CBE—Life Sciences Education
Vol. 9, 1–5, Spring 2010

Feature
Approaches to Biology Teaching and Learning

The Problem of Revealing How Students Think: Concept Inventories and Beyond

Julia I. Smith* and Kimberly Tanner†

*Department of Biological Science, Holy Names University, Oakland, CA 94619; and †Department of Biology, San Francisco State University, San Francisco, CA 94132

INTRODUCTION

A common refrain heard from many college and university biology instructors is that undergraduate science students do not seem to possess the same scientific habits of mind as their instructors, nor do they seem to have command of fundamental principles and concepts that structure the expertise of their instructors (Hestenes *et al*., 1992; Khodor *et al*., 2004; Wilson *et al*., 2006; Michael, 2007; D'Avanzo, 2008). In short, even our advanced undergraduate students do not seem to be scientifically literate—they cannot "ask and answer their own biologically relevant questions" (Wright, 2005). If we as university biology instructors are to make progress on the challenge of transforming our novice undergraduates into expert biological thinkers who are scientifically literate, then we all need tools that can aid us in revealing student thinking and in analyzing what we do in the classroom that supports or hinders the development of this scientific literacy in students. This is where classroom assessment—gathering evidence on students' thinking—is a key part of teaching at any level (Angelo and Cross, 1993; Atkin *et al*., 2001; Black and Wiliam, 1998; Huba and Freed, 2000; Sundberg, 2002; Tanner and Allen, 2004). However, there are a myriad of approaches to collecting assessment evidence from students: minute papers to gain quick insight into student thinking, reflective journal writing to promote metacognition and reveal confusions, and concept mapping to examine the structure of students' knowledge, to name just a few. Each of these many assessment approaches to monitor student thinking has its advantages and drawbacks, and some tools seem to work best for some topics or in the hands of some instructors. Here, we give an introduction to a relatively recent addition to the assessment tools in biology—the concept inventory—address its promising attributes and potential drawbacks, and raise the question of what concept inventories may actually measure. Finally, we consider potential alternative approaches to gaining insight into how students think about biology that come from the chemistry education and physics education research literatures.

DOI: 10.1187/cbe.09–12–0094
Address correspondence to: Kimberly Tanner (kdtanner@sfsu.edu).

THE PROMISE OF CONCEPT INVENTORIES

Over the past several years, concept inventories as a form of assessment have received particular attention in the biological sciences. Like many biology instructors, we have been curious about the nature of concept inventories, what they measure, and their potential use in both classroom settings and biology education research efforts. On initial consideration, concept inventories seem to be a powerful and accessible tool to support iterative improvement in faculty teaching and to enhance the scientific literacy of students. The stated goals of concept inventories have varied: to assess and build scientific literacy (Klymkowsky *et al*., 2003; Bowling *et al*., 2008a,b), to catalyze curriculum reform (Hake, 1998; Smith *et al*., 2008), and to identify student weak spots (Garvin-Doxas *et al*., 2007). In the simplest terms, a concept inventory is an outline of core knowledge and concepts for a given field and a collection of multiple-choice questions that are designed to probe student understanding of these fundamental concepts (Redish, 2000). Such inventories can potentially be used not only as a tool to provide information for instructors on how to improve teaching but also as an instrument to yield data for basic biology education research. Individual questions on a concept inventory are often born out of previous qualitative research using student interviews or open-ended essay questions that have revealed student misconceptions, incorrect thinking, or incomplete understanding regarding fundamental principles or concepts. These common misconceptions are then embedded into the choices of the multiple-choice questions as "distractors." The selection of a particular distractor by a student is intended to help instructors identify where a student is "stuck" in the mastery of a particular concept (Garvin-Doxas *et al*., 2007). Concept inventories seem to have been first developed as an instructional tool in the field of undergraduate physics, where they had a pivotal impact in advancing the field of physics education research. The force concept inventory (FCI), the first and most popular concept inventory to be developed (Hestenes *et al*., 1992), is a 29-question test focused on probing students' understanding of Newtonian and non-Newtonian concepts about force. The FCI, as it is commonly referred to, was designed to measure six conceptual dimensions of the force concept

considered essential for complete understanding (e.g., kinematics; kinds of forces; the superposition principle; and Newton's first, second, and third laws). With respect to classroom teaching, the FCI has been credited with catalyzing important reforms in undergraduate physics education, such as the development of Eric Mazur's model of peer instruction (Mazur, 1997). In addition, the FCI was key in nucleating research in physics education, such as Hake's study of normalized gain (Hake, 1998), which showed that student-learning gains were greater on the FCI with interactive pedagogy compared with traditional lecture alone. Following the success of the FCI in promoting pedagogical discussions and change, other concept inventories such as the force and motion conceptual evaluation (Thornton and Sokoloff, 1998) have been created in physics education, but none seems to have enjoyed the same widespread influence as the FCI.

Because of the significant impact of concept inventories in driving activity in physics education, biology educators have been motivated to develop concept inventories of their own, modeled on the FCI (Michael *et al*., 2008). In addition, the National Science Foundation has committed its support, providing grants for biology concept inventory development totaling >US$6 million since 1995 (www.nsf.gov/awardsearch). Unlike physics education, one of the persistent challenges in biology seems to be generating consensus among biologists about which concepts or "big ideas" are most important to assess (Garvin-Doxas *et al*., 2007; Michael *et al*., 2008). Today, there are a growing number of concept inventories and concept inventory–like assessments available for use by undergraduate biology faculty, including, but not limited to, the concept inventory in natural selection (Anderson *et al*., 2002), the biology concept inventory (Klymkowsky *et al*., 2003), the genetics concept inventory (Elrod, 2007), the genetics literacy assessment instrument (Bowling *et al*., 2008a,b), the genetics concept assessment (Smith *et al*., 2008), and numerous other inventories are under development (for overviews, see Garvin-Doxas *et al*., 2007; D'Avanzo, 2008; Michael *et al*., 2008). These inventories seem to offer numerous benefits. They encourage faculty to develop systematic classroom assessment techniques and to integrate assessment into their everyday teaching. They support faculty reflection on teaching using evidence collected systematically from students. And they help faculty gain some insight into what students may know and understand in terms of biology content knowledge. However, gaining insight into students' content knowledge or what they know is distinct from gaining insight into student thinking, their scientific literacy as defined above, and the extent to which they are accruing the scientific habits of mind that enable them to think like biologists.

THINKING CRITICALLY ABOUT CONCEPT INVENTORIES: WHAT DO THEY ACTUALLY MEASURE?

Even with their assured promise as an additional assessment tool in biology instructors' assessment toolkit, a critical analysis of concept inventories reveals several issues for consideration. Researchers in physics education have for many years deliberated upon whether concept inventories actually measure the conceptual understanding that they are designed to assess. The so-called "4-H" debate in physics education research—named for four authors of several key papers in the debate—highlights this issue (Heller and Huffman, 1995; Hestenes and Halloun, 1995; Huffman and Heller, 1995). Specifically, the debate has focused attention on whether the FCI is perhaps measuring student intuitions about questions in physics rather than a deeper conceptual understanding or way of knowing related to the six conceptual dimensions of the force concept. Hestenes and colleagues predicted that certain subsets of the questions on the FCI would map directly onto one of the six articulated dimensions of the force concept. Huffman and Heller contended that if this were true, then the questions on the test that theoretically measure a common dimension of the force concept should map mathematically onto a single factor when analyzed using factor analysis. However, just such a factor analysis of student responses on the FCI did not yield a robust mapping of test items onto their predicted conceptual dimension (Huffman and Heller, 1995). As a result, Huffman and Heller have proposed that rather than measuring conceptual understanding, the FCI may be more a measure of student familiarity with particular contexts. For example, students may be more familiar with questions about the physics of hockey pucks, and thus those questions group together in a factor analysis of students' responses because of familiarity with context rather than application or understanding of the deeper conceptual kinship. Similarly, questions regarding the same underlying conceptual dimension of the force concept presented in a less familiar context, such as rocket-related questions, do not necessarily cluster as predicted by test designers. Indeed, current research in physics education continues to explore, develop, and debate what the FCI actually measures and how to interpret it more than 17 years after its initial publication (Bao and Redish, 2001, 2006).

Another oft-stated goal of concept inventories is to measure student understanding that can inform instruction, thus making concept inventories a pedagogically useful assessment tool (Michael *et al*., 2008). However, two aspects of the construction of many (but not all) concept inventories may limit their usefulness as a classroom assessment tool: 1) the vocabulary used and 2) the format of these tests. First, a pervasive problem in concept inventories seems to be the use of jargon that obscures, rather than reveals, conceptual understanding. For example, a concept inventory question that probes students' understanding of the scientific method would seem to measure something about students' scientific habits of mind, their ways of thinking, and thus their scientific literacy. However, concept inventory questions such as this often require students to know the difference between a "positive control" and "negative control." Without a rather low level understanding of these vocabulary terms, a student would be unable to demonstrate his or her conceptual understanding of the logic inherent in the scientific method and the fair design of an experiment. As such, a student may possess a conceptual understanding of experimental design that would go unmeasured by such a vocabulary-limited concept inventory question. Second, the form of the concept inventory itself—a set of closed-ended, multiple-choice questions—would seem to be a significant limitation to gaining insight into students' thinking. A concept inventory gives a time-stamped measurement of student knowledge

before and after instruction. At best, comparison of pre- and postscores allows instructors to discern that something somehow affected student learning at some time, but not to determine what that influence was nor why it was influential. Also, by their nature, concept inventories force students to make a choice without opportunity for explanation as to why they have made that choice. As such, in their most common form, concept inventories would seem to yield little specific information about student thinking or deep conceptual understanding.

Given the debates about what concept inventories measure in the physics education community and the limitations that may be inherent in the present structure of concept inventories themselves, one wonders whether concept inventories really are measuring what they aspire to measure, deeper understanding and biological thinking. With regard to what concept inventories really measure, the heart of the issue is whether measurement of content knowledge can in some way serve as a proxy for conceptual understanding (Black, 2003). Yet, many would argue that ". . . no knowledge exists in any field, including biology, that is so essential that every literate person must know it." (Wright, 2005).

MOVING BEYOND CONCEPT INVENTORIES TOWARD MEASURING HOW STUDENTS THINK

That said, some concept inventory designers are now attempting to address many of the concerns expressed about the limitations of concept inventories based on multiple-choice questions, especially those concerns about measuring student thinking as opposed to student knowledge. One approach is a return to an assessment tool format that has been developed previously by science education researchers, namely, the "two-tier" multiple-choice instrument (for examples, see Tamir, 1971; Treagust, 1988). In these tests, questions are asked in pairs, where the first question in the pair is similar to a standard concept inventory question and the second question in the pair (the second tier) attempts to probe the likely reasoning behind the students' choice in the first question. This approach, although still predicated on using only multiple-choice questions, is an attempt to assess both students' content knowledge and the thinking and reasoning behind their answer choice.

In addition, the pioneering efforts in biology concept inventory development coupled with uneasiness about what is really being measured also seems to have stimulated the development of a related, but distinct approach: diagnostic question clusters. Diagnostic question clusters, also referred to as diagnostic question sets, do not necessarily rely only on closed-ended, multiple-choice questions. Instead, students are encouraged to approach biology questions of a variety of forms like a biological expert. Unlike concept inventories, these assessment tools seem to be specifically designed to help students improve their thinking and reasoning skills (Wilson *et al*., 2006; and Thinking Like a Biologist, www.biodqc.org/overview). Students are encouraged to develop the ability to move vertically through biological ideas from the intracellular to the organismal to the ecosystem level, thinking across scales and boundaries as needed; this is one salient characteristic of expert biological thinking.

Some concept inventory authors acknowledge that "open-ended responses, essay questions, and even lab tests provide greater insight into student knowledge, [but] they take too many resources (in terms of time and energy) to grade rigorously and objectively" (Klymkowsky *et al*., 2003). This argument assumes, however, that classroom assessment is solely in the service of the instructors, their evaluation of students, and improvement of their teaching. This stance, unfortunately, ignores one of the most important roles of classroom assessment: it is a powerful teaching tool. Assessments that challenge students to articulate their ideas—in writing, in drawing, in design of an experiment—are assessment approaches that integrate assessment into the learning process. Having students express their ideas pre- and postinstruction, and then reflect on changes in their thinking, puts assessment tools firmly in the service of learning, as tools that engage students metacognitively in fostering their own conceptual progress and growth. Fundamentally, it is what teachers and students do together within classrooms that has the potential to drive learning and progress toward scientific literacy (Black and Wiliam, 1998). Historically, Bloom (1956) taught us that if you want to develop higher-order reasoning, you have to ask higher-order questions (Allen and Tanner, 2007). If we keep testing for content knowledge because it is quick and easy, then that will limit the learning outcomes that we are able to achieve (Black, 2003). Reliance on content knowledge assessed by closed-ended, multiple-choice questions as a proxy for a more direct measure of student thinking and scientific literacy may cause instructors to overlook the very student outcomes that they hope to assess. Moving the focus of classroom assessment in biology beyond concept inventories (or other tests) that primarily assess knowledge or vocabulary or superficial aspects of the contexts being presented would seem critical to moving our students toward expert biological thinking.

ALTERNATIVE APPROACHES TO REVEALING HOW STUDENTS THINK: TALKING IN CHEMISTRY AND SORTING IN PHYSICS

The initial impetus for developing concept inventories in biology seems to have arisen from the influence of the FCI in undergraduate physics and its success in nucleating pockets of instructional reform (Klymkowsky *et al*., 2003). Examining education reform efforts and measurement tools in other science disciplines is indeed a fruitful approach. But perhaps we should not stop with concept inventories, but rather continue to explore the approaches used in chemistry and physics education to examine the impact of innovative instruction and to study the maturation of student thinking in those disciplines. Below, we review two studies that take novel approaches to measuring the maturation of student thinking in the sciences. Neither approach could easily be implemented in its original form in an undergraduate biology classroom. However, both focus squarely on the problem of measuring student thinking. And if that's the key goal of undergraduate biology education—maturation of novice biological thinkers into emerging biological expert thinkers who are biologically literate—then that is what we

need to figure out how to measure in biology, and adaptations of these approaches may offer a new way to do so.

In Their Own Words: Oral Interviews With Students

In a landmark study in chemistry education research, John Wright and his colleagues at the University of Wisconsin set out to investigate whether interactive, inquiry-based teaching in introductory undergraduate chemistry develops greater scientific competence in students compared with a traditional lecture approach (Wright *et al*., 1998). In developing a research design for their study, Wright and his colleagues decided to ask a group of skeptical science faculty what kind of evidence they would find compelling. These faculty skeptics called for a jury of their peers—external science faculty assessors—who would participate in a blind study of students' ability to think scientifically using oral interviews. Each external assessor was asked to develop his or her own oral exam and definition of student competence. Wright and his colleagues sampled students from the traditional lecture course and the active-learning infused course, such that multiple students from each octile, according to grade-based rank in the course, participated. Each faculty assessor conducted 30-minute interviews with students from both courses who were matched based on rank in their respective courses. After unblinding the results as to the course affiliation for each interviewed student, Wright and his team discovered that the independent examiners consistently ranked students from the interactive classroom higher in scientific competence compared with those from the traditional lecture classroom. A large proportion of the examiners in Wright's study identified meta-awareness—the thinking patterns of the students—as their primary criterion for judging competence. Wright and his colleagues argued that these oral discussions with students are far more effective in measuring changes in scientific thinking skills compared with standardized examinations. Oral interviews revealed students' abilities to think in original and fundamental ways. In addition, this approach enabled students to present a complete picture of their disciplinary knowledge and stance compared with a more delimited paper-and-pencil method. Indeed, Wright *et al*. (1998) contended that oral exams reflect the scientific maturity of the student and that written exams measure a student's command of the subject matter. They argued convincingly that the "habits of the mind" (e.g., thinking process, reasoning, and communication skills) inherent in successful problem solving are a key feature of students' scientific competence and are best assessed through oral discussions.

Gaining Insight into the Structure of Students' Thinking: The Problem-sorting Task

Although cognitive scientists have provided a variety of approaches to assess literacy or expertise across a range of disciplines, one study in physics education offers a unique approach to gauging the development of expertise and scientific thinking. In the early 1980s, Michelene Chi and her colleagues developed a sorting task designed for use in studying the development of expert thinking among trainees in physics (Chi *et al*., 1981). This task engaged participants in the categorization of physics problems taken from the end-of-chapter sections of a commonly used introductory undergraduate physics text. In their study, the researchers asked eight advanced physics doctoral students—classified as "experts"—and eight undergraduates who had completed an introductory course in mechanics—classified as "novices"—to sort 24 physics problems on the basis of similarity of solution. Their results strikingly revealed a distinct difference in the way that experts and novices, as defined in their study, sorted the same set of physics problems. Specifically, experts seemed to group problems on the basis of their underlying conceptual features (e.g., Newton's laws). In contrast, novices seemed to group problems on the basis of superficial, contextual features (e.g., blocks on inclined planes). Chi (2006) has argued that performance on contrived, structured tasks is a key tool in judging the nature of expertise among individuals at different stages of training within a discipline. In addition, Chi has suggested that particular tasks have the unique potential to reveal important information regarding the structure of an individual's knowledge, and subsequently their disciplinary thinking. In a study of problem solving, Smith and Good (1984) interviewed undergraduate students, graduate students, and biology instructors as they solved problems in classical genetics. Similar to the original physics study, their results suggested that novice and expert biologists exhibit differences in their representation of problems—novices focused attention on the commonality of "flower problems," whereas experts reflected on "monohybrid problems." These results suggest differences in the underlying structure of disciplinary knowledge and ways of thinking between novices and experts that are similar to the findings in physics. Studies of the transitions in thinking and knowledge structure that occur as novices mature into experts is ongoing, and the early studies of Chi and colleagues continue to influence the field.

CONCLUSIONS

In summary, attempts at constructing concept inventories in biology have sparked many new discussions about what our goals for undergraduate biology education really are and how we can measure these things. Although concept inventories are most certainly a welcome addition to the biology instructors' varied assessment toolkit, it is unclear whether they really measure what we aspire to cultivate in our students, namely, scientific habits of mind, biological literacy, and the ability to think independently like a biologist. Concept inventories have limitations: what they measure is not always clear, conceptual understanding can be obscured by jargon, and a reliance on closed-ended, multiple-choice questions necessitates that they primarily assess content knowledge rather than conceptual understanding, biological thinking, and scientific literacy. Achieving robust assessment of genuine conceptual understanding and biological thinking will require us to not only clarify what we want to endure within students' minds long after their undergraduate biology education but also determine the extent to which our assessment tools actually measure these things.

REFERENCES

Allen, D., and Tanner, K. (2007). Putting the horse back in front of the cart: using visions and decisions about high-quality learning experiences to drive course design. CBE Life Sci. Educ. *6*, 85–89.

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.

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

Atkin, J. M., Black, P., and Coffey, J. E. (ed.) (2001). Assessment and the National Science Education Standards, Washington, DC: Center for Education, National Research Council.

Bao, L., and Redish, E. F. (2001). Concentration analysis: a quantitative assessment of student states. Phys. Educ. Res. Am. J. Phys. Suppl. *69*, S45–S53.

Bao, L., and Redish, E. F. (2006). Model analysis: representing and assessing the dynamics of student learning. Phys. Rev. Spec. Top. Phys. Educ. Res. 2, 010103.

Black, P. (2003). The importance of everyday assessment. In: Everyday Assessment in the Science Classroom, ed. J. Myron Atkin and Janet Coffey, Arlington, VA: National Science Teachers Association Press, 1–11.

Black, P., and Wiliam, D. (1998). Inside the black box: raising standards through classroom assessment. Phi Delta Kappan *80*, 139–148.

Bloom, B. S. (ed.) (1956). Taxonomy of Educational Objectives: Classification of Educational Goals, Handbook I: Cognitive Domain, New York: Longman.

Bowling, B. V., Acra, E. E., Wang, L., Myers, M. F., Dean, G. E., Markle, G. C., Moskalik, C. L., and Huether, C. A. (2008a). Development and evaluation of a genetics literacy assessment instrument for undergraduates. Genetics *178*, 15–22.

Bowling, B. V., Huether, C. A., Wang, L., Myers, M. F., Markle, G. C., Dean, G. E., Acra, E. E., Wray, F. P., and Jacob, G. A. (2008b). Genetic literacy of undergraduate non-science majors and the impact of introductory biology and genetic courses. Bioscience *58*, 654–660.

Chi, M.T.H. (2006). Laboratory Methods for Assessing Experts' and Novices' Knowledge. In: The Cambridge Handbook of Expertise and Expert Performance, ed. K. A. Ericsson, N. Charness, P. J. Feltovich, and R. R. Hoffman, Cambridge, United Kingdom: Cambridge University Press, 167–184.

Chi, M.T.H., Feltovich, P. J., and Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cogn. Sci. *5*, 121–152.

D'Avanzo, C. (2008). Biology concept inventories: overview, status, and next steps. Bioscience *58*, 1079–1085.

Elrod, S. (2007). Bioliteracy. http://bioliteracy.net/Readings/papersSubmittedPDF/Elrod.pdf (accessed 5 December 2009).

Garvin-Doxas, K., Klymkowsky, M., and Elrod, S. (2007). Building, using, and maximizing the impact of concept inventories in the biological sciences: report on a National Science Foundation-sponsored conference on the construction of concept inventories in the biological sciences. CBE Life Sci. Educ. *6*, 277–282.

Hake, R. R. (1998). Interactive-engagement vs traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses. Am. J. Phys. *66*, 64–74.

Heller, P., and Huffman, D. (1995). Interpreting the force concept inventory: a reply to Hestenes and Halloun. Phys. Teach. *33*, 503–511.

Hestenes, D., and Halloun, I. (1995). Interpreting the force concept inventory: a response to March 1995 critique by Huffman and Heller. Phys. Teach. *33*, 502–506.

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

Huba, M. E., and Freed, J. E. (2000). Learner-Centered Assessment on College Campuses, Needham Heights, MA: Allyn and Bacon.

Huffman, D., and Heller, P. (1995). What does the force concept inventory actually measure? Phys. Teach. *33*, 138–143.

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., Garvin-Doxas, K., and Zeilik, M. (2003). Bioliteracy and teaching efficacy: what biologists can learn from physicists. Cell Biol. Educ. *2*, 155–161.

Mazur, E. (1997). Peer Instruction: A User's Manual, Upper Saddle River, NJ: Prentice Hall.

Michael, J. (2007). Conceptual assessment in the biological sciences: a National Science Foundation-sponsored workshop. Adv. Physiol. Educ. *31*, 389–391.

Michael, J., McFarland, J., and Wright, A. (2008). The second conceptual assessment in the biological sciences workshop. Adv. Physiol. Educ. *32*, 248–251.

Redish, E. F. (2000). Discipline-based education and education research: the case of physics. J. Appl. Dev. Psychol. *21*, 85–96.

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.

Smith, M. U., and Good, R. (1984). Problem solving and classical genetics: successful versus unsuccessful performance. J. Res. Sci. Teach. *21*, 895–912.

Sundberg, M. (2002). Assessing student learning. Cell Biol. Educ. *1*, 11–15.

Tamir, P. (1971). An alternative approach to the construction of multiple-choice test items. J. Biol. Educ. *5*, 305–307.

Tanner, K. D., and Allen, D. E. (2004). From assays to assessments—on collecting evidence in science teaching. Cell Biol. Educ. *3*, 69–74.

Thornton, R. K., and Sokoloff, D. R. (1998). Assessing student learning of Newton's laws: the force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula. Am. J. Phys. *66*, 338–352.

Treagust, D. F. (1988) The development and use of diagnostic instruments to evaluate students' misconceptions in science. Int. J. Sci. Educ. *10*, 159–169.

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. CBE Life Sci. Educ. *5*, 323–331.

Wright, J. C., Millar, S. B., Kosciuk, S. A., Penberthy, D. L., Williams, P. H., and Wampold, B. E. (1998). A novel strategy for assessing the effects of curriculum reform on student competence. J. Chem. Educ. *75*, 986–992.

Wright, R. L. (2005). Undergraduate biology courses for nonscientists: toward a lived curriculum. Cell Biol. Educ. *4*, 189–198.