score, the greater the agreement among reviewers. Note bifurcation into two distinct branches with questions with average ratings ≥3.0 clustered on the right.

1–2 range (orange to light green) and 3–4 (blue to purple) range because the average linkage distances between questions in these clusters is small (5 and 7 respectively), indicating relatively large agreement between raters. The average linkage distance between questions in the 2–3 range (darker green and light blue) is 11, indicating less agreement across the five raters on questions in the 2–3 range. This figure shows the clearest delineation between those questions with an average rating between 1 and 3 and those with an average rating between 3 and 4 (average linkage 25). This analysis suggests a strong agreement among our five raters that there are two broad clusters or categories of questions based on the taxonomy. We performed our subsequent analyses of student performance on these questions according to the two clusters produced by the hierarchical cluster analysis, Category <3 and Category ≥3.

Student Performance on Exam Questions

To compare performance of students across years, we divided all of the 2002–2007 students into an upper 27.5 percentile (overall GPA = 4.0–3.4070), a middle 45 percentile (3.4069–2.6377), and a lower 27.5 percentile (2.6376–0) group. We calculated the cut scores for each of the groups by taking the incoming GPA of each student between 2002 and 2007 and then using these GPAs to create three distinct groups of students. Using the cut scores from these three groups, we examined student performance on Category <3 questions compared with those rated Category ≥3. Table 2 shows the numbers of students in each GPA group for each semester. There were two main reasons for disaggregating students by GPA. First, there is an increasing awareness in science education of those students who leave the sciences either before or during their postsecondary education (Tobias, 1990; Seymour and Hewitt, 1997). Our assumption in creating the three GPA groups was that our lower and middle groups of students were more likely to contain more of these particular students who are at-risk for leaving. Second, the middle group contained a majority of the future K–12 science teachers. Because our research group included many science teacher educators, they were especially interested in how this group was performing on the different types of exam questions. In addition to the GPA groupings, we also disaggregated student performance according to topic, examining photosynthesis questions separately from cellular respiration questions so as to be able to

Table 2. Number of students in each GPA group per semester

Semester	Lower 27.5 percentile GPA (n)	Middle 45 percentile GPA (n)	Upper 27.5 percentile GPA (n)
FS02	115	134	79
FS03	81	137	106
FS04	86	202	121
FS05	73	165	109
FS06	101	201	116
FS07	72	138	68

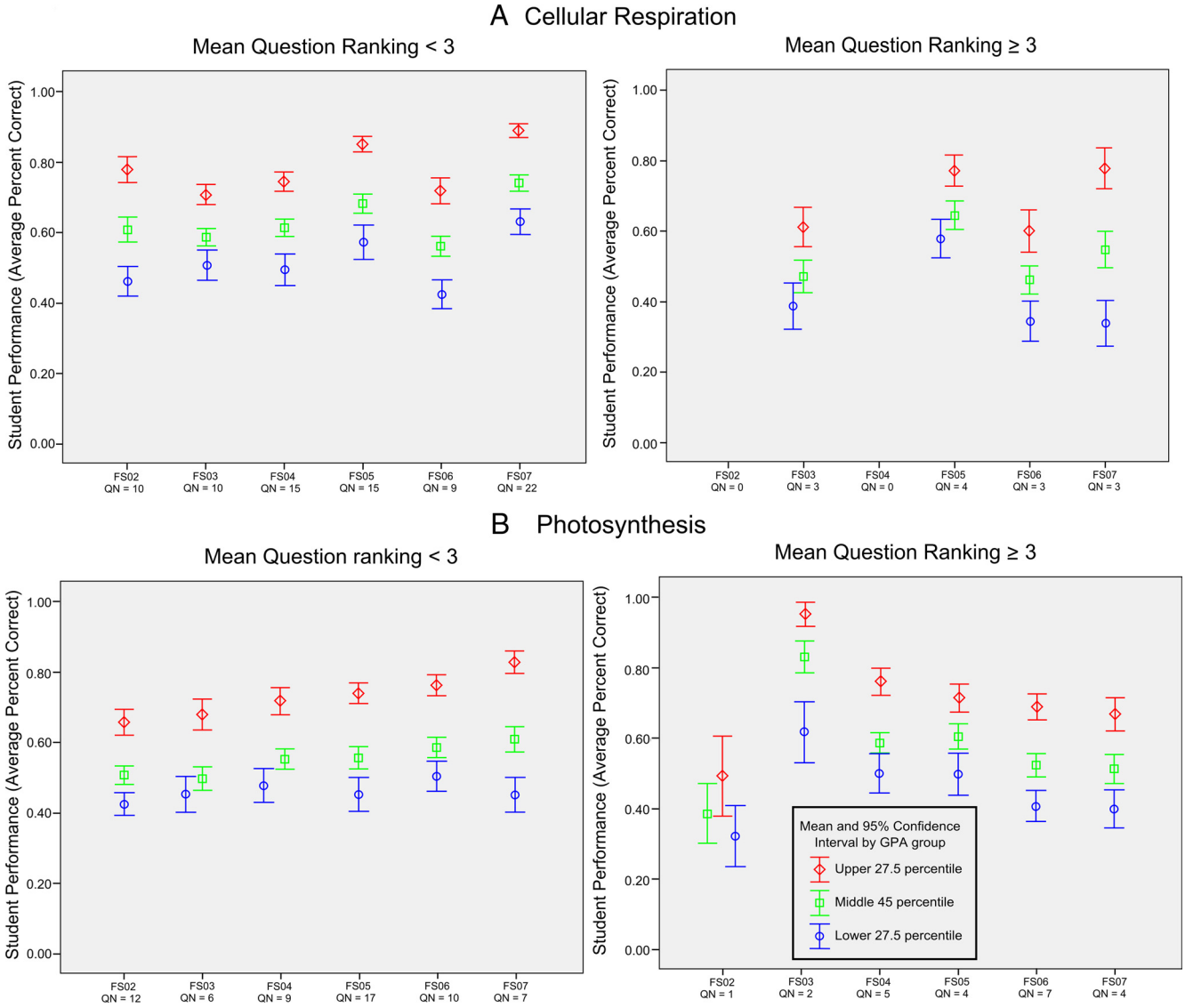

Figure 3. Student performance on cellular respiration exam questions (top of A) and photosynthesis exam questions (bottom of B) by category of question and GPA group (2002–2007). The first column contains graphs of student performance on questions rated <3. The second column contains graphs of student performance on questions rated ≥3. Performance for each of the three GPA groups (lower 27.5 percentile, middle 45 percentile, and upper 27.5 percentile) are shown with the means and 95% confidence intervals. QN is the number of questions in a particular category in a particular year.

detect any differences in student performance between these two topics.

Cellular Respiration. Figure 3A shows student performance on cellular respiration midterm exam questions (administered following the instruction on respiration) by category of question and GPA group across semesters. The semester/years are shown on the *x* axes. The percentage of correct questions are shown on the *y* axes. Performance by students in the lower, middle, and upper GPA groups are shown in the three rows of symbols indicated by color and shape of the symbol marking the mean for that year and the 95% confidence intervals shown for each group. Performance on questions with average ratings Category <3 are shown in the left column; performance on questions with average ratings categorized ≥3 are in the right column. In 2002 and 2004, there were no cellular respiration questions categorized as ≥3, so there are no values shown for those semesters in the figure.

For Category <3 cellular respiration questions, there is a trend of improvement in student performance in the upper, middle, and lower groupings between 2002 and 2007. In terms of our framework for understanding, with the exception of 2006 in which there is a statistically significant drop in student performance that we discuss below, Bio101 students are generally improving from year to year in their ability to describe/reproduce the cellular respiration model,

to manipulate the teaching model within the contextual boundaries in which it was originally presented during instruction, and also to answer cellular respiration questions not directly related to specific aspects of the teaching model. By 2007, the 95% CI bars are above the values for years 2002–2004 in all three GPA groups, showing that students were performing better than they had been in the 2002–2004 time frame ($p < 0.05$).

It is more difficult, however, to make any claims about Category ≥3 questions. Given the 95% confidence intervals and the smaller numbers of questions in this classification category, there is no trend of improvement in students' ability to answer these types of questions. In terms of our framework for understanding, this suggests that Bio101 students show no discernable year-to-year improvement in their ability to manipulate and/or apply the cellular respiration teaching model beyond the contextual boundaries in which it was originally presented during instruction.

Photosynthesis. Figure 3B shows student performance on photosynthesis midterm exam questions by category of question and GPA group. The semester/years are shown on the *x* axes. The percentage of correct questions are shown on the *y* axes. Performance by students in the lower, middle, and upper GPA groups are shown in the three rows of symbols indicated by color and shape of the symbol marking the mean for that year along with the 95% confidence intervals. Performance on questions with average ratings Category <3 are shown in the left column; performance on questions with average ratings categorized ≥3 are in the right column.

Again, for Category <3 photosynthesis questions, there is also a trend of improvement in student performance in the upper, middle, and lower groupings between 2002 and 2007, though the improvement for the lower GPA grouping is not statistically significant. In contrast with the cellular respiration questions, however, there is no drop in student performance in 2006. On the contrary, 2006 student performance follows a general trend in which Bio101 students demonstrate continuous improvement on these types of questions. In terms of our framework for understanding, Bio101 students are generally improving from year to year in their ability to describe/reproduce the photosynthesis model, to manipulate the teaching model within the contextual boundaries in which it was originally presented during instruction, and also to answer photosynthesis questions not directly related to specific aspects of the teaching model.

As is the case with Category ≥3 cellular respiration questions, student performance on Category ≥3 photosynthesis questions appears to decline; however, this is not statistically significant. In terms of our framework for understanding, this suggests that Bio101 students show no discernable year-to-year improvement in their ability to manipulate and/or apply the photosynthesis teaching model beyond the contextual boundaries in which it was originally presented during instruction.

Summary of Exam Question Types

In terms of assessing students for understanding in photosynthesis and cellular respiration, Figure 3 displays two discernible trends. First, the total number of photosynthesis and cellular respiration questions included on exams increased sharply in 2004 from approximately 30–50, and this increase was maintained between 2004 and 2007, except for 2006, when it was 30. Compared with 2002–2003, the Bio101 instructors were allotting larger portions of their 2004–2007 exams to questions regarding these two topics. We know this to be true because the number of total exam questions used each year between 2002 and 2007 remained relatively constant. Second, as a percentage of the total number of questions asked per year, the number of questions rated Category ≥3 increased substantially. Between 2002–2004, Category ≥3 questions constituted an average of 14% of the total photosynthesis and cellular respiration exam questions. Between 2005–2007, however, these same types of questions constituted nearly 30% of the topic-specific questions. Compared with 2002–2003, the Bio101 instructors were asking more questions on their in 2005–2007 exams in which the students were expected to manipulate and/or apply the teaching models in unfamiliar contexts.

Taken collectively, these two basic trends highlight 2004–2005 as a two-year period in the course in which the Bio101 instructors began a shift away from previous assessment practices. Not only did they start devoting a larger percentage of their total exam questions to the photosynthesis and cellular respiration teaching models, but they also started devoting more of the photosynthesis and cellular respiration exam questions to Category ≥3 type questions. In the language of our framework for understanding, starting in 2004–2005 Bio101 students were asked to describe, manipulate, and apply the teaching models more than Bio101 students in previous years.

DISCUSSION

Because the total number of Category ≥3 photosynthesis or cellular respiration exam questions never exceeded seven questions in a single year, we ran into difficulty performing statistically meaningful tests. Thus, in 2003, for example, the year in which students appeared to perform significantly better on Category ≥3 photosynthesis questions than in other years, there were only two questions available for use in the analysis.

The cluster analysis suggests that, while there is good rater agreement on the classifications of exam questions using the four-category system, there is little difference among Categories 1–3, which is the rationale for analyzing them in the two groups rather than all four. In addition, very few of the same questions were used in multiple years. This may account for differences in performance across years as different questions may have different levels of difficulty. Question difficulty (how students perform) is not necessarily a function of the taxonomic classification. There can be questions on which many students do well, or poorly, at any taxonomic scale. See Anderson *et al*. (2001, chapter 16) for a discussion of the empirical evidence for student performance as a function of Bloom's taxonomic classification of questions. Their meta-analyses show stronger correlations in student performance on Bloom's taxonomies are between knowledge ("lowest" level) and evaluation ("highest" level) than between adjacent levels, suggesting that student performance is *not* a function of the Bloom taxonomy of the questions.

Several findings of importance are discussed below in greater detail. These include 1) confronting a long-standing pattern observed by the course instructors regarding class averages on exam questions, 2) confronting the 2006 "dip" in student performance on Category <3 cellular respiration—but not photosynthesis—questions, and 3) confronting the overall lack of steady gains in student performance on Category ≥3 questions for both photosynthesis and cellular respiration.

The 62% Rule

Course reform efforts in Bio101, particularly between 2002 and 2007, had been significant. Focused mostly on teaching and learning of photosynthesis and cellular respiration, key concepts for the course, they had included the introduction of handheld, electronic student response devices or "clickers," the development and administration of conceptually linked "sequences" of clicker questions to address key course concepts, and the use of the DQCs (Wilson *et al.*, 2006) both in class and on exams. Despite the 6-yr-long commitment to course reform, the course instructors observed that class averages on both midterm and final exams between 2002 and 2007 rarely surpassed 62%. These averages were based on all questions, including both DQC and non-DQC questions. The "62% Rule," as it came to be called, was cited by the instructors as evidence that the extensive course reforms had not had their anticipated (beneficial) effects on student understanding, at least as measured by student performance on exams.

The results of our analysis of student performance on Category <3 and Category ≥3 photosynthesis and cellular respiration questions provide us with a different perspective on the 62% Rule. While the overall exam averages remained fairly constant between 2002 and 2007 (there were only 3 of 24 total exams in which the class average was higher than 62%), we found that students were in fact improving on Category <3 photosynthesis and cellular respiration questions and that their performance on Category ≥3 questions remained, for the most part, the same. For the class averages to remain steady at 62%, these gains in performance in Category <3 questions must have been consistently offset by students' performance on questions about other topics, such as cell cycle and DNA transcription and translation.

2006 Dip in Student Performance

One potential explanation of the 2006 decrease in student performance on the Category <3 cellular respiration exam questions (Figure 3) involves both temporal and conceptual components. In almost every year included in our study, the Bio101 instructors tested their students on cellular respiration and photosynthesis on the same midterm exam (the second of three midterm exams). In 2006, however, the initial unit on large biomolecules was dispersed throughout the course. The result was that the instructors taught the cellular respiration unit earlier in the semester and included it on the first midterm exam but kept photosynthesis on the second exam. Thus students prepared for the exams on the photosynthesis and cellular respiration models separately. It is possible that this instructional change had conceptual consequences and that the performance decrease that year indicates the significant conceptual "dividends" of addressing these two topics close together in terms of instruction and assessment. This raises the question: How might certain ways of understanding photosynthesis have desirable effects or influences on particular ways of understanding cellular respiration?

Student Performance on Category ≥3 Questions

While, with the exception of 2006, student performance on Category <3 exam questions resulted in steady gains, student performance on Category ≥3 exam questions did not. There are many possible explanations for the students' inability to apply and/or manipulate the models of photosynthesis and cellular respiration beyond the contextual boundaries in which they were originally presented during instruction. One is that while our implemented reforms led to changes in some components of the course, they may not have been as effective in promoting changes in other areas. For example, our framework for understanding and assessing the processes of photosynthesis and cellular respiration allows for a certain degree of precision and specificity not found in other existing frameworks for understanding biology. How the affordances of this framework are communicated to and implemented by instructors is a complex undertaking and thus is still a work in progress. A case in point is the Bio101 instructors' use of frameworks associated with the DQC project. The DQC frameworks for understanding organize the content associated with cellular respiration and photosynthesis around a set of specific practices—such as tracing matter, tracing energy, and keeping careful track of issues related to scale and context/location. The goal of the DQC frameworks is to help students reduce the perceived complexity of topics like photosynthesis and cellular respiration in a systematic way such that they can then apply a similar framework to problems and topics beyond photosynthesis and cellular respiration. While the instructors engaged certain dimensions of the DQC frameworks by asking students questions and/or emphasizing particular terms and concepts during the lectures, they did not explicitly present, construct, and/or otherwise explore and unpack the DQC frameworks for understanding with their students. Given their goals for student understanding of these concepts, the explicit and consistent use of the frameworks would go a long way toward helping students better understand the specific knowledge and performances that are expected of them. Such use requires significant investment of time and commitment, and we are encouraged by the changes we have observed as our colleagues have begun to take on the task of instructional reform using these frameworks and assessment tools as both instructional and diagnostic resources.

Our work has also led us to an interest in examining more closely the questions we are presently classifying as Category ≥3. It is possible that questions within this category vary in their demands on student understanding. For example, some Category ≥3 questions contain highly technical discipline-specific terminology, whereas others with the same rating consist of less technical, more commonly used terms. An example of a technical item is a question the Bio101 instructors regularly use about thylakoid membranes ("Assume a thylakoid is somehow punctured so that the

interior of the thylakoid is no longer separated from the stroma. This damage will have the most direct effect on which of the following processes?"). The mean rating of this question was 3.4; 39% of the students answered it correctly. An example of a less technical item is a question the instructors commonly use about the growth of maple trees ("A mature maple tree can have a mass of 1 ton or more—dry biomass, after removing the water—yet it starts from a seed that weighs <1 g. Which of the following processes contributes the most to this huge increase in biomass?") The mean rating of this question was 3.8; 48.5% of the students answered it correctly. It is possible that these two different types of questions place particular types of linguistic and/or conceptual demands on students that we have yet to fully account for in our present framework for understanding.

Implications and Conclusions

One of the challenges we are addressing in our work is to find more explicit and productive ways to "tune" the visual representations commonly used during instruction of photosynthesis and cellular respiration to our evolving framework for understanding. One way this tuning has occurred is that the course instructors have created more opportunities for their students to practice describing, manipulating, and applying the models in class. For example, one of the two instructors created a pedagogical intervention with the aid of iClicker® technology. Using PowerPoint and the handheld student response system, he helped students construct an important teaching model for cellular respiration from scratch by using an interactive image-driven story of a marshmallow burning over a campfire. Another way this tuning has occurred is that we have begun to foreground those principles featured in our framework and in the teaching models themselves. For example, we are working with course instructors to fine-tune standard visual representations for photosynthesis and cellular respiration (i.e., their teaching models) and align them more carefully and explicitly with the practices articulated by the DQC framework (i.e., tracing matter, tracing energy, and identifying scale and location).

Explicit instructional moves to share teaching models as well as relevant practices with students seems particularly important given what Lemke has called the "mystique of science" (Lemke, 1990). Our experiences suggest that science instructors should develop prolonged, strategic, and explicit ways of sharing what counts as target understandings of important concepts with their students. Our initial investigations, in which we explicitly situate a problem within the framework and continually engage students in making use of these practices, suggest that this mystique is removed and increased understanding results.

While we have made use of multiple-choice format questions as a vehicle for assessing students' ability to engage in principled use of teaching models, we also have used more open-ended question formats. (In fact, it was from student answers to such assessments that many of our foils for the multiple-choice items were identified.) Thus, there is no reason why these frameworks and tools could not be used with short-answer or essay questions and provide valuable feedback to instructors and students alike. In addition, our taxonomy can be adapted to any teaching model and therefore can be used by instructors to align assessment items with their teaching model and used for similar longitudinal studies.

Finally, we are pleased that our focus on principled use of models has been taken up by those in other disciplines (e.g., geology, chemistry), both at our institution and at others around the country. We have made the resources we have developed available at http://dqc.crcstl.msu.edu. We anticipate that as we further develop and refine teaching models and assessments that reflect core concepts of a discipline, we also will be able to develop more effective pedagogical strategies to support students' ability to make principled use of these models to understand science more deeply and to see the interrelationships among the different science disciplines.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the contributions of Andy Anderson, Merle Heideman, Amy Lark, Tammy Long, John Merrill, Rosa Moscarella, Ron Patterson, Aaron Russell, Duncan Sibley, and Chris Wilson to this work. We also thank Miles Loh for his assistance with the preparation of figures and the two anonymous reviewers and editor for their helpful comments. This research was supported by grants from the NSF (DUE-0243126 and Cooperative Agreement EHR 0314866) and the Carnegie Corporation (B7458). The views presented here represent those of the authors and not the funding agencies.

REFERENCES

Anderson, D. L., Fisher, K. M., and Norman, G. J. (2002). Development and evaluation of the conceptual inventory of natural selection. J. Res. Sci. Teach. *39*, 952–978.

Anderson, L. W., Krathwohl, D. R., and Bloom, B. S. (2001). A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, New York: Longman.

Angelo, T. A., and Cross, K. P. (1993). Classroom Assessment Techniques: A Handbook for College Teachers. (2nd ed.), San Francisco, CA: Jossey-Bass Publishers.

Black, P., and Wiliam, D. (1998). Assessment and classroom learning. Assess. Educ. *5*, 7–74.

Bloom, B. S. (1956). Taxonomy of Educational Objectives: The Classification of Educational Goals (1st ed.), New York: Longman, Green.

Boyer, E. (1998). The Boyer Commission on Educating Undergraduates in the Research University, Reinventing Undergraduate Education: A Blueprint for America's Research Universities. Stony Brook, NY.

Bransford, J., Brown, A., and Cocking, R. (eds.). (1999). How People Learn: Brain, Mind, Experience, and School. Washington, DC: National Academies Press. http://books.nap.edu/html/howpeople1.

Brewer, C. (2004). Near real-time assessment of student learning and understanding in biology courses. Bioscience. *54*, 1034–1039.

Caldwell, J. E. (2007). Clickers in the large classroom: current research and best-practice tips. CBE Life Sci. Educ. *6*, 9–20.

Campbell, J., and Mayer, R. (2006). Fostering cognitive activity in college lecture classes. Presented at the Annual Meeting of the American Educational Research Association (April 10, 2006), San Francisco, CA.

Cartier, J. (2000a). Using a Modeling Approach to Explore Scientific Epistemology with High School Biology Students. Research, Madison, WI: National Center for Improving Student Learning and Achievement in Mathematics and Science.

Cartier, J. (2000b). Assessment of Explanatory Models in Genetics: Insights into Students' Conceptions of Scientific Models. Research, Madison, WI: National Center for Improving Student Learning and Achievement in Mathematics and Science.

Coll, R. K., and Treagust, D. F. (2003). Investigation of secondary school, undergraduate, and graduate learners' mental models of ionic bonding. J. Res. Sci. Teach. *40*, 464–486.

Committee on Developments in the Science of Learning (2000). How People Learn. Washington, DC: National Academies Press.

Crowe, A., Dirks, C., and Wenderoth, M. P. (2008). Biology in bloom: implementing Bloom's taxonomy to enhance student learning in biology. CBE Life Sci. Educ. *7*, 368–381.

DiSessa, A. (1993). Towards an epistemology of physics. Cognition and Instruction *10*, 105–225.

Downes, S. (1992). The importance of models in theorizing: a deflationary semantic view. In D. Hull, M. Forbes, and K. Okruhlik (Eds.), *PSA 1992*, Vol. 1. Proceedings of the 1992 Biennial Meeting of the Philosophy of Science Association. East Lansing, MI: Philosophy of Science Association, 142–153.

Fitch, J. (2004). Student feedback in the college classroom: a technology solution. Educ. Tech. Res. Dev. *52*, 71–77.

Garvin-Doxas, K., and Klymkowsky, M. W. (2008). Understanding randomness and its impact on student learning: lessons learned from building the biology concept inventory (BCI). CBE Life Sci. Educ. *7*, 227–233.

Giere, R. N. (1988). Explaining Science: A Cognitive Approach, Chicago, IL: University of Chicago Press.

Giere, R. N. (1991). Understanding Scientific Reasoning (3rd ed.), Fort Worth, TX: Holt, Rinehart, and Winston.

Giere, R. N. (2004). How models are used to represent reality. Philos. Sci. *71*, 742–752.

Gilbert, S. (1991). Model building and a definition of science. J. Res. Sci. Teach. *28*, 73–79.

Gilbert, J., Boulter, C., and Rutherford, M. (2000). Explanations with models in science education. In: Developing Models in Science Education, ed. J. Gilbert and C. Boulter, Dordrecht: Kluwer Academic Publishers, 193–208.

Grosslight, L., Unger, C., Jay, E., and Smith, C. L. (1991). Understanding models and their use in science: conceptions of middle and high school students and experts. J. Res. Sci. Teach. *28*, 799–822.

Hestenes, D., Wells, M., and Swackhamer, G. (1992). Force concept inventory. The Physics Teacher *30*, 141–158.

Khodor, J., Halme, D. G., and Walker, G. C. (2004). A hierarchical biology concept framework: a tool for course design. Cell Biol. Educ. *3*, 111–121.

Klymkowsky, M. W., and Garvin-Doxas, K. (2008). Recognizing student misconceptions through Ed's tools and the biology concept inventory. PLoS Biol. *6*, e3.

Lehrer, R., and Schauble, L. (2005). Developing modeling and argument in the elementary grades. In: Understanding Mathematics and Science Matters, ed. T. Romberg, T. Carpenter, and F. Dremock, Mahwah, N.J: Lawrence Erlbaum Associates, 29–54.

Lemke, J. L. (1990). Talking Science: Language, Learning, and Values. Language and Educational Processes, Norwood, N.J: Ablex Publishing Corp.

Lemke, J. L. (1995). Textual Politics: Discourse and Social Dynamics, London, England: Taylor and Francis.

Magnani, L., Nersessian, N. J., and Thagard, P. (eds.). (1999). Model-Based Reasoning in Scientific Discovery, New York: Kluwer Academic/Plenum Publishers.

Marzano, R. J. (2001). Designing a New Taxonomy of Educational Objectives, Thousand Oaks, CA: Corwin Press.

Mazur, E. (1996). Peer Instruction: A User's Manual, Upper Saddle River, NJ: Prentice Hall.

McDermott, L.C. (1991). Millikan Lecture 1990: what we teach and what is learned—closing the gap. Am. J. Physics *59*, 301–315.

National Commission on Excellence in Education (1983). A Nation at Risk: The Imperative for Educational Reform: Washington, DC: Government Printing Office.

National Research Council (NRC) (1996a). From Analysis to Action: Undergraduate Education in Science, Mathematics, Engineering, and Technology, Washington, DC: The National Academies Press. www.nap.edu/catalog.php?record_id=9128.

NRC (1996b). National Science Education Standards. Washington, DC: National Academies Press. www.nap.edu/catalog/4962.html, 274.

NRC (1999). Transforming Undergraduate Education in Science, Mathematics, Engineering, and Technology, Washington, DC: National Academies Press.

NRC (2001). Knowing What Students Know: The Science and Design of Educational Assessment, Washington, DC: National Academies Press. www.nap.edu/catalog/10019.html.

NRC (2003a). Improving Undergraduate Instruction in Science, Technology, Engineering, and Mathematics, Washington, DC: National Academies Press. www.nap.edu/catalog.php?record_id=10711.

NRC (2003b). BIO 2010, Transforming Undergraduate Education for Future Research Biologists, Washington, DC: National Academies Press. www.nap.edu/catalog.php?record_id=10497.

National Science Board (1986). Undergraduate Science, Mathematics and Engineering Education, Washington, DC.

National Science Board Commission on Precollege Education in Mathematics, Science, and Technology (1983). Educating Americans for the 21st Century: A Plan of Action for Improving Mathematics, Science and Technology Education for all American Elementary and Secondary Students so That Their Achievement is the Best in the World by 1995, Washington, DC: National Science Foundation.

National Science Foundation (1996). Shaping the Future: New Expectations for Undergraduate Education in Science, Mathematics, Engineering, and Technology: Arlington, VA. www.nsf.gov/pubs/stis1996/nsf96139/nsf96139.txt.

Nemirovsky, R., Rosebery, A. S., Solomon, J., and Warren, B. (eds.) (2005). Everyday Matters in Science and Mathematics: Studies of Complex Classroom Events, Mahwah, NJ: Lawrence J. Erlbaum Associates.

Pickering, A. (1995). The Mangle of Practice: Time, Agency, and Science, Chicago: University of Chicago Press.

Posner, G. J., Strike, K. A., Hewson, P. W., and Gertzog, W. A. (1982). Accommodation of a scientific conception: toward a theory of conceptual changes. Sci. Educ. *66*, 211–227.

Prigogine, I. (1961). Introduction to Thermodynamics of Irreversible Processes, New York: Interscience.

Prigogine, I., and Stengers, I. (1984). Order out of Chaos, New York: Bantam.

American Association for the Advancement of Science (1993). Project 2061 Benchmarks for Science Literacy, New York: Oxford University Press.

Redish, E. F. (2003). A theoretical framework for physics education research: modeling student thinking. Proceedings of the Enrico

Fermi Summer School in Physics, Course CLVI. Italian Physical Society.

Romesburg, H. C. (1984). Cluster Analysis for Researchers, Belmont, CA: Lifetime Learning Publications.

Rouse, J. (1987). Knowledge and Power: Toward a Political Philosophy of Science, Ithaca, NY: Cornell University Press.

Rutherford, F. J. (1990). Science for All Americans, New York: Oxford University Press.

Salthe, S. N. (1985). Evolving Hierarchical Systems: Their Structure and Representation, New York: Columbia University Press.

Schrödinger, E. (1967). What is Life? Mind and Matter, Cambridge, UK: Cambridge University Press.

Seymour, E., and Hewitt, N. M. (1997). Talking about Leaving: Why Undergraduates Leave the Sciences, Boulder, CO: Westview Press.

Smith, J. I., and Tanner, K. (2010). The problem of revealing how students think: concept inventories and beyond. CBE Life Sci. Educ. *9*, 1–5.

Smith, M. K., Wood, W. B., and Knight, J. K. (2008). The genetics concept assessment: a new concept inventory for gauging student understanding of genetics. CBE Life Sci. Educ. *7*, 422–430.

Strike, K. A., and Posner, G. J. (1992). A revisionist theory of conceptual change. In: Philosophy of Science, Cognitive Psychology and Educational Theory and Practice, ed. R. A. Duschl and R. J. Hamilton, Albany, NY: State University of New York Press, 147–176.

Tanner, K., and Allen, D. (2002). Approaches to cell biology teaching: a primer on standards. Cell Biol. Educ. *1*, 95–100.

Tanner, K., and Allen, D. (2005). Approaches to biology teaching and learning: understanding the wrong answers–teaching toward conceptual change. Cell Biol. Educ. *4*, 112–117.

Tobias, S. (1990). They're Not Dumb, They're Different: Stalking the Second Tier (Occasional paper on neglected problems in science education), Tucson, AZ: Research Corporation.

Treagust, D. F., Chittleborough, G., and Mamiala, T. L. (2002). Students' understanding of the role of scientific models in learning science. Intl. J. Sci. Educ. *24*, 357.

Wilson, C., Anderson, C. W., Merrill, J. E., Heidemann, M., Merritt, B. W., and Richmond, G. (in preparation). Principled reasoning, procedural display, and misconceptions in undergraduate students' accounts of cellular respiration.

Wilson, C. D., Anderson, C. W., Heidemann, M., Merrill, J. E., Merritt, B. W., Richmond, G., Sibley, D. F., and Parker, J. M. (2006). Assessing students' ability to trace matter in dynamic systems in cell biology. Cell Biol. Educ. *5*, 323–331.

Yorke, M. (2003). Formative assessment in higher education: moves towards theory and the enhancement of pedagogic practice. Higher Educ. *45*, 477–501.

Zhang, B., Patterson, R., Richmond, G., Parker, J. M., Merrill, J. E., and Urban-Lurain, M. (2005). Using self-response system and online learning environment in large college science classes—the technologies, instructional design, and implications. Paper presented at the International Conference on Computers in Education, Singapore.