ASBMB-RCN Workshop: Assessment of Students’ Reasoning with Core Concepts and Visualizations in Biochemistry

Purdue University
Saturday, November 3, 2012
9:00 AM – 5:00 PM
9 Principles of Good Practice for Assessing Student Learning

by
Alexander W. Astin; Trudy W. Banta; K. Patricia Cross; Elaine El-Khawas; Peter T. Ewell; Pat Hutchings; Theodore J. Marchese; Kay M. McClennen; Marcia Mentkowski; Margaret A. Miller; E. Thomas Moran; Barbara D. Wright

This document was developed under the auspices of the AAHE Assessment Forum with support from the Fund for the Improvement of Postsecondary Education with additional support for publication and dissemination from the Exxon Education Foundation. Copies may be made without restriction.

1. **The assessment of student learning begins with educational values.**
   Assessment is not an end in itself but a vehicle for educational improvement. Its effective practice, then, begins with and enacts a vision of the kinds of learning we most value for students and strive to help them achieve. Educational values should drive not only what we choose to assess but also how we do so. Where questions about educational mission and values are skipped over, assessment threatens to be an exercise in measuring what's easy, rather than a process of improving what we really care about.

2. **Assessment is most effective when it reflects an understanding of learning as multidimensional, integrated, and revealed in performance over time.**
   Learning is a complex process. It entails not only what students know but what they can do with what they know; it involves not only knowledge and abilities but values, attitudes, and habits of mind that affect both academic success and performance beyond the classroom. Assessment should reflect these understandings by employing a diverse array of methods, including those that call for actual performance, using them over time so as to reveal change, growth, and increasing degrees of integration. Such an approach aims for a more complete and accurate picture of learning, and therefore firmer bases for improving our students' educational experience.

3. **Assessment works best when the programs it seeks to improve have clear, explicitly stated purposes.**
   Assessment is a goal-oriented process. It entails comparing educational performance with educational purposes and expectations -- those derived from the institution's mission, from faculty intentions in program and course design, and from knowledge of students' own goals. Where program purposes lack specificity or agreement, assessment as a process pushes a campus toward clarity about where to aim and what standards to apply; assessment also prompts attention to where and how program goals will be taught and learned. Clear, shared, implementable goals are the cornerstone for assessment that is focused and useful.

4. **Assessment requires attention to outcomes but also and equally to the experiences that lead to those outcomes.**
   Information about outcomes is of high importance; where students "end up" matters greatly. But to improve outcomes, we need to know about student experience along the way -- about the curricula, teaching, and kind of student effort that lead to particular outcomes. Assessment
can help us understand which students learn best under what conditions; with such knowledge comes the capacity to improve the whole of their learning.

5. **Assessment works best when it is ongoing not episodic.** Assessment is a process whose power is cumulative. Though isolated, "one-shot" assessment can be better than none, improvement is best fostered when assessment entails a linked series of activities undertaken over time. This may mean tracking the process of individual students, or of cohorts of students; it may mean collecting the same examples of student performance or using the same instrument semester after semester. The point is to monitor progress toward intended goals in a spirit of continuous improvement. Along the way, the assessment process itself should be evaluated and refined in light of emerging insights.

6. **Assessment fosters wider improvement when representatives from across the educational community are involved.** Student learning is a campus-wide responsibility, and assessment is a way of enacting that responsibility. Thus, while assessment efforts may start small, the aim over time is to involve people from across the educational community. Faculty play an especially important role, but assessment's questions can't be fully addressed without participation by student-affairs educators, librarians, administrators, and students. Assessment may also involve individuals from beyond the campus (alumni/ae, trustees, employers) whose experience can enrich the sense of appropriate aims and standards for learning. Thus understood, assessment is not a task for small groups of experts but a collaborative activity; its aim is wider, better-informed attention to student learning by all parties with a stake in its improvement.

7. **Assessment makes a difference when it begins with issues of use and illuminates questions that people really care about.** Assessment recognizes the value of information in the process of improvement. But to be useful, information must be connected to issues or questions that people really care about. This implies assessment approaches that produce evidence that relevant parties will find credible, suggestive, and applicable to decisions that need to be made. It means thinking in advance about how the information will be used, and by whom. The point of assessment is not to gather data and return "results"; it is a process that starts with the questions of decision-makers, that involves them in the gathering and interpreting of data, and that informs and helps guide continuous improvement.

8. **Assessment is most likely to lead to improvement when it is part of a larger set of conditions that promote change.** Assessment alone changes little. Its greatest contribution comes on campuses where the quality of teaching and learning is visibly valued and worked at. On such campuses, the push to improve educational performance is a visible and primary goal of leadership; improving the quality of undergraduate education is central to the institution's planning, budgeting, and personnel decisions. On such campuses, information about learning outcomes is seen as an integral part of decision making, and avidly sought.
9. **Through assessment, educators meet responsibilities to students and to the public.** There is a compelling public stake in education. As educators, we have a responsibility to the publics that support or depend on us to provide information about the ways in which our students meet goals and expectations. But that responsibility goes beyond the reporting of such information; our deeper obligation -- to ourselves, our students, and society -- is to improve. Those to whom educators are accountable have a corresponding obligation to support such attempts at improvement.

To request more information about the Assessment Forum, contact:

One Dupont Circle, Suite 360 Washington DC 20036-1143
Phone 202/293-6440 Fax 202/293-0073 info@aahe.org
© 2003 American Association for Higher Education
Student assessment, in all its forms, is central to the educational process. In this paper in the series, “Bridging the Gap”, I describe how assessment can be used as a powerful instrument for influencing how and what students learn and how and what instructors teach in a manner that is conducive to educationally sound curriculum change and improvement. In this way assessment is seen as a useful strategy for colleagues interested in bridging the gap between educational research and its application in teaching practice.

Keywords: Summative, formative assessment, teaching, learning, curriculum, biochemistry.

I have chosen to launch this column with a series of articles on assessment because assessment is of central importance in education and is arguably the most powerful tool that educators have at their disposal to influence and improve the teaching process [1]. Since the terms “assessment” and “evaluation” are used rather loosely at times and, in some cases, differently in different countries, it is appropriate to commence by defining them. For the purpose of this paper, I will use the definitions of Scriven [2] and Taras [3] who refer to assessment as, “judgements of students’ work,” and evaluation as, “judgements regarding courses or course delivery, or the process of making such judgements.”

This article, by way of introduction to assessment, will focus on the power of assessment, not only as a tool for grading students, but also as an effective instrument that can be used by educators to inform, control, and enhance the teaching and learning process for the benefit of all stakeholders. In the next article, I will focus on the meaning and assessment of conceptual understanding and the often neglected cognitive skills that are so essential for the development of expert biochemists and molecular biologists. This will then be followed by an article on assessing visual literacy and visualization skills, a topic of extreme importance to modern biochemists, whose success relies very heavily on visual tools to represent the microscopic and molecular world they investigate [4]. Finally, the last article on assessment will expose readers to a variety of useful tools that can be easily used to evaluate the quality of assessment tasks so that they meet required standards and are valid measures of students’ achievement of course objectives. In all these articles, I shall make use of selected examples from biochemistry and molecular biology to illustrate the main theories and principles presented. In this way, I hope to facilitate the implementation of these ideas by biochemistry and molecular biology educators and go some way towards promoting the “bridging of the gap”-goals of this column.

Fig. 1 outlines the crucial relationship that exists between the four key components of the educational process, namely course objectives, teaching, learning, and assessment. This paper will focus mainly on the details pertaining to the assessment component while at the same time emphasizing the important relationship between all the four components. Thus, the figure is not only recommended as a guiding framework for this paper but is also intended to serve as a useful diagram for facilitating discussions among faculty and students about the relationship between the components and how they affect or inform each other. The diagram occupies a permanent place on our notice board and is quite often used to stimulate educational debate.

SUMMATIVE ASSESSMENT: MEASURING (GRADING) ACHIEVEMENT OF COURSE OBJECTIVES

Summative assessment has been described by Taras [3] as “a judgement which encapsulates all the evidence up to a given point.” When that point is at the end of the course, summative assessment becomes a powerful tool for measuring or grading whether a student has achieved all the course objectives (see Fig. 1). In designing a course, it is important to first establish the course objectives, in terms of the knowledge and skills you would like your students to acquire. As illustrated in...
Fig. 1. Relationship between objectives, teaching, learning, and assessment. The single-headed solid arrows indicate “...should determine...,” the double-headed solid arrows indicate “...should determine...” or “...enables/measures achievement of...” in the downward or upward directions, respectively; and, the double-headed, broken arrows indicate “...should be in line with...”

Fig. 1, this should directly inform what you teach and, therefore, what you assess. It is important to always check that what you teach will enable students to achieve the course objectives, and what you assess is indeed measuring students’ achievement of the specified objectives. In my experience, educators often think they are doing this but find otherwise when they take a closer look at their course notes and compare these to the stated course objectives and their test and examination questions set during the course. For example, one objective of a course in metabolism might be to develop students’ ability to apply their knowledge of the principles of metabolism to other pathways that they have not encountered before. However, on examination of the coursework and the summative assessment, it becomes clear that the students had not been given any tasks involving the interpretation of “unseen” pathways. On further investigation, it emerges that the instructor had assumed that knowledge of the principles implies automatic ability to transfer and apply such knowledge to other situations, something that science education research has shown to not necessarily be true. Indeed, researchers such as Solomon and Perkins [5] as well as Meyer [6] have shown that transfer skills need to be explicitly taught through extensive problem-solving exercises and teaching strategies such as the cuing of students to transfer their knowledge from one context to another. How to assess and promote the development of transfer skills in students will be one of the focuses of another. How to assess and promote the development of explicit teaching through extensive problem-solving exercises and teaching strategies such as the cueing of students to transfer their knowledge from one context to another. How to assess and promote the development of transfer skills in students will be one of the focuses of another.

The grades obtained by students through summative assessment can also be used to evaluate the effectiveness of the instructor and of the course in achieving the desired objectives and, therefore, whether there will be any need for course and/or instructor improvement in the future. This assumes of course that the assessment used is fair and valid as, for example, the use of very easy questions can create a perception that the instructor is far more effective than he or she actually is.

When summative assessment is performed on a regular basis throughout the course, it is often called continuous assessment. Continuous assessment allows students frequent measures of student performance (grades) and their progressive achievement of course objectives. This allows students the opportunity to monitor their progress and, if necessary, to try and improve their grades before the course ends. Continuous assessment also affords instructors the opportunity to phase in new types of questions so that students can get repetitive practice at answering such questions and the opportunity to develop new problem-solving strategies before they encounter them in examinations. This helps to progressively build student confidence and competence during the course and better prepares them for the final examination. Continuous assessment need not be followed by any action taken by faculty to improve student learning based on feedback from student answers. If it does, then this constitutes a situation in which continuous assessment has been used in combination with formative assessment (see also under formative assessment section).

There is a current worldwide move to use a continuous assessment process, with no final examination. Instead, students progressively accumulate grade points during the course by doing regular quizzes, tests, and assignments, which are designed to keep them studying consistently and well informed as to the progress of their learning. A portfolio of all assessment tasks done by each student during the course is supplied to the examiner, who is then in a much better position, than in the case of a single examination, to obtain a more accurate and fairer measure of a student’s performance and knowledge. Indeed, the science education literature contains extensive criticism of summative-type final examinations that count towards a high percentage of the final grade (e.g., [7]). This is because a student’s performance on the day of the final examination might be affected adversely by factors such as stress-induced mental blocks, off days, or feeling unwell. In addition, the sudden-death nature of examinations gives students no chance to recover and improve and no time for the instructor to help them after they have left the course. And, as stated above, students are discouraged from working consistently during the course and encouraged to cram at the last moment and to spot and develop strategies to pass that are not in the interests of promoting sound learning.

All types of summative assessment have important social implications, particularly in the case of new graduates who do not have the benefit of any relevant work experience and related letters of recommendation to present to potential employers. In this case, without some summative measure of an individual’s capability or competence, we would have no way of determining their potential for a particular job or career. Thus, in such cases, as discussed by Howson [8], society tends to be more interested in the grades students obtain through summative assessment than the learning and assessment process by which the grades were achieved.

FORMATIVE ASSESSMENT: MONITORING AND IMPROVING STUDENT LEARNING

Consideration of the process of teaching and learning brings us to formative assessment. Black and coworkers have published extensively on the nature and value of
formative assessment (e.g., [9–11]). They describe formative assessment as all the activities undertaken by instructors during the course that yield feedback information on teaching and learning and, most importantly, in which such information is explicitly used to adapt the teaching to improve and promote student learning before the course ends [9, 10] (Fig. 1). Thus, formative assessment is not the same as continuous assessment (see above), which simply involves frequent summative assessment plus feedback on student grades, without any attempt to use feedback information to render significant improvements in student learning during the course. A combination of formative and continuous assessment is seen by various educational researchers as providing a more complete teaching and learning experience for students.

There is extensive evidence (e.g., [9]) that formative assessment can raise standards of learning and develop faculty’s professional practice in a way that they value. A key factor is the quality of the feedback information yielded by the assessment tasks. Good feedback should indicate the existence of any “gap” between the actual level of a student’s work being assessed and the required standard [12]. In this way, students can easily identify whether they need to improve in certain areas, while instructors can decide whether they need to adjust their teaching approach, spend longer on a particular aspect, or give certain students extra tutorials. In addition, formative assessment affords faculty the opportunity to monitor and, if necessary, improve the quality and the standard of the assessment tasks they are giving to students during the course. Through analysis of student answers, any questions that prove either too difficult, too easy, unfair, unclear, or misleading can be restructured or replaced with better questions. Article 5 of this series on Bridging the Gap will focus more specifically on methods such as item analysis that can be used to screen the quality of questions. Thus, in this way, formative assessment can afford biochemistry students and instructors the opportunity to identify problems that students may be experiencing at an early stage, before the examinations arrive and it becomes too late to correct them.

A fairly recent innovation, which considerably enhances the goals of formative assessment, has been the development of various personal response systems (PRSs), more commonly called “clickers” (e.g., [13, 14]). Such systems, which are being increasingly used worldwide by biochemistry instructors, considerably enhance the active participation of students in learning activities by requiring them to respond to multiple-choice questions (MCQs) in real time during class. Their responses are then instantly recorded and processed on the instructor’s computer, allowing for immediate student feedback as to what percentage of students selected each option. This affords the students and the instructor the opportunity to actively debate and critically evaluate the various options presented and to justify their particular choices with sound arguments. This not only helps students develop the interactive and critical thinking skills that are so important for their development as scientists, but also allows them to address any conceptual or reasoning difficulties before they become firmly ingrained in a student’s cognitive structure and, therefore, resistant to correction [15]. In addition, Burnstein and Lederman [16] showed that the use of a PRS increased student attendance, kept them more alert during class, and improved their overall performance and motivation for learning.

The science education literature contains numerous reports (e.g., [17]) in which educators formalize this monitoring/feedback/action process by doing research with student response data obtained from formative assessment [11]. The research outcomes can then, for instance, be used to inform improvement of teaching and assessment practice, the remediation of student conceptual difficulties, and curriculum change. This constitutes one form of a much larger area of research that is commonly called, action research [18] and is a much more rigorous way of monitoring the quality of one’s teaching practice. This approach is recommended to biochemistry educators as a simple, effective tool for investigating an educational problem, and taking action to resolve it. Participants choose a problem to focus on (e.g., Why are some students producing such poor concept maps in my metabolism class?), gather data, reflect on and share their findings, plan for action, carry it out, check their results, and plan for further action until the problem has been resolved. This cycle of steps—act, observe, reflect, plan—can be used more broadly to continually develop and improve the quality of course curricula and teaching programs, something that has now become a basic requirement at most universities. The publication of action research in a journal widely read by research scientists would be of value to the whole biochemistry and molecular biology teaching community. In a future article of this column, I shall discuss some of the many other applications of action research to the educational process and introduce readers to quantitative and qualitative educational research methods that can be used to rigorously evaluate the quality of their teaching and student learning.

Formative assessment can, through regular monitoring of student progress and the giving of feedback, also act as a powerful and effective instrument for promoting learning (studying) during the course (Fig. 1). This is because the more often you assess students, the more they will have to think and learn about their subject and the more often they will be motivated to get down and learn in anticipation of the next assessment session. Thus, formative assessment encourages students to work consistently throughout the course, rather than cramming at the last moment, which is so typical of the period before examinations.

ASSESSMENT-DRIVEN STUDENT LEARNING: LEARNING TO THE TEST

Assessment is an extremely powerful instrument for dictating to students what and how they should learn for tests (see Fig. 1). This helps them develop more effective

---

1 The abbreviations used are: PRS, personal response system; MCQ, multiple-choice question; SFES, science faculty with education specialties; PCK, pedagogical content knowledge; AERA, American Educational Research Association.
learning approaches. Most students will not only use lecture notes supplied by the instructors to inform what they learn but will particularly base what they learn on what knowledge and skills past tests and tutorials have focused on. Furthermore, how such knowledge and skills are assessed (i.e., the type of questions used) will also inform students how to learn (Fig. 1). For instance, if essays are commonly used, students will practice writing essays, while if MCQs dominate tests then students will practice doing these types of questions. Thus, this approach used by students, commonly called “spotting” or learning to the test, offers instructors a powerful tool for directing student learning as they see fit. Whereas it is normally very difficult to pressure students into changing their learning styles, manipulation of the nature of the assessment task to dictate a particular learning approach, offers faculty an easier means of motivating students to change their learning styles so that they can satisfactorily master the question.

As long ago as 1983, Entwistle and Ramsden [19] described two main learning approaches that students tend to use, depending on the type of questions given to them (i.e., how we assess and what students learn will both influence how they learn; Fig. 1). If the assessment simply requires the mindless regurgitation of memorized facts, students will probably use what Entwistle and Ramsden [19] call a surface approach to learning science. In a surface approach, they suggest that the student’s intention is to master the examination system rather than learn as much as possible; to complete the task requirements; memorize the information needed for assessments; treat the task as something not interesting or challenging; focus on separate facts without integration of knowledge; and, not think about why (purpose) and how (strategy) the topic should be studied. For example, during my undergraduate years, I was given the following type of question that encouraged me to adopt a surface approach to learning about metabolism.

**Question 1**

“Give details of the glycolytic pathway, with respect to the chemical structures of all intermediates and the names of relevant enzymes, coenzymes, and cofactors.”

Such a question clearly encourages the memorization of facts and does not require the students to demonstrate their understanding of the functioning of a metabolic pathway. In contrast, I have used the following type of question in my metabolism course [20], which requires extensive understanding of the functioning of a pathway, including transfer and application of knowledge of thermodynamics and kinetics, to solve the problem.

**Question 2**

The discovery that glyceraldehyde-3-phosphate dehydrogenase is irreversibly inhibited by iodoacetaete was important in the history of research on glycolysis. Explain the effect of this inhibitor in muscle on:

a. The overall flux through glycolysis;

b. The relative concentrations of intermediates of the glycolytic pathway;

c. The net production/utilization of ATP by glycolysis and the pathway efficiency;

d. The synthesis of triacylglycerol;

e. If iodoacetate was able to inhibit lactate dehydrogenase instead of glyceraldehyde-3-phosphate dehydrogenase, would the muscle be able to generate ATP by glycolysis at a high enough rate for strenuous physical activity? Explain.

To minimize the need for students to memorize information, I would also allow students access to their textbook when answering this question, or supply them with a diagram of the glycolytic pathway details, that is, all the information that they would have had to supply if answering Question 1 above. Question 2 promotes what Entwistle and Ramsden [19] call a deep approach, rather than a surface approach to learning science, since such types of assessment require more mindful understanding of concepts and the ability to integrate and apply knowledge to novel situations. Good conceptual questions also allow for a range of scientifically correct answers, rather than a single answer, so that the students are encouraged to be more creative and give their own answers, rather than the answer they think the lecturer might expect. This helps students feel confident that their thoughts and ideas are welcomed and will be judged fairly.

MCQs, such as those presented by Szeberényi [21], can also foster a deep approach by students to learning. However, such questions also allow students to adopt a surface approach to problem solving and to obtain the correct answer without understanding why it is correct. For example, when the student selects the correct option in an MCQ, how does the instructor know whether the student actually understood the whole problem and the relevant science, or whether he or she had simply guessed the answer or rote learnt enough facts to be able to select the correct answer without understanding the science? One way to minimize such problems is to always ask students to give an explanation for the choice they made. This encourages the student to think more deeply about the question, while the instructor can then easily establish whether a correct choice is backed up by a sound explanation and, therefore, deep conceptual understanding of the problem. Question 3 below, is an example of such a question that, like question 2, also probes an aspect of student understanding of the functioning of metabolism.

**Question 3**

Consider the following part of the glycolytic pathway functioning in a cell.

![Glycolytic Pathway Diagram](Image)

If 6-phosphofructokinase is totally and irreversibly inhibited by a toxic substance:
i. The reactions AFTER the inhibited 6-phosphofructokinase reaction will:
   Speed up/continue at the same rate/slow down/stop (Circle the correct answer)
   Because . . .

ii. The reactions BEFORE the inhibited 6-phosphofructokinase reaction will:
   Speed up/continue at the same rate/slow down/stop (Circle the correct answer)
   Because . . .

iii. The overall flux through glycolysis will:
   Stay the same/increase/decrease/be zero (Circle the correct answer)
   Because . . .

Thus, in summary, in a deep learning approach, the student's intention is to understand, question, critique, and debate the course content; relate new ideas to previous knowledge; relate concepts to everyday experience; relate evidence to conclusions; examine the logic of the argument; and apply knowledge to solving problems. Since many students have a background of rote learning from their earlier education, the sudden expectation that students should think and reason out cognitively demanding problems can be a stressful experience for them [20]. Therefore, care should be taken to gradually introduce students to assessment that requires a deep approach to learning and to ensure that they are given plenty of practice at answering such questions. It is also important to not continually reuse the same “deep” questions from one year to the next as, in my experience, many students will not continually reuse the same “deep” questions from one year to the next. In fact, this was one of the approaches successfully used by numerous medical schools worldwide when they introduced problem-based and theme-based learning into their curricula. To ensure greater success in such ventures, various scientists and educational researchers (e.g., [22]) have recommended that departments consider hiring Science Faculty with Education Specialties (SFES), into science departments who can assist in all areas of the educational process and facilitate ongoing curriculum development, backed up by sound discipline-based educational research. This issue was extensively discussed in paper 1 of this column [23] and has been strongly supported by many faculty as an indispensable strategy for facilitating the bridging of the research-teaching practice gap.

Although assessment-driven curriculum change, or in its simplest form “teaching to the test,” is considered by many science educators as undesirable, there is no doubt that the questions often given at the end of textbook chapters, or offered as textbook supplements, can and do, have a profound influence on what faculty teach. Thus, in an informal way, assessment-driven change is happening all the time, so why not formalize it as a “bridging-the-gap” activity?

**PEDAGOGICAL CONTENT KNOWLEDGE (PCK): WHAT WE TEACH AND ASSESS DETERMINES HOW WE TEACH AND ASSESS**

The relationship between the “what’s” and the “how’s” of teaching, learning, and assessment (Fig. 1) has been the target of extensive science education research since Lee Shulman, of Stanford University, first coined the term pedagogical content knowledge (PCK) in his 1985 Presidential address at the American Educational Research Association (AERA). In his seminal paper [24], he suggests that faculty not only need science content knowledge and pedagogical knowledge to be competent instructors but also a third type of knowledge, which he termed PCK. To distinguish between the three types of knowledge, he describes content knowledge as knowledge of scientific and mathematical principles, laws, theories, and concepts; thinking, reasoning, critical, analytical, synthetic, and process (i.e., cognitive) skills; practical and technical skills; and attitudes and values. Pedagogical knowledge is described as knowledge of teaching and assessment techniques; curriculum material development; student learning and theories of learning; and management of learning environments. Finally, he defines PCK as knowledge of the specific teaching and assessment approaches that are most appropriate for each topic, principle, concept, or skill. In other words, how you teach and assess should depend on what you are teaching and assessing (Fig. 1), as different content areas, topics, or skills need a different pedagogical approach. More simply, it is the additional knowledge that new lecturers and even tutors realize they need when teaching subject matter for the first time—how to communicate it in the most appropriate manner for optimal learning.

According to Shulman and colleagues, PCK includes knowledge of the most useful and appropriate ways of representing (e.g., verbal descriptions, diagrams, simulations, animations, equations, and graphs), teaching, and assessing each individual topic or concept and skill, so as to make it optimally comprehensible and understand-
able to students. It also includes knowledge of what makes specific topics, concepts, and skills easy or difficult for students to learn. This, in turn, includes knowledge of students' prior knowledge or misconceptions that they bring with them from different educational backgrounds; students' alternative conceptions and incorrect ways of reasoning; and remediation strategies or instructional conditions that would most likely correct all these difficulties.

Clearly, all biochemistry faculty possess some PCK, particularly concerning the topics they are most experienced in teaching, and that their PCK will continually develop throughout their teaching careers. The big question is whether their PCK of a particular topic requires improvement and, if so, how can they go about improving their PCK? Since few, if any, biochemistry textbooks or accompanying instructor's manuals explicitly describe the PCK that faculty could implement when teaching each topic, instructors need to find alternative ways of developing their PCK. One way is to gather a group of colleagues together, who are all experienced in teaching the topic of interest, and brainstorm what the optimal way of teaching that topic is in terms of the best teaching and assessment approaches, including the most useful representations and strategies for preventing or remediating student difficulties. For example, regarding the topic of metabolism, as discussed above, I would recommend teaching and assessment approaches that minimize the need for rote learning chemical structures and rather emphasize understanding of the functioning of pathways. Other components of my PCK regarding metabolism would, for example, include the use of animations to illustrate chemical and cellular processes; the use of thermodynamics and kinetics principles and calculations to explain metabolic regulation, reaction direction, rate, etc; the use of concept maps [25] to develop students' integrated knowledge of metabolic concepts; the use of examples of enzyme inhibitors or gene mutations to understand how pathways respond to stress; and the use of unseen pathways and biotechnology processes to develop students' ability to transfer and apply their knowledge of metabolic principles to novel situations. Considering my past research on student difficulties with metabolism [see for example [20] and [26]], I would also present to colleagues how I believe such difficulties could be prevented or remediated during teaching.

In summary, Shulman and Sherin [27] stated that, "...one of the most significant factors influencing the effectiveness of teaching is the teacher's own subject matter knowledge and pedagogical content knowledge." In presenting some examples of PCK, Bucat [28] reinforces this opinion by emphasizing the importance and urgency of doing some education research that probes and documents the topic-specific PCK of respected chemistry faculty. Since, as already discussed, what and how instructors assess needs to be in line with what and how they teach, assessment is not only an integral part of PCK but can be used as a powerful instrument to promote the development of PCK in faculty [29]. Thus, it is clearly of utmost importance that biochemistry educators also develop and document their own PCK for each biochemistry topic, concept, or skill that they teach and that assessment can be used as a powerful tool for promoting this development.

CONCLUSION

In this article, I have shown that assessment can be used by biochemistry and molecular biology educators and educational developers as a powerful instrument for "bridging the gap" [23] between research knowledge on curricula, teaching, and learning and its application in the educational process. More specifically, assessment can be used in multiple ways as a powerful tool for affecting student learning and facilitating both faculty development and curriculum change. Since the major aim of this article, and all the articles in this column, is to facilitate the application of ideas from the science education research literature to teaching practice, the following is a summary of the multiple ways in which I recommend biochemistry and molecular biology educators could harness the power of assessment.

- Use summative assessment to measure (grade) student achievement of course objectives;
- Use summative assessment to evaluate instructor and course effectiveness and to inform faculty and curriculum development;
- Use formative assessment to monitor student progress during the course and give feedback that is explicitly used to improve and promote teaching and student learning;
- Use PRSs to give immediate feedback that promotes active learning, critical debate, and the remediation of any difficulties among students;
- Use action research to monitor and inform teaching practice, student learning, and curriculum change;
- Use continuous assessment, with no final examination, to promote consistent studying and to allow students to recover from setbacks and improve their accumulative grades during the course;
- Use what you assess to influence what students learn (learning to the test);
- Use how you assess to determine whether students will use a surface or deep approach to learning;
- Use what and how you assess to influence what and how faculty teach (teaching to the test);
- Use assessment to promote the development of instructors' PCK.

In conclusion, it is perhaps fitting to quote the words of Graue [30] who states, "Choosing the appropriate assessment strategy is a validity concern; the tool must be relevant to the task at hand ... ." In discussing the important relationships between teaching, learning, and assessment, this article has introduced many important aspects of what amounts to the issue of validity. This topic will be the focus of more detailed coverage in the next two articles of this column, where we will illustrate how different forms of assessment need to be used, depending on what knowledge and skills are being assessed.

Acknowledgment—John Rogan (Univ. KwaZulu-Natal, Pietermaritzburg), Adele Wolfson (Wellesley, Massachusetts), Judith G. Voet (Swarthmore College, Pennsylvania), Duane W. Sears
REFERENCES

Bridging the Gap

Bridging the Educational Research-Teaching Practice Gap

CONCEPTUAL UNDERSTANDING, PART 1: THE MULTIFACETED NATURE OF EXPERT KNOWLEDGE*

Received for publication, May 6, 2008

Trevor R. Anderson‡¶ and Konrad J. Schönborn§

From the ‡Science Education Research Group (SERG), School of Biochemistry, Genetics, Microbiology and Plant Pathology, University of KwaZulu-Natal, Pietermaritzburg, South Africa, and §Didaktik der Biologie, Zentrum für empirische Unterrichts- und Schulforschung (ZeUS), Georg-August-Universität Göttingen, Germany

The term “conceptual understanding” has been used rather loosely over the years in educational practice, with a tendency to focus on a few aspects of an extremely complex phenomenon. In this first article of a two-part miniseries on conceptual understanding, we describe the nature of expert (versus novice) knowledge and show how the conceptual understanding of experts is multifaceted in nature requiring competence in a wide range of cognitive skills. We then discuss five such facets of conceptual understanding that require competence in the cognitive skills of memorization, integration, transfer, analogical reasoning, and system thinking. We also argue for the importance of explicitly teaching and assessing such facets of understanding as part of all molecular life science curricula so as to better prepare our students to become experts in the field. Examples of the assessment tasks that can be used to promote the development of multifaceted conceptual understanding in students are presented in Part 2 of this series.

Keywords: Assessment, conceptual understanding, cognitive skills, expert versus novice knowledge, meaningful learning.

The past decade has seen a dramatic increase in biochemical knowledge, leading to extensive debate as to what we should be teaching in terms of the core knowledge and key concepts of our field. Yes, we have a wide range of educational resources and curriculum documents prepared by eminent biochemists and biologists (e.g., see [1]), but, however, these are broad and detailed, and they often do not clearly identify the fundamental core conceptual knowledge of our discipline. In response to this concern, Hamilton et al. [2] have founded the IUBMB-driven Concept Inventory Project, which is currently being piloted by members of an Australian Carrick Grant award (see [3]). These and various other smaller projects, focusing on specific themes such as structure and function of biomolecules, properties of amino acids, reversible equilibrium [4], and metabolism (Degerman and Tibell, personal communication) could have an important influence on education in the molecular life sciences, leading to graduates with more focused and meaningful conceptual knowledge for tackling developments at the forefront of our field.

Content knowledge stored in long-term memory remains inert unless individuals possess the cognitive ability to put such knowledge to active use. This premise is supported by Mayer [5, p. 226] who suggests that successful learning includes not only acquiring knowledge but also having the cognitive skills to use such knowledge in a variety of new situations. Similar sentiments have been expressed by Chattopadhyay [6] who suggests that the objective of genetics learning should be to promote conceptual understanding and thinking (cognitive) skills that encourage students to apply their knowledge to real-life situations. This is supported by the goals of a U.S. Teagle Foundation sponsored project being conducted by Wolfson et al. [7], which, inter alia, addresses the need to teach more cognitive skills to undergraduates, so that they are better prepared for the workplace. In essence, the development of cognitive skills in students can be viewed as fostering “meaningful learning,” which Mayer [5] defines as going beyond the mere learning of factual knowledge to include the development of conceptual understanding in students.

Over the years, numerous authors have attempted to describe what it means to “understand a concept” (e.g., [8]). Unfortunately, “conceptual understanding” continues to be used rather vaguely, focusing only on parts of what is a very complex idea. In support of this, Orgill and Bodner [9] are of the opinion that most biochemical concepts are multifaceted in nature and that true understanding of...
a concept stems from students being able to integrate all facets of a concept into some intelligible whole.

The idea of conceptual understanding being multifaceted was first proposed by White and Gunstone [10], some 15 years ago, who expressed concern that if the definition of conceptual understanding is limited by instructors to only “low-order” cognitive skills, such as memorization, important facets of its meaning would be neglected. For example, a particular student might perform well in a question requiring a definition of the concept of chemical equilibrium (requiring memorization skills), but perform poorly if asked to apply such knowledge to an explanation of the direction of metabolic pathways (requiring application and transfer skills) or to the development of a computer or physical model that represents the equilibrium process (requiring analogical reasoning). For this reason, White and Gunstone [10] also suggested that it is extremely difficult to compare two students’ understanding of a concept, because one might do better in a test of one facet of understanding and another student do better on another. As a solution to their concerns, they suggest instructors teach and assess several facets of conceptual understanding, with each facet of understanding of a concept requiring competence in a different cognitive skill, so that students can develop a more comprehensive understanding of a particular science concept, and faculty can better understand deficiencies in student knowledge.

In this article, we will not attempt to focus on all the facets of understanding of a concept and related cognitive skills ever documented in the literature, but rather on a selected few that we recommend to instructors should be formally taught and assessed as part of all biochemistry course curricula. Our two major arguments for this recommendation are first, such knowledge and skills are essential components of expert knowledge and skill competence and therefore, important to develop in our own students and, second, students have shown various documented difficulties pertaining to each facet, which, therefore, need to be formally addressed and remedi-ated. Thus, this first article of a two-part miniseries on conceptual understanding will aim to address the following specific questions:

- What is the nature of expert (versus novice) knowledge?
- What types of cognitive skills are crucial for students’ development of various facets of expert conceptual understanding?

THE NATURE OF EXPERT (VERSUS NOVICE) KNOWLEDGE

A major goal of education is to develop students from novices into experts. This raises the question of what we mean by an “expert” as opposed to a “novice.” In our view, novice and expert knowledge is defined by a continuum, because there is no clear boundary between when a “novice” has sufficient knowledge to be called an “expert,” and, in the same vein, there is no clear absolute ceiling to expert knowledge. Nevertheless, we can suggest some types of knowledge and skills that tend to characterize an expert when compared with a novice and then focus our teaching on developing such knowledge in our students, so that we “move” or “scaffold” [11] them along the continuum toward expert thinking and understanding.

The science education literature contains extensive research on the nature of expert as opposed to novice knowledge and, therefore, the type of knowledge and skills which would be desirable to develop in modern molecular life sciences curricula. For example, experts display a wide range of higher-order cognitive skills that include the ability to synthesize, critically analyze, and evaluate information [12] that has been gained in a situated context [11]; to be creative, to transfer and apply knowledge to other novel contexts [4], to reason analogically about concepts [13], to reason locally as well as globally (system thinking) [14], and to visualize abstract structures and processes at different levels of biological organization from molecular through to microscopic and macroscopic levels [15]. In addition, through their wider experiences, experts possess a far richer set of connections between ideas than novices do [16]. That is, experts construct more developed and interconnected mental models, explanatory frameworks, and schemata of concepts than novices, who often struggle to integrate their mental models into coherent and powerful conceptual frameworks [17]. In support of this, Khodor et al. [18] suggest that experts’ knowledge is integrated around certain core and fundamental concepts, which provide them with a sophisticated knowledge structure for solving novel problems. In contrast, Stevens et al. [19] maintain that novices, by virtue of their more fragmented knowledge, have a “lower ability to frame the problem” to recognize the importance of problem elements and to prioritize solution strategies. These workers have also found that “novice strategies are often ineffective (they fail to reach the correct answer) and inefficient (they require more steps, more time, and more reference material), whereas experts are more efficient in the use of resources and in deriving the correct answer.” For example, Nahum et al. [17] have pointed out that students often confuse intramolecular and intermolecular bonds, overgeneralize bonding concepts, and memorize terminology without understanding the underlying conceptual relevance of bonding phenomena. Chi et al. [20] also showed that when solving physics-related calculations, novices tended to play “hunt for the formula,” whereas experts qualitatively analyzed the problem and performed order of magnitude estimates before trying to obtain a solution quantitatively. Indeed, many of our own biochemistry students favor using a formula, over first-order principles, to calculate thermodynamic parameters and seem unperturbed when answers reflect unphysiological values such as molar concentrations of cellular metabolites.

Because much of the cognitive processing used by experts to solve problems is “automatic” or part of their tacit knowledge, instructors need to make a directed effort to explicitly explain such crucial processes to novices [21]. In this regard, Novak and Cañas [22] have suggested that getting experts to construct concept maps can be a very useful tool for “exposing” their tacit knowledge, whereas Marbach-Ad et al. [23] used a curriculum
development exercise involving faculty with expertise in host–pathogen interactions to access their tacit knowledge. The following is another approach, which can be a very useful “Bridging-the-Gap” [24] workshop exercise (Grayson, personal communication) for helping colleagues identify their own expert knowledge and skills for development in students:

1. Without going into specific details of content knowledge, address the following question in groups of three or four persons and summarize all the responses on one side of a large board.
   - What type of knowledge and skills do you think characterize your practice as an expert biochemist?
2. Repeat (1) above, but instead, address the following question and record the responses on the other side of the board.
   - What type of knowledge and skills do you explicitly teach in your biochemistry courses?
3. Discuss and compare the responses to the two questions.

In the first author's experience, colleagues usually find a profound difference between the responses to questions (1) and (2), particularly with respect to cognitive skills (e.g., application and linking of knowledge for solving problems) and the requirements for deep conceptual understanding, which tend to predominate in response to question 1 but are mainly absent from responses to question 2. This helps colleagues realize the importance of adjusting their course curricula to be more focused on cognitive skills and the development of expert knowledge in their students. Crucially, what also typically emerges is the realization that being an expert is not about how much content can be "reeled off," but rather about having a basic conceptual "vocabulary" and the cognitive skills required to make sense of new or existing knowledge. Because of the nature of science [25, 26], it is also essential for novices to comprehend that scientific knowledge is dynamic and can be added to or replaced rather quickly, whereas the cognitive skills essential for expert practice develop slowly and remain relatively constant.

To help students develop expert-like knowledge and skills, it is important to promote what Entwistle and Ramsden [27] call a deep approach rather than a surface approach to learning. As discussed in a previous article of this column [28], a deep learning approach requires students to develop understanding through the use of various cognitive skills such as integration, critical analysis, and application of knowledge. This relates well to the ideas of Bloom et al. [29] who, some 50 years ago, published a taxonomy of educational objectives, which has made a significant impact on teaching, learning, and assessment. Allen and Tanner [30] have demonstrated how this taxonomy can be used effectively in cell biology for the design of questions that probe different aspects of students’ knowledge and cognitive skills. More recently, Anderson et al. [12, 31] have revised the taxonomy to include both a knowledge and cognitive process dimension. Aspects of the cognitive process dimension were developed by educational psychologist, Richard Mayer, who, in a subsequent work [5], illustrates how the idea of deep learning is closely related to fostering a meaningful learning approach as a major educational goal. According to Ausubel [32], and in line with a constructivist view of learning [33], meaningful learning occurs when students have sufficient prior knowledge upon which to anchor new ideas. Fisher et al. [34] extend the definition by stating that “meaningful learning occurs when individuals actively incorporate new ideas into their mental structures, ask themselves what the implications of the idea are, and assess whether those implications make sense” (p. 80). Mayer [5] maintains that a key outcome of meaningful learning is that students should not only become competent in the cognitive process of remembering (recognizing and recalling) scientific knowledge but also develop other cognitive competencies that characterize experts such as understanding, applying, analyzing, evaluating, and creating. In this regard, Novak [35] views learning on a continuum, which can vary from "extreme rote" to "highly meaningful," with the latter involving higher-order cognitive skills. Most importantly, Mayer [5] recommends that such cognitive competencies should be explicitly taught as a primary educational objective. Meeting this objective is where assessment can show great power as a "Bridging the Gap" tool. Because of the well-established importance of aligning course objectives with teaching, learning, and assessment (see [28]), specifically designed assessment tasks can be extremely effective in directing and promoting the teaching, learning, and development of such cognitive skills in our students, so that they progress toward becoming experts [28].

THE MULTIFACETED NATURE OF CONCEPTUAL UNDERSTANDING

On the basis of the above argument, we pose the question, which facets of conceptual understanding compose expert understanding of a concept? Because it is neither feasible nor practical to teach and assess all facets of understanding of a concept, it is important to try and select the most important ones. There are facets of conceptual understanding that instructors should always focus on regardless of the nature of the concept and others that are of greater relevance to a particular concept or educational context. Table I lists a selection of what we consider to be important cognitive skills cen-

---

**Table I**

Selected cognitive skills that are central to various facets of expert conceptual understanding in biochemistry and molecular biology

<table>
<thead>
<tr>
<th>Understanding a concept means the ability to:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memorize knowledge of the concept in a mindful manner, as distinguished from rote learning</td>
</tr>
<tr>
<td>Integrate knowledge of the concept with that of other related concepts so as to develop sound explanatory frameworks</td>
</tr>
<tr>
<td>Transfer and apply knowledge of the concept to understand and solve (novel) problems</td>
</tr>
<tr>
<td>Reason analogically about the concept</td>
</tr>
<tr>
<td>Reason locally and globally about the concept (system thinking)</td>
</tr>
</tbody>
</table>

---
tural to various facets of expert conceptual understanding in biochemistry and molecular biology.

Bloom’s revised taxonomy [12] advocates the following standard format for stating educational objectives, which we also use in Table I: “The student will be able to, or learn to, verb noun,” where the verb indicates the cognitive process and the noun the concept or knowledge that is intended to be learnt. A key aspect of these objectives is that they should be teachable and assessable, and the type of assessment task used should be aligned with course objectives (see [28]). In this section, we briefly discuss the five cognitive skills that are central to various facets of expert conceptual understanding (Table I), while examples of assessment tasks that can be used to measure and promote development of such cognitive skills in our students are presented in Part 2 of this miniseries.

Mindful memorization (Table I) is not rote learning, but a crucial cognitive skill required for the understanding of all concepts. Fisher et al. [34] defines “mindfulness” as paying attention with intention to understand or make sense of information. Thus, mindful memorization can be defined as an essential cognitive process in which information is memorized with some specific intention or purpose in mind, namely to understand, use, or apply the information in problem-solving activities that require higher-order cognitive skills. Therefore, it is important that the memorization precedes any other higher-order cognitive processes, such as transfer and application, because all cognitive processes require “something to process”—in this case—memorized information. Memorization skills are a crucial component of all expert competence (Table I), and we need to teach and assess (see part 2 of this series) such basic skills in our students, so that they have meaningful underlying knowledge upon which to exercise their higher-order skills. Knowing facts or being able to regurgitate a concept should, however, not be confused with understanding the nature of the concept. One increasingly important facet to know in this era of instant electronic retrieval of information and open-book tests is to know that something is known (e.g., the citric acid cycle) and where to find it and should you need it. This is not exactly memorization but is related to it and is an important facet of learning (anonymous reviewer, personal communication).

The memorization of isolated facts leads to knowledge that exists in the mind as “disconnected chunks” of information [36]. For memorized knowledge to become useful for understanding phenomena and for solving problems, the “chunks” need to be integrated into a meaningful network or schema of concepts that serve as a sound explanatory framework for a particular topic (Table I). Such explanatory frameworks and the integration skills necessary for developing them characterize an important facet of expert conceptual understanding and problem-solving knowledge, and, therefore, we need to explicitly teach and assess such skills in our students. Concept mapping [37, 38] is recommended to biochemists as an example of a useful tool for illustrating the nature and extent of an expert’s versus a novice’s integrated knowledge. Concept maps are composed of concepts (that form the “nodes” of the map) linked to each other by propositional statements (written along mono or bidirectional arrows) that describe relationships between concepts. They are thought to mimic the storage of knowledge in the mind [38] and to reflect the nature, elements, links, network, and structure of a person’s explanatory framework about a particular topic and, therefore, their understanding of the relevant concepts. Thus, concept maps are also very useful for revealing evidence of poor integration skills, deficiencies in knowledge, or alternative conceptions [39] that require remediation [40]. More extensive visualization of large groups of concepts can be achieved by generating semantic networks [41]. Semantic networks allow for visualizing both 3D and bidirectional connections between concepts, allowing for the development of sophisticated concept descriptions. For example, Gorodetsky and Fisher [16] reported that biology experts typically demonstrate 20–30 bidirectional links to big ideas such as “DNA” or “protein,” whereas students usually create far fewer links. In Part 2 of this series, we will discuss how concept maps and semantic networks may, with the aid of various software and map scoring methods, be used as tools for promoting and assessing students’ conceptual understanding.

The facet of expert understanding requiring competence in the cognitive skills of transfer and application of knowledge (Table I) is considered by Mayer [5] to be a fundamental goal of education and is arguably the most important facet that biochemistry and molecular biology experts draw on in their practice. Transfer has been defined by Mayer and Wittrock [42] as the ability to use or apply knowledge of a concept to solve new problems, answer new questions, or facilitate learning of new subject matter. For transfer and application of knowledge about a concept to be successful, a person must first memorize and integrate the key concepts into an explanatory conceptual framework. That is, their prior conceptual knowledge must be adequate and soundly constructed and must have been initially learned and well understood in a meaningful and situated context before they abstract, transfer, and apply it to solving problems in another context [5, 11, 43, 44]. Science education research (e.g., [45]) has shown that students have numerous difficulties regarding knowledge transfer. Therefore, it is essential to explicitly and formally teach and assess transfer and application skills as part of our biochemistry and molecular biology curricula, so that such difficulties can be corrected and so that our students develop the problem-solving skills that characterize experts.

According to Orgill and Bodner [9], the cognitive process of analogical reasoning (Table I) involves comparing a familiar domain (the analog) with a less familiar domain (the target). Analogies enable students to make connections between abstract concepts (the target) and more concrete concepts (the analog) with the intention of promoting understanding of the abstract concepts. Because learning in biochemistry depends very heavily on understanding the abstract world of molecular structures and processes, developing analogical reasoning skills is absolutely crucial for becoming an expert biochemist. In addition, modern biochemistry textbooks and accompa-
nying computer and web-based resources make great use of analogies to explain and visualize biochemical phenomena while instructors make extensive use of them during teaching. Examples of analogies used in biochemistry (e.g., [9, 46]) include the “lock-and-key” analogy representing enzyme-substrate binding; the “hand-and-glove” analogy for induced fit in enzyme-substrate binding; the term “chaperone” for proteins that direct the folding of other proteins; the term “fluid mosaic” for the dynamic structure of a cell membrane; DNA as a “recipe” for creating a human organism; ATP as the energy “currency” in cells; and, the mitochondrion as the “power plant” of the cell. Analogies play a variety of educational roles including clarifying thinking and inducing sound understanding of a concept or phenomenon [47], adjusting or changing an alternative conception (e.g., [48, 49]), or addressing inappropriate reasoning difficulties [50]. It is important to note one caveat about the use of analogies. They can also sometimes be misleading, because students may perceive the analog literally and transfer the wrong features of the analog to the target. It is important that biochemistry instructors teach and assess analogical reasoning skills as part of all course curricula to minimize student conceptual difficulties and promote development of expert skills. In Part 2 of this series, we present examples of tasks that can be used to assess and promote such development in our students.

The ability to reason locally and globally (Table I) about a concept or phenomenon and to grasp the implications of local effects on a complete living system is a crucial, but often disregarded, facet of experts’ biochemical knowledge. Local and global reasoning skills are essential for understanding concepts and phenomena in the molecular life sciences. For example, it is important to understand that any local changes, such as activation or inhibition of one reaction in a metabolic pathway might also have global effects on all the reactions in a pathway as well as on the system as a whole. In this regard, Anderson and Grayson [51] obtained empirical evidence of a localized reasoning difficulty in which several biochemistry students correctly thought that the irreversible inhibition of glyceraldehyde-phosphate dehydrogenase would stop this reaction (i.e., would have a local effect) but failed to predict that this inhibition would have a global effect on the overall flux through glycolysis. Such difficulties have important implications for students’ thinking when studying systems biology [14, 52] and for understanding metabolic control analysis where extensive kinetic and thermodynamic information from single reactions (local) are used in mathematical modeling to predict the behavior of a (global) system. System thinking requires students to no longer think linearly and in a reductionist manner about metabolic pathways and about dated phenomena such as single rate-limiting steps [53]. Instead, system thinking involves considering the dynamic nature of metabolism and signal transduction pathways and the role of all reactions in a pathway in determining metabolic flux. Empirical studies in science education suggest a strong link between the development of system thinking and conceptual understanding [54], whereas system thinking has also been shown to be fundamental to students’ construction and holistic integration of mental models in the life sciences [55]. Indeed, Evagorou et al. [55] have recently reported on the use of interactive simulations for developing system thinking in fifth and sixth graders at elementary school. These workers cite the following six abilities of expert system thinkers (p. 2), which we suggest could also be applied to higher education contexts of the molecular life sciences:

- Analyze interrelationships between different objects and explore emergent properties;
- Analyze phenomena and problems in wider contexts;
- Consider multiple cause-and-effect relationships;
- Discover and represent dynamic processes (delays, feedback loops, and oscillations) that underlie patterns of the system’s behavior;
- Anticipate the long-term consequences and possible effects of present actions;
- Understand changes in a system over time [55].

In view of the importance of system thinking as an expert skill, we recommend that it be taught and assessed as part of formal biochemistry curricula, so that our students can develop this facet of understanding of biochemical concepts. To promote the development of system thinking in our students, we will need to develop assessment tasks that address each of the above abilities. Examples of such tasks are presented in Part 2 of this miniseries on conceptual understanding.

CONCLUSION

The aim of this article was to expose readers to literature regarding the nature of conceptual understanding. We are of the view that “conceptual understanding” is a term that is used rather loosely both in science and education, because it is multifaceted, complex, and most instructors tend to focus only on selected aspects of the phenomenon. To optimize students’ understanding of a concept, so that it becomes comparable to that of experts, instructors need to teach and assess as many of the facets of the concept as feasible. However, clearly neither students’ nor experts’ understanding of a concept will ever be absolute as there will always be other facets of understanding that require competence in other cognitive skills. This fits well with a life-long learning philosophy and the idea of “moving” along the novice-expert continuum. In this regard, White and Gunstone [10] state that “understanding of a concept is not a ‘dis-
nitive skill competencies and, therefore, the conceptual understanding of their students:

- Consider conceptual understanding as a multifaceted phenomenon that requires competence in specific higher-order cognitive skills;
- Identify what facets of understanding and skill competence constitute expert knowledge of the topic you are teaching and aim to develop as many of these as possible in your students;
- Explicitly, teach and assess these facets of understanding and cognitive skills as a formal part of every course in the biochemistry and molecular biology curriculum.

In Part 2 of this miniseries on conceptual understanding, we will use specific examples of assessment tools to illustrate how one might use assessment as a powerful "Bridging the Gap" tool for promoting the teaching, learning, and development of multifaceted knowledge in our students.

Acknowledgments—Judith G. Voet (Swarthmore College, Pennsylvania) and two anonymous referees are thanked for their valuable critique of earlier versions of this article and Patricia McLean (Univ. KwaZulu-Natal, Pietermaritzburg) for English proofing.

REFERENCES AND FURTHER READING


Bridging the Gap

Bridging the Educational Research-Teaching Practice Gap

CONCEPTUAL UNDERSTANDING, PART 2: ASSESSING AND DEVELOPING STUDENT KNOWLEDGE*

Received for publication, July 10, 2008, and in revised form, July 28, 2008

Konrad J. Schönborn‡ and Trevor R. Anderson§¶

From the ‡Division of Visual Information Technology and Applications (VITA), Department of Science and Technology (ITN), Linköping University, Norrköping, Sweden, and §Science Education Research Group (SERG), School of Biochemistry, Genetics, Microbiology and Plant Pathology, University of KwaZulu-Natal, Pietermaritzburg, South Africa

The first paper [1] in this two-part miniseries on conceptual understanding discussed expert and novice conceptual knowledge, the multifaceted nature of conceptual understanding, and the cognitive skills essential for constructing it. This second article presents examples of instruments for the assessment and development of five facets of conceptual understanding that require competence in the cognitive skills of mindful memorization, integration, transfer, analogical reasoning, and system thinking. We also argue for the importance of explicitly assessing these facets of conceptual understanding as part of all biochemistry and molecular biology curricula so as to develop expert knowledge in our students.

Keywords: Assessment tasks, measuring and developing conceptual understanding, cognitive skills, meaningful learning.

The first article [1] in this miniseries on conceptual understanding described how cognitive skills are essential for the construction of expert knowledge. We argued that conceptual understanding and the associated cognitive skills are multifaceted in nature, and that it is necessary to explicitly develop such knowledge and cognitive abilities in our students in order to scaffold them along the novice-expert continuum. In this regard, the following five cognitive skills were considered crucial to the development of knowledge about each scientific concept [1] and to the fostering of meaningful learning in biochemistry and molecular biology students:

1) Memorize knowledge of the concept in a mindful manner, as distinguished from rote learning;
2) Integrate knowledge of the concept with that of other related concepts so as to develop sound explanatory frameworks;
3) Transfer and apply knowledge of the concept to understand and solve (novel) problems;
4) Reason analogically about the concept;
5) Reason locally and globally about the concept (system thinking).

Because assessment can be used as a powerful “Bridging-the-Gap” tool for promoting learning [2], we argue that specifically designed tasks, which focus on each of the above cognitive skills [1], could be effective in both promoting and grading students’ cognitive competence and their multifaceted understanding of concepts. Indeed, assessment should focus on stimulating learners to actively engage in constructing meaning, rather than passively recalling knowledge (e.g. [3, 4]). In support of this, Klymkowsky et al. [5], founders of the Biology Concept Inventory (BCI),1 have stated that biology education requires instruments that go beyond testing rote learning. Fostering such a “bioliteracy” is intertwined with measuring conceptual understanding “as a means of assessing student fluency in a subject area and as a measuring stick of, and impetus for curricular reform aimed at improved teaching and learning” [5, p. 156]. These workers [6] have stressed the need for assessment tasks that can enhance students’ understanding of the key biological ideas. The objective of this article is to address the following questions:

*This work was supported by the NRF (GUN Number 2053218) and Linköping University.
¶ To whom correspondence should be addressed. Science Education Research Group (SERG), School of Biochemistry, Genetics, Microbiology and Plant Pathology, University of KwaZulu-Natal, P/Bag X01, Scottsville, Pietermaritzburg 3209, South Africa. Tel.: +27332605464; Fax: +27332606127; E-mail: anderson@ukzn.ac.za.

1The abbreviations used are: BCI, Biology Concept Inventory; POE, predict–observe–explain; FLAG, Field-tested Learning Assessment Guide; MCQ, multiple-choice questions; BCF, biology concept framework; carbon monoxide, CO; oxygen, O2; THF, tetrahydrofolate; dTMP, thymidine monophosphate; DHF, dihydrofolate; DHFR, dihydrofolate reductase.
• What types of instruments are available for assessing conceptual understanding in science education?
• What assessment tasks can be used to measure and develop the five cognitive skills central to each facet of conceptual understanding in biochemistry and molecular biology?

INSTRUMENTS FOR ASSESSING CONCEPTUAL UNDERSTANDING

The science education literature contains numerous examples of assessment instruments that are deemed reliable and valid tools for measuring students' conceptual understanding. As conceptual understanding is multifaceted [1], use of one type of instrument will only yield information about a particular aspect of students' knowledge and will never serve as a complete indicator of understanding [7]. Therefore, it is important to use a range of instruments that probe a variety of cognitive skills.

White and Gunstone [7] have described various ways of assessing students' conceptual understanding. These include concept mapping (see Qu. 3 later), word association, predict–observe–explain (POE), and drawing-related tasks. In addition, Sundberg [8] has outlined a range of “quantitative” and “qualitative” assessment tools and suggests that modern teaching requires multiple forms of assessment if instructors are to truly monitor student understanding. Types of assessment suggested include pretest–posttest instruments, Likert scale items, observation, individual and group interviews, and concept mapping. In addition, short-answer essays (see Qu. 2, 6–8 later) and empirically based diagnostic tests (see Qu. 4 later) can also be helpful. Resources such as the Field-tested Learning Assessment Guide (FLAG) [9] can also provide instructors with practical information about different assessment techniques [8].

Multiple-choice questions (MCQs) are popular in the biological sciences [10] for assessing students' conceptual understanding, particularly in large classes where grading is time-consuming. MCQs are particularly useful for the development of concept inventories and related diagnostic tests that aim to assess students' conceptual understanding. The prevalence of such tests is expanding and they include, for example, the BCI [5, 6], the conceptual inventory of natural selection [11], as well as question banks focusing on photosynthesis [12], diffusion and osmosis [13], cellular transport [14], and visualized biochemical structures [15]. In addition, Khodor et al. [16] have formulated a hierarchical biology concept framework (BCF) that shows the relationships between concepts, highlights concepts that are difficult for novices to learn, and demonstrates what concepts are important for teachers to assess. Klymkowsky et al. [5] are developing items for the BCI that can be used to assess students' conceptual understanding of the fundamental ideas in molecular, cell, and developmental biology. These authors view their inventory as a “lever for moving our current educational system in a direction that delivers a deeper conceptual understanding of the fundamental ideas upon which biology and biomedical sciences are based” (p. 155). In this regard, because students come to class with their own uniquely constructed ideas and beliefs about how the world works (e.g. [17, 18]), such prior knowledge often conflicts with scientifically acceptable knowledge, preventing students from engaging in meaningful learning.

Although popular, MCQs may or may not yield useful information on students' conceptual understanding, depending on their particular design. There are at least three variants of MCQs: The common one-tier, one-tier with a written explanation for the choice made [19], and two-tier MCQs [20]. One-tier MCQs do not always provide reliable measures of conceptual understanding because one student's selection could be based on a sound understanding while another student's selection could be the result of simple “guesswork.” In addition, one-tier MCQs also “suffer” from the erroneous assumption that at least one of the response options is a mirrored account of a student's understanding of a concept. To help alleviate this problem, one-tier MCQs that require students to give an explanation for their choice (see Qu. 4), induce expression of actual understanding. As students often “know” the right answer but often do not know the reason why, this design allows the instructor to independently grade the “MCQ part” and/or the “explanation part,” or use the explanation to check whether exposed difficulties are real. Another approach is to ask students to rate how confident they feel about their choice(s). In two-tier MCQs [20], the first tier is a multiple choice content question based on propositional statements (the “what”) and the second tier is a further multiple-choice group of reasons corresponding to the first tier (the “why”). The second tier is composed of alternative conceptions and one scientifically sound reason that are often generated from educational research. The second tier sometimes includes a further “choice” in which learners can provide an alternative written answer should they feel none of the second-tier statements are “correct.” Several studies have shown that two-tier MCQs are a reliable and valid means for diagnosing conceptual and reasoning difficulties or learning gains (e.g. [12–14]).

Azer [21] has expressed concern that the literature contains few guidelines for designing MCQs that assess deep conceptual understanding and higher-order thinking skills in medical education. In response to this concern, Azer provides examples of common pitfalls to avoid in constructing MCQs, including avoiding double-negative statements in distractors. Azer also provides practical tips for helping examiners design MCQs that test cognitive skills required for deeper understanding. These include suggestions that MCQs should address specific educational objectives, encourage integration and application of knowledge, address factors that affect the validity of questions, and that students are exposed to examples prior to assessment. In addition, Haladyna et al. [22] have developed a comprehensive taxonomy that describes 31 general MCQ-writing guidelines that particularly focus on content, formatting, style, writing the “stem,” writing the “choices,” and ensuring plausible distractors.
While not always practical for instructors, the ideal scenario is to adopt a science education research approach that first uses open-ended questions to establish real alternative conceptions, which are then incorporated into MCQs as reliable and valid distracters [19]. This is extremely important as there is growing evidence (e.g. [23]) that distracters can be considered by students as scientifically plausible causing them to develop new alternative conceptions. The advantage of open-ended questions is that they can provide rich qualitative data on aspects of students’ understanding, including misconceptions and reasoning difficulties [19]. However, open-ended questions cannot be easily administered to large groups of learners, whereas MCQs are particularly suitable for such classes and can readily generate information suitable for direct quantitative analysis.

White [24] has called for instructors to employ “concept questions” for measuring “real” understanding in biochemistry and molecular biology. Generally speaking, concept questions do not require the passive regurgitation of knowledge and they are largely unaffected by “open-book” assessment conditions. In our view, conceptual questions that are of the short-answer (see Qu. 1, 4), open-ended, and short-essay types (see Qu. 2, 6–8), diagram-related type, or observable type (see Qu. 5) can yield deep insights into students’ understanding. Unlike one-tier MCQs, where response options are fixed, these instruments allow for a potentially unlimited range of scientifically-sound (and soundless) answers. In this regard, we recommend the use of an “evolving model answer” in which scientifically correct responses other than that of the instructor’s solution are pooled as the grading of scripts proceeds and are given equal chance of scoring full marks. This approach encourages students to be more creative and confident in their thinking as they realize that their own ideas will receive fair treatment. In turn, this minimizes typical “what the lecturer wants to hear” responses. Hence, open-ended type questions are sometimes more useful than MCQs for assessing conceptual understanding.

ASSESSING AND DEVELOPING THE FIVE FACETS OF CONCEPTUAL UNDERSTANDING IN STUDENTS

In this section, we present specific assessment tasks that can be used to measure and promote the development of student knowledge with respect to the five facets of conceptual understanding and the corresponding cognitive skills [1]. Our objective is to illustrate how readers might use assessment as a powerful “bridging-the-gap” tool for fostering students’ development of multifaceted knowledge indicative of experts. The presented tasks stem from three sources, our own educational research, the science education literature, and from two experienced biochemistry educators. Our aim is to present a limited number of tasks that clearly illustrate how each cognitive skill might be assessed and developed in students. We do presuppose that biochemistry students would be familiar with the conceptual area pertaining to each of the tasks. Suggested “optimal” solutions to each of the tasks are briefly presented in the Appendix.

MEMORIZING KNOWLEDGE OF THE CONCEPT IN A MINDFUL MANNER

According to Mayer [3, p. 228], the ability to memorize or remember factual information is comprised of two related cognitive processes. Recognizing (identifying) is described as “locating knowledge in long-term memory that is consistent with presented material.” Recalling is described as “retrieving relevant knowledge from long-term memory” [3, p. 228]. Although the engagement of both recognizing and recalling processes is essential for acquiring factual knowledge, mindful memorization requires going a bit beyond simply “knowing,” recalling or being able to “repeat” facts, and it is more concerned with employing a richer level of memory [1, 25]. Mindful memorization involves acquiring factual information with some specific intention or purpose in mind—for employment in higher-order cognitive activities such as the integration and application of knowledge for the solving of problems. Thus, although we wish to avoid overemphasizing the role of memorization in the development of students’ conceptual understanding, we feel that it is crucial to explicitly distinguish this process from rote learning (e.g. [24]) and emphasize that “parroting” is not synonymous with deep understanding (anonymous reviewer, personal communication). In this regard, we suggest that the following two tasks developed by H.B. White (personal communication) are representative examples for assessing students’ mindful memorization of knowledge:

Qu. 1: What is the primary source of energy for a 100-m race?

Qu. 2: There are a number of antifolate drugs, including methotrexate, trimethoprim, and sulfanilamide. Select one of these. Describe how it works by explaining what medical conditions it is used to treat.

Responding to Qu. 1 requires the student to have some or other purpose for the “energy source” in mind. By the same token, Qu. 2 calls on the ability to “make sense” of information by first describing how an antifolate drug might function, followed by expressing an understanding of the implications of this knowledge for medical treatment. Thus, assessing students’ mindful memorization of knowledge is an important facet of conceptual understanding in the biomolecular sciences because it is a cognitive priori to building knowledge of a concept, one that precedes other “higher-order” cognitive skills such as those discussed later.

INTEGRATE KNOWLEDGE OF THE CONCEPT WITH THAT OF OTHER RELATED CONCEPTS SO AS TO DEVELOP SOUND EXPLANATORY FRAMEWORKS

The promotion of students’ explanatory frameworks and accompanying integration skills are crucial for fostering expert knowledge in the life sciences [1]. In this regard, Novak [26] has shown that concept mapping is a versatile tool for promoting the development of students’ integration skills and integrated knowledge, and for assessing changes in students’ conceptual understanding during a course. Concept mapping tasks can take
several forms. For example, students could be supplied with specific concepts, a diagram representing related concepts, or asked to identify key concepts essential for explaining a specific topic. A favored map structure (e.g., “top down hierarchies,” “spokes,” “chains” or “nets”) may also be specified, or students could be provided with a partially completed map and asked to “add in” concepts or supply necessary links. In the following task kindly supplied by H.B. White (personal communication), students have to “complete” a map to demonstrate how different concepts and processes are interrelated:

Qu. 3: Carbon monoxide, an odorless, colorless, and tasteless gas, is deadly. An atmosphere containing 0.1% carbon monoxide (CO) can lead to death within an hour. CO binds to human hemoglobin 220 times more tightly than does oxygen (O₂). In 2003, Nascar driver Rick Mast retired because of acute and chronic CO poisoning that resulted in severe headaches and other symptoms in his work environment (New York Times, February 2, 2003). Analysis of the blood of Americans shows that between 0.5 and 2% of the hemoglobin in the blood of rural nonsmokers has bound CO. This level can increase to as much as 5% in urban nonsmokers and 9% in heavy smokers. The value is never 0%, because CO is produced naturally in the body in the conversion of heme to bilirubin.

Complete the concept map below by filling in the empty boxes with the most appropriate words:

![Concept Map](image)

Solving Qu. 3 requires students to engage the higher-order skill of integrating a range of potentially “disconnected chunks” of information [1], about concepts pertaining to hemoglobin and related physical, molecular, and cellular processes, into an explanatory framework that captures the understanding of hemoglobin and CO poisoning. Thus mapping tasks such as Qu. 3 are beneficial for both developing and assessing students’ integration skills and integrated understanding of a topic. The literature contains several avenues for using concept maps to assess such student knowledge. For instance, concept maps can be used in formative assessment [2] to provide feedback to students and instructors. This can involve comparing student maps with expert maps through the use of resources such as Cmap [27] which offers a “compare concept maps” facility. In this case, an expert map can be used as a “control” for identifying “gaps” in the students’ knowledge as well as so-called “critical” [28] or “threshold” [29] concepts that underpin higher-order concepts. Identification of such concepts, as well as any inappropriate links between concepts, can inform strategies for correcting the student’s difficulties. Student maps may also be scored during summative assessment [2] by awarding students grade points that include the number of relevant concepts, the number of valid links between concepts, and map complexity [28]. Examples of more complex scoring systems are outlined by Wandersee [30] and Ruiz-Primo [31]. Although such scoring systems are useful for grading, concept maps tend to be unique to each individual and, therefore, not always easy to assess (anonymous reviewer, personal communication).

It is important that students practice generating maps, as the tool in itself requires procedural skills that need to be explicitly developed. In this regard, practical protocols for constructing “paper-based” concept maps have been...
described by White [32] and Allen and Tanner [33], while software such as Cmap [27] and Inspiration [34] could be useful to instructors. More advanced visualization of the relationships between students' concepts can be achieved through the use of SemNet [35] software, which externalizes learners' knowledge as highly sophisticated "semantic networks" (anonymous reviewer, personal communication). Relationships in semantic networks are $n$-dimensional, meaning that they can point (bidirectionally) in an infinite number of directions in 3D space. Gorodetsky and Fisher [36] observed that biology students who used SemNet recalled and accurately used twice as many concepts as students who did not use the tool.

Transfer and Apply Knowledge of the Concept to Understand and Solve (Novel) Problems

The application of knowledge about a concept to the solving of (novel) problems is a crucial facet of conceptual understanding and meaningful learning [1, 3]. Such transfer processes stimulate students' construction of integrated knowledge, thereby rendering the knowledge more accessible and useful [37]. This is essential as it is well known that students tend to, for example, separate what they learn in physics and chemistry from biochemistry (e.g., [4]) and only "open" one "content box" at a time, rather than transfer knowledge across disciplines. It is necessary to formally teach and assess transfer skills so that students develop problem-solving proficiencies that characterize experts [1].

The science education literature suggests that there are at least three prerequisites for successful transfer (e.g., [38]). First, knowledge must be memorized mindfully [25]. Second, knowledge must be integrated into an explanatory framework and contextualized [3, 37]. Third, students need to be made aware that knowledge learnt in one context is often applicable to other contexts. Assessment can be used as a powerful tool for promoting the development of transfer skills in students [1]. For example, consider the following MCQ (Qu. 4) developed and validated by Garvin-Doxas and Klymkowsky [6, p. 232]. We believe that this question is a very effective means for assessing students' understanding of the diffusion concept and promoting the development of transfer skills in students, through requiring them to apply their knowledge of the concept, learned in chemistry (or physics), to the context of a biochemical process, while avoiding inappropriate transfer described in the distracters. We have modified the question by requiring students to provide a reason for their choice:

Qu. 4: Imagine that you are an ADP molecule inside a bacterial cell. Which best describes how you would manage to "find" an ATP synthase so that you can become an ATP molecule?

- a) I would follow the hydrogen ion flow
- b) The ATP synthase would grab me
- c) My electronegativity would attract me to the ATP synthase
- d) I would be actively pumped to the right area
- e) Random movements would bring me to the ATP synthase

Provide a reason for your selection above . . .

Thus Qu. 4 requires students to transfer their knowledge of the concept that "the random process of diffusion takes place all the time" to the context of a (novel) biological process [6]. Some readers may feel that the anthropomorphic nature of Qu. 4, that is, that the ADP molecule has some or other "conscious intentionality" (G.E. Höst, personal communication), may be a potentially misleading device that should be avoided in the design of assessment items (anonymous reviewer, personal communication). However, research by Zohar and Ginosar [39] has provided arguments for the potential pedagogical value of anthropomorphic reasoning, provided students are explicitly guided as to the role that a particular anthropomorphism is playing in the communication of a concept.

In conclusion, although biology education research has shown that knowledge transfer is often a demanding process for learners to engage in [40], it is perhaps the most essential expert facet of conceptual understanding in biochemistry and molecular biology and it should be explicitly developed and assessed in our students.

Reason Analogically About the Concept

Analogies are used extensively to reason about phenomena [1], a process that is not automatic in novices [41]. In agreement with Orgill and Bodner [42], we propose that this integral facet of conceptual understanding be explicitly taught and assessed as part of formal biochemistry curricula. To facilitate analogical reasoning, it is key that instructors explicitly explain the relationship between the analog and target concept(s) to ensure that students avoid interpreting the analogy as a literal depiction of reality. This is important as the nature of the analogy, students' conceptual knowledge and their analogical reasoning skills are all pivotal factors in determining whether an analogy will foster student understanding (e.g., [43]). Given that analogical reasoning is an important component of expert knowledge, the following assessment task could contribute to promoting this facet of conceptual understanding:

Qu. 5: Use the three different-strength magnets and iron nails provided to develop an analogy for explaining the transfer of electrons between electron carriers during oxidative phosphorylation. Fully describe how your analogy is related to what you think are the "target" concepts. Also suggest how the features of your analogy are similar and different to how electron transport occurs in vivo.

The analogical reasoning required to respond to Qu. 5 involves at least three mappings between analog and target concepts. First, students should map between the idea of each magnet being representative of a single electron-carrier and each nail being representative of an electron. Second, students should map the idea of the differences in magnetic attraction between nails and
magnets of different strengths, to the idea of differences in "electron affinity" between electron carriers of different redox potentials that determines the sequence of electron transport. Third, to gain understanding about how electrons are physically transferred, it is necessary for students to map the idea of a collision between magnets (and corresponding transfer of one nail from one carrier to another) to the idea that, during redox reactions, electrons are transferred through random collisions between carriers rather than "jumping" from one statically positioned carrier to another.

Apart from stimulating learners to expose the limitations and strengths of a particular analogy when promoting analogical reasoning, it is also very important for instructors to explicitly point out the dissimilarities (and similarities) between analog and target concepts (anonymous reviewer, personal communication). In addition, Clement [44] has proposed a useful teaching strategy for remediating alternative conceptions involving the use of bridging analogies during active learning environments. Such analogies are considered as an intermediate between the source analog and the target concept and share features with both of them. The bridging analogy divides the process of matching the source to the target analogy into smaller, easier to comprehend steps. In this way, students can identify both the similarities and dissimilarities between analog and target concepts so that they correctly map the one to the other (anonymous reviewer, personal communication). Thus in summary, we suggest that it is vitally important to teach and assess analogical reasoning as part of curricula so that the students gain this important facet of expert conceptual understanding.

**Reason Locally and Globally About the Concept (System Thinking)**

Local and global reasoning skills are fundamental to biological understanding. However, students often focus only on the local effects, and tend to neglect the global effects of dynamic processes [1]. For example, Cohen et al. [45] found that when students were asked what happens when one part of an electric circuit was altered, 27% of students considered only the local effects of the change and did not think about global changes to the whole circuit. Within the domain of metabolic pathways, explaining the control properties of networks of coupled enzyme-catalyzed reactions in steady state requires systems thinking and a theoretical framework such as metabolic control analysis (e.g. [46]). The reason for this is that an emergent, systemic property (such as the degree of control that a particular enzyme exerts over the flux through the pathway of which it is part) depends on the local properties of all the individual enzymes and the way they are connected, and it cannot be explained in terms of the properties of that single enzyme alone (J.-H.S. Hofmeyr, personal communication). The following three tasks (kindly supplied by Hofmeyr) illustrate the type of questions that could be used to assess and facilitate the development of students’ system thinking in biochemistry.

Qu. 6: An idea that still permeates much of the biochemical literature and textbooks is that a sequence of coupled reactions must contain one step, the so-called rate-limiting step, which fully determines (controls) the metabolic flux through the pathway. Explain what properties an enzyme-catalyzed step must have in order to be rate-limiting, that is, fully control the steady-state flux through the sequence of reactions of which it is part.

Qu. 7: A central result of metabolic control analysis is that the control of flux can be shared between the steps in the system, and that no step need be fully rate-limiting. Using a simple two-step coupled reaction system, explain what properties the two enzymes must have in order for the second to have 60% control over the flux and the first the remaining 40%.

Qu. 8: Explain why a near-equilibrium reaction cannot control the flux of a metabolic pathway.

Finding a feasible solution to Qu. 6 requires students’ to first develop the “local” thinking that, for a step to be “rate-limiting,” it must be insensitive to changes in its substrate and product concentrations. Qu. 7 then calls on the development of a more “global,” system-type thinking to understand that the flux of a particular step in a metabolic pathway depends on the properties of all the steps in the system. Finally, Qu. 8 can be used to promote more advanced system-type thinking in that, although a near equilibrium reaction step in a pathway is very sensitive to changes in its substrate and product concentrations, the degree to which it is from equilibrium must be considered in the light of the overall metabolic system of which it is part.

Overall, students’ local and global reasoning about a concept is an important component of expert knowledge. Such emphasis is supported in a recent study by Verhoeff et al. [47], who have shown that students’ coherent understanding in cell biology shares an intimate relationship with a “systems thinking competence” (p. 543). Consequently, tasks such as those aforementioned could be used to assess and develop this type of reasoning as part of formal curricula in biochemistry and molecular biology.

**CONCLUSION**

The aim of this article was to provide examples of tasks that can be employed to assess and develop five facets of students’ conceptual understanding and the corresponding cognitive skills [1]. An ideal assessment scenario would be to develop questions that are both valid and reliable measures of conceptual understanding as well as being efficient to grade (J. Voet, personal communication). In this regard, we have presented eight examples of tasks that could partly satisfy these requirements. Overall, well-designed MCQ variants or “short answer” items could be suitable for large classes where instructors seek grading efficiency without compromising question validity. In contrast, “open-ended” instruments are particularly suited to smaller classes, where instructors have more time to access the student’s reasoning and knowledge construction. Nevertheless, it is
advantageous to use a range of assessment tasks so as to promote the development of as many facets of conceptual understanding in students as possible so that they progressively develop from being novices into experts [1]. As students do not acquire such expert knowledge and skills automatically, instructors need to explicitly teach them as part of all course curricula. In this way, assessment would be used as a powerful “bridging-the-gap” tool [2] for promoting the development of conceptual understanding and related cognitive skill competencies.

The next article of this column will consider the nature and assessment of students’ interpretation and visualization of external representations in biochemistry and molecular biology. In this regard, we will aim to show how assessment could be used to promote students’ skills necessary for interacting with the array of visual representations that communicate our science.

Acknowledgments—The authors thank Judith G. Voet (Swarthmore College, Pennsylvania) and two anonymous referees for their valuable critique of this article, and are grateful to Harold B. White (Univ. Delaware) and J.-H.S. Hofmeyr (Swarthmore College, Pennsylvania) and two anonymous referees that communicated our science.

REFERENCES

APPENDIX: BRIEF, SUGGESTED SOLUTIONS TO ASSESSMENT TASKS

Qu. 1: Creatine phosphate (H.B. White, personal communication).

Qu. 2: Tetrahydrofolate (THF) is an essential cofactor in the synthesis of thymidine monophosphate (dTMP) and purines. Methotrexate, a tight-binding analog of dihydrofolate (DHF), strongly inhibits human dihydrofolate reductase (DHFR), causing the conversion of THF to DHF and the accumulation of DHF. This has a cascade effect on other folate-dependent reactions that can only use THF. Without dTMP and purine nucleotide synthesis, DNA synthesis is disrupted. Thus cancer cells and other rapidly growing cells are most severely affected by methotrexate. Trimethoprim also inhibits DHFR with similar effects in bacteria and is thus used as an antibiotic for certain bacterial infections. The antibiotic, sulfanilamide, has a different mode of action. It is a structural analog of para-aminobenzoate (PABA), a biosynthetic precursor of folate in bacteria. Thus, it inhibits the synthesis of folates (H.B. White, personal communication).

Qu. 3: The following 24 terms correspond to each of the spaces in the concept map for insertion in a left-to-right direction from top to bottom, in sequence of the hierarchical position of each space on the map—ferrohemoglobin, oxyhemoglobin, blood, heme catabolism, purple, red, CO, O₂, lungs, incomplete combustion, CO₂, carboxyhemoglobin, respiration, cigarette smoke, H₂O, pink, fires, car exhaust, ATP, CO poisoning, death, peripheral tissues, dizziness, severe headaches (H.B. White, personal communication).

Qu. 4: Choice (ε). Diffusion occurs in living systems all the time. Diffusion is a random but underlying process responsible for many emergent (macroscopic) biological phenomena. Therefore, ADP “movement toward” ATP synthase is governed by the random process of diffusion rather than by any other artificial “driver” (see Garvin-Doxas and Klymkowsky [6]).

Qu. 5: See text.

Qu. 6: For a step to be truly rate-limiting, that is, have a flux-control coefficient of 1, it must be insensitive to changes in its substrate and product concentrations. This implies that changes in the activity of any other enzyme in the pathway, which would cause changes in substrate and product concentrations, cannot be communicated to the rate-limiting step. In the language of control analysis this means that the elasticity coefficients of the rate-limiting step with respect to its substrates and products must approach zero; this will be obtained if the enzyme is saturated with substrate, is far from equilibrium, and is not subject to product inhibition. If the enzyme catalyzes the committing step of a pathway, the substrate of which is well-buffered, then a zero product elasticity suffices to make it rate-limiting (J.-H.S. Hofmeyr, personal communication).

Qu. 7: The ratio of flux-control coefficients of the two steps depends on the values of the elasticity coefficients of the two steps with respect to the metabolite “X” that links the two reactions (i.e. the sensitivities of their rates with respect to changes in the concentration of “X”) as follows: \( \frac{C_F}{C_P} = \frac{v}{x} \) (in terms of absolute elasticity values). An elasticity ratio of 2/3 will give the required distribution of flux control. This demonstrates that the flux-control coefficient of a particular step depends on the properties of all the steps in the system (J.-H.S. Hofmeyr, personal communication).

Qu. 8: The rate of a near equilibrium step is extremely sensitive to changes in its substrate and product concentrations. This means that their elasticity coefficients with respect to its substrates and product are very high. Any substrate or product elasticity contains a thermodynamic and a kinetic term. When near equilibrium, the thermodynamic terms approach plus infinity (substrate) and minus infinity (product) and, therefore, dominate the kinetic terms, which can be neglected. The absolute value of the ratio of substrate and product elasticities then approaches one. Applying this condition to the mathematical expression of control coefficients in terms of elasticity coefficients results in a zero flux-control coefficient, as well as zero concentration control coefficients of the near-equilibrium. This is a good example of where control analysis can demonstrate something that is difficult to explain by using physical arguments. Note also that the statement that a near-equilibrium reaction cannot control the flux must be viewed with caution. The degree to which it is out of equilibrium must always be compared with that of the other steps in a system. Again, control analysis provides the tools with which to do this (J.-H.S. Hofmeyr, personal communication).
Bridging the Gap

Bridging the Educational Research-Teaching Practice Gap

FOUNDATIONS FOR ASSESSING AND DEVELOPING BIOCHEMISTRY STUDENTS’ VISUAL LITERACY

Received for publication, March 25, 2010, and in revised form, June 25, 2010

Konrad J. Schönborn‡ and Trevor R. Anderson§¶

From the ‡Department of Science and Technology (ITN), Division of Media and Information Technology, Linköping University, Sweden, §Science Education Research Group (SERG), School of Biochemistry, Genetics and Microbiology, University of KwaZulu-Natal, Pietermaritzburg, South Africa

External representations (ERs), such as diagrams, animations, and dynamic models are vital tools for communicating and constructing knowledge in biochemistry. To build a meaningful understanding of structure, function, and process, it is essential that students become visually literate by mastering key cognitive skills that are essential for interpreting and visualizing ERs. In this article, first we describe a model of seven factors influencing students’ ability to learn from ERs. Second, we use this model and relevant literature to identify eight cognitive skills central to visual literacy in biochemistry. Third, we present simple examples of tasks as a foundation for designing more sophisticated and complex items for assessing and developing students’ visual literacy. We conclude that visual literacy is fundamental to the development of sound conceptual understanding and it is crucial to develop visual skills in parallel with meaningful learning outcomes in all biochemistry curricula.

Keywords: Visual literacy, developing interpretation and visualization skills, assessment, external representations, meaningful learning.

In two previous “Bridging the Gap” articles [1, 2] we focused on the assessment of various cognitive skills necessary for conceptual understanding. In addition, biochemists require other cognitive skills for visualizing and interpreting the myriad of external representations that communicate our science [3]. We use external representations (ERs) to describe the range of visual tools used to communicate scientific knowledge in the external world (e.g. [4]). ERs can be static or dynamic and include diagrams, pictures, physical models, animations, simulations, multimedia, and virtual realities. In the submicroscopic, abstract world of biochemistry, interpreting ERs is often the key to unlocking a meaningful conceptual understanding of structure, function, and process [3]. However, the ERs that we use can be graphically complex in terms of their constituent symbolic language and therefore difficult for students to interpret. Thus, it is crucial to develop our students’ visual literacy [3] to scaffold our students along the novice to expert continuum [1]. According to Bamford [5], visual literacy encompasses the skills required to read and write visual or symbolic language including the ability to, (i) decode and interpret ERs, (ii) encode and construct meaningful ERs, (iii) visualize objects in the “mind’s eye” and, (iv) comprehend ERs generated by others.

Voet and Voet [6] have pointed out the importance of visual literacy in modern biochemistry in light of an improved ability to visualize and study protein structure and function with modern computer-based technologies (e.g. [7]). Examples of useful educational resources include a java-based visualization environment constructed by Bottomley et al. [8] and a Jmol resource for interactive molecular visualization by Herraéz [9]. More recently, Hodis and Sussman [10] have developed an open resource (wiki) called Proteopedia that links descriptive text to manipulatable 3-D structures. Although numerous other visualization resources continue to enter the biochemistry education scene, only limited empirical research exists on students’ interpretation and visualization of ERs in our science. Recent studies include students’ interpretation of, and learning with, physical models (e.g. [11, 12]), static pictures (e.g. [13, 14]), animations (e.g. [15, 16]), and virtual environments (e.g. [17]). Although the increase in the number of studies in the field is encouraging, little attempt has been made to identify the specific cognitive skills associated with expert-level visual literacy, let alone the assessment thereof. Avgerinou and Ericson [18] support the urgent need to explicitly distill such skills and surmise that, “higher order visual literacy skills do not develop unless they are identified and taught” (p. 288).

This paper is available online at http://www.20436.org DOI 10.1002/bmb.20436
FIG. 1. Model of seven factors, including four interactive factors, affecting students’ ability to interpret and visualize ERs in biochemistry [4].

To induce “bridging the gap” between empirical science education research on visual literacy and its actual application to improving learning and teaching in biochemistry, this article addresses the following three questions:

- What factors affect students’ ability to interpret and visualize ERs in biochemistry?
- What cognitive skills are central to the visual literacy of expert biochemists?
- What simple examples of tasks can offer points of departure for assessing and developing students’ visual literacy in biochemistry?

FACTORS AFFECTING STUDENTS’ ABILITY TO INTERPRET ERs

Our research has modeled at least seven factors that affect students’ ability to interpret, visualize, and learn from ERs in biotechnology [4]. These factors may be important to consider when investigating the notion of an expert-level visual literacy in biochemistry. The factors of the model are expressed in the form of a Venn diagram presented in Fig. 1.

The conceptual factor (C) of the model represents a student’s conceptual knowledge, of relevance to an ER, whereas the reasoning factor (R) encapsulates the repertoire of cognitive skills that a student might utilize when interpreting and visualizing an ER. The representation mode factor (M) characterizes the actual nature of the ER, including the symbolic language that composes the ER. As shown in Fig. 1, these three factors are interdependent in that students cannot use their cognitive skills to do any reasoning (sense-making) without something to reason with, namely the ER (Factor R-M) and/or their own conceptual knowledge (Factor R-C). Thus, R-M represents a student’s ability to decode the symbolism making up the ER, whereas R-C represents the ability to employ the appropriate conceptual knowledge necessary for interpreting the ER. The (C-M) interactive factor represents the scientific (propositional) knowledge represented by the ER and its constituent symbolism. Finally, the (C-R-M) interactive factor encompasses a student’s simultaneous engagement of all the factors essential for the successful interpretation of an ER [4].

This model has important implications for biochemistry learning and teaching. First, it prompts instructors to realize that a certain minimum amount of prior conceptual knowledge (C) is indispensable to the interpretation of an ER. If such knowledge is lacking, or unsound due to alternative conceptions, then the student will compromise their ability to learn from the ER and successfully acquire the knowledge it represents (C-M). Thus, it is important that instructors establish the state of students’ prior knowledge before they expose them to particular ERs. Second, the model reminds us that learning from an ER is highly dependent on the nature and quality of the ER and its constituent symbolism (M) in effectively representing the scientific knowledge that it intends to represent (C-M). Here, an important message for instructors is to evaluate the soundness of an ER before exposing students to it. Indeed, a major area of visualization research consists of identifying what criteria are important when designing pedagogically effective ERs (see [19]). Furthermore, even if the ER is sound, instructors should also confirm whether it is intelligible to students and, if not, explicitly explain the meaning of the constituent symbolism as well as the limitations of the ER. Third, the model reminds instructors that having the necessary prior conceptual knowledge (C) and a highly effective ER (M) is still insufficient if students’ lack the cognitive skills (R) to both engage the appropriate conceptual knowledge necessary for interpreting the ER (R-C), and to decode the symbolic language used in the ER (R-M). Thus, it is of utmost importance for instructors to develop such cognitive skill competence (R) in their students, a fundamental goal that is the focus of the next section.

IDENTIFICATION AND ASSESSMENT OF COGNITIVE SKILLS CENTRAL TO EXPERT VISUAL LITERACY

In this section, we address the second and third questions raised in this article namely, what cognitive skills (R-C and R-M factors, Fig. 1) are central to the visual literacy of expert biochemists, and what tasks can assess and develop students’ visual literacy in biochemistry?

A synthesis of literature from the past 10 years has led to the identification of eight visual skills (Table I) associated with the notion of an expert visual literacy. Although some of the skills overlap in terms of objectives, we have purposely kept them separate so that instructors can more easily develop and assess each individual competency in their students. In so doing, one should recognize that no task can exclusively assess a single skill. As will be shown in this article, all tasks require students to simultaneously engage more than one of the skills listed in Table I, as well as several other cognitive skills, discussed elsewhere (See [1, 2]). In the interests of clarity, we have purposely pitched the tasks presented in this article at an introductory biochemistry level, with the idea that instructors could use them as a basis for designing more sophisticated tasks for higher educational levels and different biochemistry contexts. All the tasks were trialed in an introductory course on protein structure and function. Student responses to each task were screened to establish whether the particu-
lar visual skill was being tested. Preliminary findings suggest that our goals are being achieved although an empirical educational research study is required to fully validate the tasks. In this regard, we are currently conducting clinical interviews to meta-tag these and several other more sophisticated questions for assessing and developing visual skill competence in students. Such studies are beyond the scope of the present article.

**Decode the Symbolic Language Composing an ER**

All ERs are composed of symbolic language that needs decoding during the interpretation of an ER (Table I). In biochemistry, ERs can be particularly challenging for students to decode because of the great diversity of often idiosyncratic symbolism that may look aesthetically pleasing to viewers but does not always convey the intended scientific meaning [3]. For example, the same concept may be represented by several different symbols or, one symbol might be used to represent several different concepts (e.g. [13]). Consequently, there is extensive evidence for student difficulties with the decoding of the symbolism in ERs (Table I). For example, some students misinterpret ERs by focusing only on certain salient markings (e.g. brightly colored symbols) at the exclusion of others (e.g. [14]). Furthermore, where ERs are part of assessment tasks, a poorly designed ER, and not necessarily student conceptual knowledge, may be the cause of incorrect responses [41]. In the case of animations, the interpretation of moving graphical markings within restricted times can also create cognitive difficulties for students [34].

Despite these problems, many instructors make little effort to specifically explain an ER and its symbolism to students as they (mistakenly) assume that, because the ER is clear to them, the same will automatically hold for their students [19]. Thus, it is important to develop students’ skills for decoding ERs [25, 41]. One way to achieve this is to give students tasks that require them to use symbolism “keys” as a tool for decoding ERs. For example, the Protein Chart of Garratt and Orengo [42] can be used to interpret ribbon symbolism representing major domains (e.g. β-barrels), motifs (e.g. leucine zipper), and oligomeric proteins (e.g. proteosome) and to link such structures to cell functions. Another way to develop students’ symbolic language and skills is to give them extensive practice at answering formative assessment tasks that require them to decode symbolism, as well as combinations of symbolism, inherent in ERs from a wide range of topic areas. Qu. 1 is a simple example of such a task. Note that decoding a ball-and-stick representation does not only involve a surface-level perceptual process (factor R-M), but also the engagement of conceptual knowledge represented by the symbolism (factor R-C) [4].

**Qu. 1: Consider the peptide represented in Fig. 2.**

(a) Describe what the different “balls” and “sticks” represent in terms of peptide structure.

(b) Label the N- and C- termini of the peptide.

(c) How many peptide bonds are present in the structure?

**Evaluate the Power, Limitations, and Quality of an ER**

Closely related to the decoding of symbolism is the importance of developing students’ ability to evaluate the representational power, limitations and overall quality of ERs (Table I). Representational power is about how successfully a particular ER achieves its intended goal(s) (factor C-M), whether this be helping to develop a mental model of a concept, structure or process, or as a tool for solving a problem. As a first step in evaluating the power of the ER, it is crucial to ascertain what the goals of the ER are by deducing the limitations of the ER in terms of what parts of the phenomenon are, and are not, represented by the ER. Such limitations are not necessarily a weakness of the ER as the real power of a model is often (but not always) in its simplicity rather than its complexity (for example, Fig. 2 only depicts selected features of peptide structure). Thus, the aim of ER designers should be to develop multiple representations (Table I) of a phenomenon with each representation depicting selected features of the entire phenomenon in as clear a manner as feasible. Having established the goals of an ER, the

---

**Table I**

<table>
<thead>
<tr>
<th>Visual literacy encompasses the ability to:</th>
<th>Selected references for further reading:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decode the symbolic language composing an ER</td>
<td>13–15, 20, 21</td>
</tr>
<tr>
<td>Evaluate the power, limitations, and quality of an ER</td>
<td>4, 7, 15, 22, 23</td>
</tr>
<tr>
<td>Interpret and use an ER to solve a problem</td>
<td>4, 11, 16, 17, 20</td>
</tr>
<tr>
<td>Spatially manipulate an ER to interpret and explain a concept</td>
<td>12, 17, 24–26</td>
</tr>
<tr>
<td>Construct an ER to explain a concept or solve a problem</td>
<td>4, 23, 27–29</td>
</tr>
<tr>
<td>Translate horizontally across multiple ERs of a concept</td>
<td>12, 15, 21, 30, 31</td>
</tr>
<tr>
<td>Translate vertically between ERs that depict various levels of organization and complexity</td>
<td>30, 32–35</td>
</tr>
<tr>
<td>Visualize orders of magnitude, relative size, and scale</td>
<td>36–40</td>
</tr>
</tbody>
</table>

**FIG. 2.** ER depicting a peptide in a conventional “ball-and-stick” format. Display generated with Viewer-Lite 5.0, Accelrys Software Inc.
second crucial step in determining the power of an ER is to evaluate how accurately and effectively the ER, and its constituent symbolism, conveys the intended scientific knowledge so that these goals are achieved. Indeed, as already discussed, symbolism can often be confusing and misleading and significantly affect the overall quality of the ER. Hence, it is important to develop students’ representational competence [43] by giving them practice at answering assessment tasks that require critical evaluation of the quality of (“good” and “bad”) ERs. Qu. 2 is a simple example of such a task. Other more sophisticated tasks could of course involve the use of molecular viewing software as well as animated and stereo tools.

Qu. 2: Consider the ER in Fig. 2 and list which structural features of the peptide:

(a) Are represented by the ER.
(b) Are not represented by the ER.
(c) Comment on how clearly you think the ER and its constituent symbolism, represents the structure of the peptide.

Interpret and Use an ER to Solve a Problem

Interpretation of an ER can only occur once all the symbolic language composing the ER has been decoded (R-M). Once this has occurred, the individual can construct a mental model of the ER as an integrated whole and can interpret it by linking it (R-C) to the conceptual knowledge it represents (C-M). Having interpreted the ER, the individual is then in a position to use the ER for a range of tasks including, solving a (novel) problem, making a prediction, or constructing new conceptual knowledge (Table I). Qu. 3 is a simple example of a task that uses an ER to assess and promote the development of problem-solving skills in biochemistry.

Qu. 3: Consider the peptide represented in Fig. 2:

(a) Identify the amino acid residues from the N- to the C-terminus.
(b) How many groups would be protonated at a pH of 7.0?
(c) How many double bonds are there in this peptide? Identify them.
(d) Which residues in the peptide can display hydrophilic properties? Explain why.

Spatially Manipulate an ER to Interpret and
Explain a Concept

Much chemistry education research has shown that spatial visualization skills are essential for interpreting 2-D ERs that portray 3-D objects, and that students show extensive difficulties in performing these processes [41]. In biochemistry, only a limited number of educational studies have been done on spatial visualization skills (e.g. [26]) despite the diversity of complex biomolecules that require visualization via various computer and physical ERs. In addition, spatial skills are required for interpreting, inter alia, animations of cellular processes, electron micrographs of cellular structures, and Cartesian graphs [4].

Tuckey and Selvaratnam [44] identified several skills as being important for the visualization of 3-D molecules from 2-D ERs. These include the ability to decode symbolic depth cues, understand the spatial relationships (width, depth, and height) represented in the 2-D ER, mentally manipulate an ER in the “mind’s eye,” and visualize its transformation. Assessment tasks that make use of modeling software offer the opportunity to develop such skills in our students. Moreover, modern virtual reality environments that incorporate haptic feedback (see [17]) provide a unique alternative for developing spatial skills through the sense of touch. Related to this, Roberts et al. [12] have shown that students can develop spatial skills through the tactile manipulation of physical models, while Harris et al. [26] have shown that a combination of physical and computer-generated ERs can achieve this goal. Importantly, a recent article [7] on macromolecular visualization points out that research is only starting to establish how best to use physical models in structural biology education. As a further step in this direction, albeit in a simple form, Qu. 4 is an example of a task that could be used to develop students’ various spatial skills including, depth perception (Qu. 4a), the visualization of the relative configuration of atoms (Qu. 4b), and the mental rotation of molecules (Qu. 4c). Note that such tasks could of course be extended for use within more sophisticated dynamic modeling environments.

Qu. 4: Consider the peptide represented in Fig. 2:

(a) Which one of the oxygen atoms would be closer to you in the present orientation?
(b) Identify whether each of the two peptide bonds is in cis or trans configuration.
(c) Sketch what the peptide would look like after a 180° rotation about the y-axis.

Construct an ER to Explain a Concept or
Solve a Problem

Biochemists make extensive use of ER-construction as a problem-solving tool (Table I) to capture a research method overview (e.g. as a flow diagram), illustrate mechanisms of cellular regulatory processes (e.g. as an animation), and model 3-D structures. In addition, biochemists might modify or manipulate an ER to, for example, predict what affect an activator or inhibitor might have on the bioactivity of a protein. Thus it is important to develop our students’ ability to construct, modify, and use their own ERs as part of their practice as a biochemist. To achieve this, students should receive extensive opportunities to perform some of the above-mentioned activities. In addition, simple hand-drawing tasks, or exercises requiring the construction of physical models of biomolecules using various materials (see [23, 24]) are ways of developing this competence. An exciting innovation that could also promote ER construction skills is the online protein structure game called Foldit [45]. Foldit requires participants to use various visual skills to fold a ribbon representation of a random coil into its most stable conformation.
Qu. 5 is an example of a simple task that could also be used to develop ER-construction skills, as well as other visual skills (Table I), ranging from decoding (Qu. 5a; Factor R-M) through to ER interpretation (Qu. 5b and 5c) during which adequate conceptual knowledge needs to be engaged (Factor R-C).

Qu. 5: Examine the following representation of a polypeptide sequence:


(a) Use a sketch to predict how this peptide might fold so that the three indicated amino acids are brought into close proximity.

(b) Suggest what type of secondary structure might result from the folding. Explain why.

(c) Suggest what effect(s) could lead to an interaction between the three amino acid residues.

Translate Horizontally Across Multiple ERs of a Concept

It is common practice in the submicroscopic world of biochemistry to use different ERs to represent the same phenomenon or different features of a phenomenon [3]. For example, the concept of enzyme-substrate interaction can be represented by a wide range of different ER modes, including abstract (e.g. Michaelis-Menten formula), symbolic (e.g. chemical equation), graphical (e.g. Michaelis-Menten plot), stylized (e.g. animation of enzyme-substrate binding), and realistic modes (a crystal structure). To develop an integrated understanding of this concept, students would need to “move” or translate across these different ERs, including decoding the different symbolism in each ER and linking the relationships between each ER [46]. Schönborn and Bögeholz [30] have built on the idea of translation in a biology context by defining a horizontal translation, which deals with interpreting ERs that represent a phenomenon at the same level of biological organization, whether it be at the macro-, micro- or molecular level. Qu. 6 is an example of a simple task aimed at developing students’ horizontal translation skills, in this case involving ERs that all depict the molecular level of biological organization (Fig. 3). In this example, translating across all three ERs facilitates the development of a more complete mental model of the structure of the subunit (Qu. 6). Clearly, students will also make use of other skills presented in Table I during this process.

Qu. 6: Consider the left, central, and right ER in Fig. 3 in which each represents different structural features of one beta subunit of human carbonmonoxy hemoglobin. Interpret each of the three ERs and compare them with respect to what they do, and do not represent in terms of the various structural features of hemoglobin.

Translate Vertically Between ERs that Depict Different Levels of Organization and Complexity

All biochemists require the skills to translate vertically “between” levels of biological organization and
complexity [30]. For example, in the case of a biological tissue, it might be necessary to translate between ERs that represent the visible and tangible macro-level, the microscopic level, as well as the molecular level of biological organization [35]. The “distance” of vertical translation required may not always involve a complete macro-micro-molecular transition but could also comprise smaller distances between levels of complexity [30]. For example, Qu. 7 requires students to remain within a molecular domain but to still translate between different levels of complexity.

Various studies have shown that students have difficulties developing the vertical translation skills that characterize expert-level visualization (e.g. [32, 33]). Thus, practice in doing tasks, such as Qu. 7, that explicitly test such skills could improve student competence in this area. In this example, translating vertically between the three ERs allows students to gain a greater appreciation of the additive effect of “moving” through increasing levels of biological complexity.

Qu. 7: Consider the top, central, and bottom ERs in Fig. 3 representing various structural features of human carbonmonoxy hemoglobin. Interpret the three ERs and compare them with respect to the different levels of complexity that they represent.

Visualize Orders of Magnitude, Relative Size, and Scale

Visualizing orders of magnitude and scale is related to vertical translation processes since the latter often includes grasping the absolute and relative size of organs, cells, organelles, and biomolecules. However, competence in the former may achieve much more since it also enables the visualization of the relative quantity of a wide range of biostructures and parameters of relevance to living systems. These could include the approximate number of mitochondria or nuclei in different cell types, the typical concentrations of biomolecules necessary to regulate metabolic systems or to sustain life without reaching toxic levels, as well as the typical magnitude of kinetic constants (e.g. $K_m$, $V_{max}$) and thermodynamic values (e.g. Gibbs energy, redox potentials) that would be realistic in a living system. Visual skills are also required for interpreting log scales used in graphical representations of assays, while the properties of many substances (e.g. enzymes) in living systems change as the scale of parameters such as pH and temperature change. Finally, understanding nanotechnology also depends heavily on skills for visualizing scale (see [22, 37]) while in biology, scale is considered a “threshold” concept in that it is seen as an essential prerequisite for mastering other concepts [36]. Qu. 8 is an example of a simple task that could develop and assess this competency in students.

Qu. 8: Arrange the following structures in order of decreasing size and match each of them with their approximate diameter:

- Proteasome, glucose molecule, red blood cell, water molecule, mitochondrion, ribosome, hemoglobin molecule, typical bacterium, DNA double helix, typical virus
  - Diameters: 7 $\mu$m, 1–3 $\mu$m, 0.5–1.0 $\mu$m, 80 nm, 25 nm, 11–15 nm, 6.4 nm, 2.4 nm, 1 nm, and 0.3 nm.

CONCLUDING REMARKS

The eight skills discussed in this article should not be viewed as a complete set of competencies required for optimal visual literacy as numerous other (and often complementary) cognitive skills may influence the visualization process (e.g. [1, 2, 35]). However, by developing competence in these skills, we believe that students will go a long way toward optimizing their ability to interpret and use ERs as effective knowledge-building and communication tools. Clearly, although the success of this endeavor will depend heavily on the nature of the formative and summative assessment tasks [47] that instructors design to respectively develop and grade each skill competence. In this regard, this article provides examples of simple tasks for the assessment of each visual skill that instructors could use as a foundation for the development of more sophisticated tasks. Unfortunately, educators often place little emphasis on actively teaching the skills necessary for interpreting and visualizing ERs [18]. Since visual literacy is fundamental to the development of sound conceptual understanding, a central pedagogical goal of all biochemistry instructors should be to teach and assess students’ visual skills in parallel with the development of all learning outcomes.

- In summary, the main messages conveyed in this article are as follows:
  - Ensure that the ER is an accurate and sound representation of the intended knowledge (C-M);
  - Confirm whether the ER is clear and intelligible to students; if not, explain the ER and its limitations;
  - Ensure that students have the appropriate conceptual knowledge to interpret the ER;
  - Check which of the three factors—soundness of an ER (M), prior conceptual knowledge (C), or cognitive skill competence (R) are limiting, if students show difficulties interpreting the ER, and take appropriate action;
  - Check whether students’ reasoning (R) difficulties are due to inappropriate decoding of the symbolic language composing the ER (R-M), or to inadequate engagement of their conceptual knowledge of relevance to the ER (R-C), and take appropriate action;
  - Design assessment tasks that aim to specifically assess students’ competence in each of the identified visual skills (Table I). Validate each task by means of student interviews to ensure that they are actually assessing the particular visual skill;
  - Use the tasks to formatively develop students’ visual literacy during a course, and to summatively assess their attainment of such competencies at the end of the course;
  - Integrate visual literacy development into all courses across the biochemistry curriculum.
Acknowledgment—we thank Duane Sears and Lizanne Huy-samen for permission to use adaptations of their tasks (in Qu. 3 and Qu. 5, respectively), and Richard Garratt and Gunnar Höst for valuable discussions.

REFERENCES


Qu. 1: Each “ball” represents the center of an atom and each “stick” represents the length and orientation of the covalent bond between two atoms. The grey, red, and blue balls represent carbon, oxygen, and nitrogen atoms, respectively. (b) The C-terminus is located at the carboxyl group represented by the only carbon (grey) atom bound to two (red) oxygen atoms. The N-terminus is the amino group represented by the nitrogen (blue) atom joined to a single alpha carbon atom. (c) 2.

Qu. 2: (a) The following are represented—carbon, oxygen, and nitrogen atoms; covalent bonds between atoms; bond angles; peptide bonds; an aromatic ring; amino acid residues; amino acid side chains; the overall spatial configuration of the actual positions of each atom relative to another. (b) The following are not represented—hydrogen atoms; double bonds in the keto- and carboxyl groups and the aromatic ring; the partial double bond character of the peptide bond; hydrogen bonds; partial charges on the atoms; the surface topography and volume occupied by atoms, groups and the whole peptide molecule. (c) The symbolism in the ER clearly represents the structure of the peptide with the exception that double bonds are represented in an identical mode to that of single bonds. By representing only selected features of the peptide, the ER is simpler and less complex and, therefore, clearer, and easier to interpret. However, multiple ERs are required to represent all features of the peptide to build a more realistic and integrated mental model of the peptide.

Qu. 3: (a) Ser-Lys-Tyr. (b) 1 (alpha NH$_3^+$) + 1 (epsilon NH$_3^+$) = 2 protonated groups. (c) Two C=O (peptide bonds) + one C-terminal carboxyl + three C=C (Tyr ring) = 6 double bonds. (d) All three residues can show hydrophilic properties: Water can bind via hydrogen bonds to the epsilon NH$_3^+$ of lysine, the alpha NH$_3^+$, and hydroxyl group of serine, and the alpha carboxyl and hydroxyl group of tyrosine.

Qu. 4: (a) The hydroxy oxygen atom of the tyrosine residue. (b) The peptide bond between Ser and Lys is in cis configuration and the peptide bond between Lys and Tyr is in trans configuration.

Qu. 5: (a) The peptide would most likely fold into an “S” shape β-meander with the flexible P-G-G-P and G-G-G regions forming beta turns. (b) The folding would lead to the formation of anti-parallel beta sheets. (c) Phe, ile, and Met could interact through the hydrophobic effect, which, by favoring hydrophilic interactions between solvent molecules, leads to hydrophobic clustering.

Qu. 6: Each ER represents structural features of one beta subunit that the other two do not. The wireframe ER on the left depicts the relative atomic coordinates (and corresponding covalent bonds) that define the overall spatial configuration of the subunit. In contrast, the ribbon ER in the center depicts the subunit’s overall conformation or tertiary structure as well as some secondary structure, consisting in this case of α-helices. The ER on the right portrays the molecular volume occupied by the subunit and therefore the size and shape of the structure’s molecular surface (also see [22]).

Qu. 7: The bottom ER depicts the non-protein prosthetic group found in all four subunits of hemoglobin. The ER in the center represents a higher level of complexity by displaying the location of the prosthetic group relative to the overall conformation of the subunit. The top ER adds a further level of complexity by representing how two subunits would be spatially associated to one-another and to each of their prosthetic groups, through noncovalent interactions.

Qu. 8: Red blood cell (7 μm), typical bacterium (1–3 μm), mitochondrion (0.5–1.0 μm), typical virus (80 nm), ribosome (25 nm), proteasome (11–15 nm), hemoglobin (6.4 nm), DNA double helix (2.4 nm), glucose molecule (1 nm), and water molecule (0.3 nm).
**Bridging the Gap**

**Bridging the Educational Research-Teaching Practice Gap**

**TOOLS FOR EVALUATING THE QUALITY OF ASSESSMENT INSTRUMENTS**

Received for publication, June 19, 2009, and in revised form, October 2, 2009

Trevor R. Anderson‡ and John M. Rogan

*From the Science Education Research Group (SERG), School of Biochemistry, Genetics and Microbiology, University of KwaZulu-Natal, Pietermaritzburg, South Africa*

Student assessment is central to the educational process and can be used for multiple purposes including, to promote student learning, to grade student performance and to evaluate the educational quality of qualifications. It is, therefore, of utmost importance that assessment instruments are of a high quality. In this article, we present various tools that instructors could use, both to improve instrument design and validity before presentation to students and, to evaluate the reliability and quality of the assessment after students have answered the questions. In keeping with our goals of the Bridging-the-Gap series, we also present various ideas from the educational literature regarding the evaluation of assessment quality and end with a list of criteria that instructors could use to guide their evaluation process.

**Keywords**: Assessment, quality, design, item analysis, validity, reliability.

There is a world-wide move in educational institutions towards the establishment of more formal structures for the evaluation of the quality of education [1]. This change has been motivated by the demands of various stakeholders in the educational process including, funding and accreditation bodies, potential employers of graduates, and the students themselves demanding qualifications that are recognized for their high educational standards, and which will thereby enable them to succeed in an ever-increasing competitive world. One way to achieve this goal is to evaluate the quality of the assessment instruments employed to assess the achievement of such outcomes. Evaluation of assessment quality also offers universities a more accurate means of evaluating faculty teaching and course quality [2] than student evaluation questionnaires which, despite often yielding unreliable measures of teaching quality [3], are the dominant instrument used to judge teacher success and promotions.

In this article in the “Bridging-the-Gap” series on assessment, we focus on procedures that can be employed to improve the quality of student assessment. The first part of the paper deals with steps that can be taken in the design of valid test instruments, while the second part deals with the calculation of reliability and the performance of an item analysis that can be done after the students have written the test. These procedures provide evidence for retaining good items, discarding bad ones, and generally feed back into the improvement of assessment design when the instrument is used the next time around.

**BEFORE ASSESSMENT—DESIGNING THE TEST INSTRUMENT**

**Issues of Validity**

The importance of validity has been stressed throughout this series of articles, if not by name, by example. Simply stated, if all the items in an assessment instrument are designed to match one or more objectives of a course, then there is evidence for the validity of that instrument. For example, three selected learning outcomes from a course on introductory metabolism taught by one of the authors are given in the box below, together with examples of test items that can be used both formatively to promote the development of an outcome during the course and summatively to check whether students have indeed achieved the outcome [2]. Note that, in line with our previous papers [4, 5], each outcome consists of both a cognitive skill (e.g. apply, transfer, or reason) and a knowledge component.
Examples of learning outcomes and corresponding assessment tasks for an introductory course on metabolism

**Outcome 1.** Apply the concepts of thermodynamics to explain metabolic regulation, including the direction of metabolic pathways; Explain why a metabolic pathway is able to proceed in a particular direction.

**Outcome 2.** Reason locally and globally about the concept (system thinking); Use the supplied metabolic charts to answer the following question. If the enzyme triosephosphate isomerase was irreversibly inhibited by a toxic substance, predict what effect this would have on:

(a) The flux through glycolysis;
(b) The concentration of the intermediates in glycolysis before the point of inhibition;
(c) The production of NADH by the Citric Acid Cycle during aerobic carbohydrate metabolism;
(d) The production of ATP per mole of glucose during anaerobic carbohydrate metabolism;
(e) The utilization of oxygen by the electron transport chain.

**Outcome 3.** Transfer and apply your knowledge of chemistry to understand metabolism; Use your knowledge of chemistry to suggest which factors determine whether two biochemical compounds will react in an enzyme-catalyzed reaction?

The task addressing Outcome 1 requires students to apply their knowledge of thermodynamics to explain under what circumstances all the consecutive-linked reactions in a pathway are able to proceed in the same forward direction. The task addressing Outcome 2 requires students to exercise their system thinking skills to think locally about the effects on the individual enzyme-catalyzed reaction that is inhibited and about the global implications of this inhibition on the functioning of glycolysis, in which the enzyme occurs, and other related pathways. In contrast, the task addressing Outcome 3 requires students to transfer and apply their knowledge of chemistry, studied in the previous year, to answer a question about metabolic reactions, something which, in our experience, students find very difficult to do as they tend to store such knowledge in ‘closed boxes’ and only transfer it if cued to do so [6].

From the above, it seems obvious that the assessment instruments designed by those who teach the course should automatically be valid—and this is usually the case. However, at times we sabotage our own goals by not thinking about metabolism; something which, in our experience, students find very difficult to do as they tend to store such knowledge in ‘closed boxes’ and only transfer it if cued to do so [6].

After the above, it seems obvious that the assessment instruments designed by those who teach the course should automatically be valid—and this is usually the case. However, at times we sabotage our own goals by not thinking about the effects on the individual enzyme-catalyzed reaction that is inhibited and about the global implications of this inhibition on the functioning of glycolysis, in which the enzyme occurs, and other related pathways. In contrast, the task addressing Outcome 3 requires students to transfer and apply their knowledge of chemistry, studied in the previous year, to answer a question about metabolic reactions, something which, in our experience, students find very difficult to do as they tend to store such knowledge in ‘closed boxes’ and only transfer it if cued to do so [6].

From the above, it seems obvious that the assessment instruments designed by those who teach the course should automatically be valid—and this is usually the case. However, at times we sabotage our own goals by not thinking about the effects on the individual enzyme-catalyzed reaction that is inhibited and about the global implications of this inhibition on the functioning of glycolysis, in which the enzyme occurs, and other related pathways. In contrast, the task addressing Outcome 3 requires students to transfer and apply their knowledge of chemistry, studied in the previous year, to answer a question about metabolic reactions, something which, in our experience, students find very difficult to do as they tend to store such knowledge in ‘closed boxes’ and only transfer it if cued to do so [6].

From the above, it seems obvious that the assessment instruments designed by those who teach the course should automatically be valid—and this is usually the case. However, at times we sabotage our own goals by not thinking about the effects on the individual enzyme-catalyzed reaction that is inhibited and about the global implications of this inhibition on the functioning of glycolysis, in which the enzyme occurs, and other related pathways. In contrast, the task addressing Outcome 3 requires students to transfer and apply their knowledge of chemistry, studied in the previous year, to answer a question about metabolic reactions, something which, in our experience, students find very difficult to do as they tend to store such knowledge in ‘closed boxes’ and only transfer it if cued to do so [6].

From the above, it seems obvious that the assessment instruments designed by those who teach the course should automatically be valid—and this is usually the case. However, at times we sabotage our own goals by not thinking about the effects on the individual enzyme-catalyzed reaction that is inhibited and about the global implications of this inhibition on the functioning of glycolysis, in which the enzyme occurs, and other related pathways. In contrast, the task addressing Outcome 3 requires students to transfer and apply their knowledge of chemistry, studied in the previous year, to answer a question about metabolic reactions, something which, in our experience, students find very difficult to do as they tend to store such knowledge in ‘closed boxes’ and only transfer it if cued to do so [6].

From the above, it seems obvious that the assessment instruments designed by those who teach the course should automatically be valid—and this is usually the case. However, at times we sabotage our own goals by not thinking about the effects on the individual enzyme-catalyzed reaction that is inhibited and about the global implications of this inhibition on the functioning of glycolysis, in which the enzyme occurs, and other related pathways. In contrast, the task addressing Outcome 3 requires students to transfer and apply their knowledge of chemistry, studied in the previous year, to answer a question about metabolic reactions, something which, in our experience, students find very difficult to do as they tend to store such knowledge in ‘closed boxes’ and only transfer it if cued to do so [6].

From the above, it seems obvious that the assessment instruments designed by those who teach the course should automatically be valid—and this is usually the case. However, at times we sabotage our own goals by not thinking about the effects on the individual enzyme-catalyzed reaction that is inhibited and about the global implications of this inhibition on the functioning of glycolysis, in which the enzyme occurs, and other related pathways. In contrast, the task addressing Outcome 3 requires students to transfer and apply their knowledge of chemistry, studied in the previous year, to answer a question about metabolic reactions, something which, in our experience, students find very difficult to do as they tend to store such knowledge in ‘closed boxes’ and only transfer it if cued to do so [6].

From the above, it seems obvious that the assessment instruments designed by those who teach the course should automatically be valid—and this is usually the case. However, at times we sabotage our own goals by not thinking about the effects on the individual enzyme-catalyzed reaction that is inhibited and about the global implications of this inhibition on the functioning of glycolysis, in which the enzyme occurs, and other related pathways. In contrast, the task addressing Outcome 3 requires students to transfer and apply their knowledge of chemistry, studied in the previous year, to answer a question about metabolic reactions, something which, in our experience, students find very difficult to do as they tend to store such knowledge in ‘closed boxes’ and only transfer it if cued to do so [6].
tors we need to be equally concerned with how accurately we are measuring the progress of our students. Reliability, when applied to psychometric measures, is an intuitive concept, which shares some of the same meaning as in the world of physical measurement. If a claim is made that the pH of a certain solution is 4.5, then we would want to know how accurate this measurement is (i.e. what is the reliability of the instrument used?), and in particular, the size of the error associated with this claim—for example 4.5 + 0.15. We would also want to know whether this result would be obtained consistently—whether we would get the same reading if we were to take the measurement again.

A caveat needs to be emphasized at this stage. The explanation of reliability below only applies to assessment instruments whose items are closed—where there is clearly one correct answer. The reliability of complex and open-ended assessment techniques cannot be calculated in the manner described below.

When it comes to finding the reliability of a test, we are measuring its consistency. Theoretically, if we gave the same test twice to the same group of students and each student received the same score in the second as in the first test, we could say that the test is reliable. This method of establishing reliability, correlating the first set of scores with the second, is known as the test–retest method. In real life students would likely remember some of the items and do better the second time around. Hence this method is seldom used in practice. Common calculations of reliability of assessment instruments are measured its consistency. Theoretically, if we gave the same test twice to the same group of students and we could say that the test is reliable. This method of establishing reliability, correlating the first set of scores with the second, is known as the test–retest method. In real life students would likely remember some of the items and do better the second time around. Hence this method is seldom used in practice. Common calculations of reliability of assessment instruments are calculated in the manner described below.

One important application of reliability is the calculation of the SEM. The scores that we give our students (obtained score) on a test or examination may be considered reliable to the extent that they reflect the students’ true scores—those scores which accurately assess students’ ability or performance if we were able to eliminate measurement from the performance of each individual on each item.

Reliability measures consistency not accuracy. A pH meter might consistently measure the pH of a solution incorrectly if it has not been calibrated. In terms of a test, validity ensures accuracy, while reliability deals with consistency. A test which is not reliable cannot be valid. However, a test which is reliable might not be valid as it might consistently be measuring the wrong objective.

**The Standard Error of Measurement**

One important application of reliability is the calculation of the SEM. The scores that we give our students (obtained score) on a test or examination may be considered reliable to the extent that they reflect the students’ true scores—those scores which accurately assess students’ ability or performance if we were able to eliminate all sources of error. The true score of a student may be thought of as follows:

\[ \text{True score} = \text{Obtained score} - \text{Error of measurement} \]

The error of measurement will vary from student to student, and indeed for the same student from day to day. For example, if a certain student wrote a test on Wednesday, it could be that the error of measurement for that student was small. However, say that on Thursday the student was involved in a car accident. If she were to take a test on Friday, it is likely that her troubled mental state would result in a larger error of measurement.

It is obviously not possible to know what the error of measurement is for each individual student each time they are assessed. However, for a given test, we can calculate an overall estimate of the error of measurement—a quantity known as the SEM. The SEM can be calculated as follows:

\[ \text{SEM} = s \sqrt{1 - \text{reliability}} \]

where \( s \) is the standard deviation.
Note that if the reliability is one, then the SEM becomes zero. An example of how the SEM may be interpreted is given below.

For a 100-item multiple-choice test, the average is 65, the standard deviation 15, and the reliability 0.5. The SEM can be calculated:

\[
SEM = 15\sqrt{(1 - 0.5)} \\
= 15(0.71) \\
= 10.65
\]

The score for a person getting 70 on the test could be expressed as 70 ± 10.65. This may be interpreted that there is about a 68% probability that this person’s true score lies somewhere between 60 and 80—in other words a person’s score is presented, not by a point, but by a range.

From the above example, it is clear that for a high-stakes examination, such as one that determines whether a student passes or fails a course, a high coefficient of reliability is desirable. When determining the fate of a student, looking at the range in which the student’s true score might lie is preferable to using a point score.

There are a number of factors that contribute to enhancing or diminishing the reliability of an instrument, and hence increasing the SEM. Some are random and thus beyond the control of the instructor. However, there are two that are important and within the control of an instructor. The one is that students know and understand the criteria on which their work will be assessed. The second is the quality of the items making up the instrument. A reliable instrument is of necessity made up of “good” items. Some of the statistical packages identify the most problematic item, and recalculate the reliability of the instrument if that item were to be deleted. However, no information is provided on why that item is problematic. For this an item analysis of each item is both necessary and desirable.

### Item Analysis

Item analysis provides one with certain characteristics of each item in the instrument. This information can be used to identify weak items and hence improve both the items and the instrument as a whole. We have found that undertaking this task with the help of colleagues is an invaluable exercise. As with the calculation of reliability, item analysis can only be performed once at least one group has answered the items in the instrument (see Nagata [17] for a full treatment of item analysis).

**Item Difficulty** — The index of item difficulty is simply defined as the proportion or percentage of students in a group getting the item correct. In the case of a weighted item (one in which more than one possible score is possible, e.g. 1, 2, 3, or 4) it is the average for that item expressed as a proportion or percentage. The ideal value of this index depends on the purpose of the test. For example, in a test that is designed to identify just a few winners from a large group (e.g. a science Olympiad), the index for most items will typically be quite low. However, for the normal semester test one would expect the indices of items to range between about 0.4 and 0.8. An index of say 0.1 would certainly raise a flag and cause one to examine that item carefully, and to ask the question, “Why are so many students unable to answer this item? Was this concept not taught?”

**Item discrimination** — The index of item discrimination is the difference between the proportion of the top one-third (approximately) of students getting that item correct minus the proportion of the bottom one-third of students getting that item correct. In other words:

\[
\text{Index of discrimination} = \text{Item difficulty}_{\text{top third}} - \text{Item difficulty}_{\text{bottom third}}
\]

The top and bottom thirds are usually selected on the basis of the test being analyzed, but could be selected on the basis of a cumulated semester grade. The index can range from 1 (all top students and none of the bottom students get the answer correct) to –1. A negative index indicates a problematic item—one on which the better students are performing worse than the poorer students. It is often an indication of a poorly worded or misleading item. Indices of between 0.2 and 0.4 are considered satisfactory. Note that an item with a difficulty index of one necessity will of have a discrimination index of zero. The same is true of an item with a difficulty index of zero.

<table>
<thead>
<tr>
<th>Student</th>
<th>Total score on the whole test</th>
<th>Item 1</th>
<th>Item 7</th>
<th>Item 13</th>
<th>Item 19</th>
<th>Item 23</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>95</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>90</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>87</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>82</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>80</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>78</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>75</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>H</td>
<td>72</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>I</td>
<td>65</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>J</td>
<td>62</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>K</td>
<td>60</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>L</td>
<td>60</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>M</td>
<td>58</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>N</td>
<td>56</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>O</td>
<td>54</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P</td>
<td>52</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Q</td>
<td>49</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>R</td>
<td>46</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>S</td>
<td>45</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>45</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>U</td>
<td>41</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>V</td>
<td>40</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>W</td>
<td>38</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>X</td>
<td>35</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: A–H constitutes the top group and Q–X the bottom.
the correct option. It involves calculating the proportion or percentage of students who select each of the distracters as well as whether each distracter is achieving its intended purpose. Distracter analysis offers a means of establishing the quality of the item, rendering it easier than it should be. Distracters may be obtained from various sources. Ideally, the distracters should stem from misconceptions that have already been identified by research on students, usually through open-ended questions and/or clinical interviews [18]. If this is not possible, then misconceptions published in the educational literature that have been identified in other student contexts may also be used as distracters. However, as not all misconceptions are universal, some of them may need to be discarded if very few of your students select them. It is also acceptable to derive distracters from student difficulties that have been identified based purely on intuition and teaching experience. Nonsensical distracters should be avoided as they can be confusing to students and should in any case be eliminated by item analysis. Thus, it is important that distracters be authentic and plausible to a reasonable number of students. Highly unlikely distracters will lower the quality of the item, rendering it easier than it should be. Distracter analysis offers a means of establishing whether each distracter is achieving its intended purpose and contributing to the overall quality of the item. It involves calculating the proportion or percentage of students who select each of the distracters as well as the correct option.

### Example of a multiple-choice item

(A hypothetical percentage of students selecting each statement is given in parenthesis)

Indicate which of the following statements are correct descriptions of a spontaneous enzyme-catalysed reaction. Circle the number(s) next to right answer(s) corresponding to correct statement(s) and give reasons for your choice(s).

A spontaneous enzyme-catalysed reaction:

1) Always proceeds immediately on its own without external stimuli, other input or driving forces; [15%]
2) Occurs rapidly; [13%]
3) Has the potential to occur but might not occur immediately; [62%]
4) Always proceeds with the release of energy because it is exergonic (ΔG is negative); [1%]
5) Occurs without the need for activation energy. [9%]

Reasons...

The multiple-choice question above, is probing student understanding of the thermodynamic concept of spontaneity. As spontaneity is not about kinetics, it does not tell us about when (e.g. immediately) or how rapidly a spontaneous reaction will occur; only that it has the potential to occur in the forward direction, when all the necessary requirements, including activation energy, for reaction have been met. As shown in the box, 62% of students selected the correct option 3, suggesting that this statement is clear and intelligible to a majority of students. In contrast, only 1% selected distracter 4, suggesting that this statement, seeming obviously wrong to most students, does not contribute to the quality of the item and should be replaced with a more plausible option. The remaining distracters are clearly achieving their purpose in exposing misconceptions among a significant number of the students and are, therefore, making an important contribution to the quality of the item.

A Final Thought on Item Analysis—Item analysis is a useful instrument provided it is used judiciously rather than mechanically. The procedure points to items which might require further scrutiny and modification. It would be a mistake to make a decision such as, “I will discard all items that 75% of the students fail.” There might be a way of modifying some of these items in such a way that more students are able to get the right answer. There might also be sound reasons for retaining some items that most students fail, as these items probe the understanding of some important principles that students should know. Retaining them sends an important message to students that there are concepts that they cannot ignore.

### CONCLUSIONS

**Why Bother with Things Like Validity and Reliability?**

Clearly, the development of valid assessment instruments and the calculation of their reliabilities is a time-
It is conceivable that once a microscope was developed, research might show our instrument lacks reliability. If it were taken to develop the instrument and what was done to ensure its validity and its reliability. Most peer-refereed journals in science education will require this information for any article, where the results and conclusions are based on an assessment instrument.

Fourth, collaborating with colleagues, preferably in group sessions, in both the design of valid assessment instruments and the analysis of student results (e.g. reliability and item analysis) can be a stimulating form of professional growth and development [2] that can lead to a common teaching philosophy, improved cohesion between courses, more logical progression between educational levels, and comparable standards across programs. To fully realize the potential that the design and analysis of assessment instruments, and indeed a complete department-wide assessment program, has for collaboration and professional growth, we conclude by offering some sample discussion points (Table I) that could provide the basis for productive bridging-the-gap workshop sessions [19].

Having worked through the table, workshop participants could then address the following questions to decide whether there is any need to improve the quality of their assessment and, if so, what aspects need to be focused on:

### Table I

A list for discussion of criteria that instructors may use to evaluate the quality of their assessment instruments and overall program

<table>
<thead>
<tr>
<th>To what extent does the instrument satisfy the following criteria?</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A. Before assessment: design of the instrument</strong></td>
<td></td>
</tr>
<tr>
<td>1) Does the instrument assess at least one of the specified learning outcomes/objectives (i.e. does it assess what you think it is assessing—is it valid)? Some possible sub-questions:</td>
<td></td>
</tr>
<tr>
<td>(a) What specific concept(s) do you think your question is designed to probe?</td>
<td></td>
</tr>
<tr>
<td>(b) Does it assess conceptual understanding?</td>
<td></td>
</tr>
<tr>
<td>(c) Does it assess any cognitive skills and, if so, which ones (see e.g. [4, 5, 20])?</td>
<td></td>
</tr>
<tr>
<td>(d) Does it allow for a range of scientifically correct (creative) answers?</td>
<td></td>
</tr>
<tr>
<td>2) Is the instrument compatible with the type of knowledge or skill being assessed (i.e. use appropriate forms of assessments depending on what is being assessed)?</td>
<td></td>
</tr>
<tr>
<td>3) If the question includes a diagram:</td>
<td></td>
</tr>
<tr>
<td>(a) Do you think the diagram and its constituent symbolism is clear and not too complex for the student to understand?</td>
<td></td>
</tr>
<tr>
<td>(b) Do you think the diagram will help the student to answer the question?</td>
<td></td>
</tr>
<tr>
<td>4) Do students have the necessary prior knowledge and skills to answer the question?</td>
<td></td>
</tr>
<tr>
<td>5) Have students had practice at answering this type of instrument?</td>
<td></td>
</tr>
<tr>
<td>6) Will students understand the expectations and nature of the task? (i.e. do they understand the question? Is the language clear and unambiguous?) Do they understand the criteria that will be used to assess/grade their answers [10]?</td>
<td></td>
</tr>
<tr>
<td>7) Is the standard of the assessment appropriate for what will be assessed (e.g. assessment for mastery of concepts, skills, principles; for competence regarding use of equipment; and, for adequate proficiency regarding general course information)?</td>
<td></td>
</tr>
<tr>
<td>8) Is there a marking memorandum that will ensure that the answers can be fairly and reliably graded? If appropriate, is there a rubric?</td>
<td></td>
</tr>
<tr>
<td><strong>B. After assessment: analysis of student responses</strong></td>
<td></td>
</tr>
<tr>
<td>1) Was the instrument reliable i.e. what was its coefficient of reliability?</td>
<td></td>
</tr>
<tr>
<td>2) Was the standard of the assessment appropriate for the level of study? Did you do an item analysis to identify items that may be too easy or difficult? Did most items discriminate between the top and bottom third of the class?</td>
<td></td>
</tr>
<tr>
<td>3) Did it reveal evidence of student difficulties and misconceptions?</td>
<td></td>
</tr>
<tr>
<td>4) In the case of multiple-choice questions did you do a distracter analysis to check the effectiveness of the distracters?</td>
<td></td>
</tr>
<tr>
<td>5) Did you give qualitative feedback to students regarding their level of understanding and any difficulties they showed (i.e. not just grades obtained) [10]?</td>
<td></td>
</tr>
<tr>
<td><strong>C. Overall evaluation of the assessment plan for the course</strong></td>
<td></td>
</tr>
<tr>
<td>1) Have all the outcomes/objectives of the course as a whole been adequately assessed?</td>
<td></td>
</tr>
<tr>
<td>2) Was any one of the outcomes/objectives over-assessed at the expense of some of the others?</td>
<td></td>
</tr>
<tr>
<td>3) Does the assessment program as a whole comprise a sufficient range of assessment techniques or types?</td>
<td></td>
</tr>
<tr>
<td>Does this range include assessing the same major concepts in different ways, and cater to differential strengths of students?</td>
<td></td>
</tr>
</tbody>
</table>

Consuming process. With many demands on our time, what benefits might be derived from undertaking these tasks? We would like to suggest four reasons. We would also like to suggest that not all instruments are equally in need of close scrutiny—priorities might need to be established.

First, it is the nature of science to strive for perfection. It is conceivable that once a microscope was developed with a 100× magnification that biologists could have decided, "good enough." But this was not the case, and more powerful microscopes in turn led to more insightful research. So it is with assessment. Better assessment can lead to both better teaching and learning [2], and to better research.

Second, we have a legal and moral obligation to ensure that high-stake assessment is both valid and reliable. We place ourselves in an untenable situation if we fail a student on a final examination with a score of 59 (assuming the pass mark to be 60) when subsequent research might show our instrument lacks reliability. If it were to have a standard error of say 15, then the true score of that student might well be 70.

Third, while the thrust of this article is on ways to improve assessment, both formative and summative, at the classroom level, there are other situations to consider. If an assessment instrument is used as part of a research project, then it is imperative to show what steps were taken to develop the instrument and what was done to ensure its validity and its reliability. Most peer-refereed journals in science education will require this information for any article, where the results and conclusions are based on an assessment instrument.
1) How well do your assessment instruments meet the criteria in Table I? Specify any areas of concern.
2) Would you make any changes to the way you design certain assessment instruments? If so, what specific changes would you make?
3) In what way do you think these changes would impact on or improve teaching and student learning?

A Final Point

The analysis techniques described in this article should be accessible to all, and are perfectly adequate for the analysis of the kind of assessment instruments typically used in departments in a science faculty. More rigorous analysis of the kind of assessment instruments typically used in large-scale curriculum projects and international comparisons of scientific achievement.

Acknowledgment—Patricia McLean (University of KwaZulu-Natal, Pietermaritzburg) is thanked for English proofing the final version.

REFERENCES

Definitions of Outcomes Assessment

In the early 1980s the term assessment, or more accurately, outcomes assessment, was adopted in the United States to refer to information obtained from students, graduates, and other stakeholders that may be used to improve academic programs and student services within universities. In many other countries, this process is called evaluation, or program evaluation. The term assessment is preferred in the U.S. to distinguish the process designed to improve programs and services from evaluation, a process designed to gauge the achievements of academic staff for purposes of awarding promotions, tenure, and merit pay.

This author views outcomes assessment as a prudent step in a process that begins with planning what we wish to do. Plans are implemented and simultaneously appropriate data can be collected for use in assessing progress. If assessment findings are used to improve our processes, our plans may be adjusted, and the cycle of planning, implementing, assessing, and improving begins anew. Assessment in this context may be defined as a process of providing credible evidence of the outcomes of higher education that is undertaken for the purpose of improving programs and services within an institution. A second, simpler definition focuses squarely on the paramount college outcome, student learning. Former vice president of the American Association for Higher Education, Theodore Marchese, calls assessment “a rich conversation about student learning informed by data” (personal communication, January 7, 2004). This definition may provide the best context for the study of assessment currently underway in Scotland.

Assessment of Individuals and Groups

When academic staff hear the term assessment, they think most often in terms of assessing individual student development. They assess basic skills such as the ability to write, communicate orally, or use mathematics, for the purpose of advising students about appropriate placement in courses. They review student performance in their classes or modules using assignments, papers, and projects. And as students complete some programs, they are given comprehensive written and/or oral exams that test what they have learned throughout their years of study. Important outcomes of assessing individual student development include the following: (1) faculty can assign marks or grades to students, (2) students learn about their own strengths and weaknesses so that they can correct them and improve their future performance, and (3) students acquire skills in self-assessment that they can use throughout their lives. Assessment of individual student development is a critically important component of the higher education experience.

For purposes of conducting outcomes assessment, we need a second look: at aggregated student work in a class or module, in sections of the same class, and even across classes in a curriculum. Looking at student work collectively, we can tell where learning is satisfactory and where gaps in learning exist. We may also obtain some clues about which approaches to instruction produce the most learning for which students. These group assessment activities consist of classroom assignments, tests, and projects—all the same sorts of measures that are used to assess individual student development. But with group assessment we can add a variety of other measures, such as questionnaires for students, graduates, and employers. Interviews and focus groups yield helpful data. We can look at program completion data to see how many students complete our courses and curricula and how long it takes them. We can look at the placement of students in further education or careers. By tracking our graduates, we can see
how successful they are in post-graduate programs or on the job and if they have received
awards or recognition for their performance. Finally, we can use the results of group, or
outcomes, assessment to improve our programs and to demonstrate accountability to external
stakeholders.

To summarise, assessment of individual student development can assist students in mastering
content as well as in learning to assess their own strengths. Group, or outcomes, assessment
can help faculty improve instruction and enable institutions to demonstrate their accountability.

Good assessment, or evaluation as many call it, embodies the same principles as does good
research. In both we pose an important question, determine an appropriate approach to
answering the question, collect data, analyse the findings, and issue a report. Assessment goes
a step farther in that the findings are utilized to improve instruction in individual classrooms as
well as entire academic programs and university-wide services.

Preparation Academic Staff to Conduct Assessment

Since most academic staff are not trained as teachers, faculty development is an important
prerequisite for conducting good assessment. Faculty development can help instructors:

• write clear objectives for student learning in modules and curricula,
• individualise instruction using a variety of methods and materials, and
• develop assessment tools that test higher order intellectual skills.

In determining appropriate approaches to assessment, it is very helpful to write goals and
objectives for student learning using action verbs. For instance, if we want students to improve
their writing skills, an appropriate assessment of their progress would be a written assignment. If
we want them to develop skills in locating reliable information, we could give them a project
incorporating the use of such skills in order to assess their Internet search and analysis
strategies.

Bloom's Taxonomy of Educational Objectives (Bloom, 1956) consists of six increasingly complex
categories that describe what Bloom has called the cognitive domain. These extend from
knowledge and comprehension at the lowest level of complexity through application, analysis,
synthesis, and evaluation. Action verbs may be associated with each of these levels of the
domain. For instance, if we develop an objective for students using a verb such as identify,
define, or describe, this learning objective is at the knowledge level. If we ask them to
demonstrate, compute, or solve, students will be performing at the application level. If we expect
them to criticize, compare, or conclude, the students will be developing skills at the evaluation
level. In faculty development, discussing the use of verbs from the various levels of Bloom’s
Taxonomy can be a helpful step in developing the ability to assess learning outcomes.

The use of action verbs in learning objectives may be illustrated more specifically as follows: If
we ask a student in an English course to demonstrate how language influences intellectual and
emotional responses, we are testing the student’s application skills. Synthesis skills would be
illustrated in the following objective: Synthesize diverse issues and responses raised in
collaborative discussions of texts. Learning outcomes in science might include the following:
Define and explain basic principles, concepts, and theories of science (knowledge level); solve
theoretical and experimental problems in science (application level); and evaluate scientific
arguments at a level encountered by informed citizens (evaluation level).

A matrix can be useful in a number of ways in promoting conceptual thinking about assessment.
A matrix format with six columns that has been used successfully at many colleges and
universities in the United States is one that has as a heading for the first column, “What general
outcome are you seeking (e.g., critical thinking)?” The second column is headed “How would you
know it (the outcome) if you saw it—that is, what would the student know or be able to do?” The
third column heading is “How will you help students learn the concept, in class or out of class?” And the fourth heading is “How could you measure each of the desired behaviours listed in column 2?” The fifth column heading reads “What are the assessment findings?” And the sixth asks “What improvements are or might be based on assessment findings?” Completing such a matrix can enable faculty to explain to students and other stakeholders (1) specific learning outcomes of a module or a course of study, (2) collective student outcomes, and (3) actions undertaken to improve student learning based on assessment findings.

Classroom, Unit, and University-Wide Levels of Assessment

Outcomes assessment occurs at a number of levels. It begins with the individual student in a classroom. Aggregating the work of all students in a classroom will provide information to inform classroom assessment. Aggregating student work across various classes or modules can provide assessment (evaluation) of the impact on learning of an entire course of study. Looking at student products across the disciplines in a college provides assessment at that level. Assessment findings from various academic units within a university can provide a measure of institutional effectiveness that can be used to demonstrate accountability at the state, regional, or national level.

A distinction must be drawn between direct and indirect measures of student learning. Direct measures are those assignments, exams, projects, and papers that enable us to see what students actually know and can do. Indirect measures include questionnaires, interviews, and focus groups that enable us to assess the process of learning or other aspects of the student experience. Direct measures of learning are critical if we are to assess acquisition of knowledge and skills. But no test score will tell us why certain components of students’ knowledge are strong or weak. Thus indirect measures are needed to help us understand why weaknesses are occurring and what might be done to address them. Good assessment includes both direct and indirect measures.

Citing some examples of assessment at various levels may add clarity to this concept. Fast feedback, or classroom assessment, can be used at the individual classroom level. Students are asked during the last five minutes of a classroom session to state the most important thing they learned in the class that day and to tell the instructor what is still unclear. Then they may be asked about the helpfulness of the advance reading assignments for the day’s work. Finally, they may be asked for suggestions for improving the class and/or the assignments. In an illustration from the Graduate School of Business at the University of Chicago, students responded to the last question in that sequence by suggesting the following improvements: (1) install a portable microphone, (2) increase the type size on transparencies, (3) leave lights on when using a projector, (4) don’t cover the assigned reading in great detail, but instead (6) provide more examples from actual practice in class lectures and discussion (Bateman and Roberts, 1993).

We can adapt the typical course evaluation to include questions about the student experience. Are students encountering in the course principles of good practice in undergraduate education (Chickering and Gamson, 1987)? We might ask, for instance, if in a given module or in an entire curriculum (1) learners held high expectations for one another, (2) learners interacted frequently with academic staff in and outside class, (3) learners participated in learning teams, (4) learners respected diverse talents and ways of learning (Cournoyer, 2001).

Primary Trait Scoring

Primary trait scoring is an assessment method that can be used in both direct and indirect measures, and at all levels (Walvoord and Anderson, 1998). Instructors identify the traits or attributes that are necessary for success in an assignment, then compose a scale or rubric that gives clear definition to each point, and finally evaluate student work according to the rubric. For example, a project that involves developing and presenting a research paper encompasses at
least the following primary traits: (1) an appropriately narrow topic or purpose, (2) a bibliography, (3) an outline, (4) a first draft, (5) a final draft, and (6) an oral defence. For each of the traits of this assignment we might develop a three-point rubric, defining each point carefully and explicitly. The bibliography, for instance, might be assessed as follows:

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Outstanding: References current, appropriately cited, representative and relevant</td>
</tr>
<tr>
<td>2</td>
<td>Acceptable: References mostly current, few citation errors, coverage adequate, mostly relevant</td>
</tr>
<tr>
<td>1</td>
<td>Unacceptable: No references or containing many errors in citation format, inadequate coverage, or irrelevant</td>
</tr>
</tbody>
</table>

If one creates a matrix containing the primary traits of an assignment as row titles and the levels of each rubric as column headings, such a matrix can serve three purposes. First, it can be shared with students prior to an assignment so that they will understand the criteria being used to judge their work. Second, it can be completed for each student on the basis of the work submitted and thus provide detailed feedback when returned to the student. Third, if the instructor places a check mark in the appropriate box of the matrix for every mark assigned in evaluating the work of all students, the matrix can indicate to the instructor where there are weaknesses in student learning and suggest what changes may need to be made to enable every student to reach the desired learning outcomes.

Another matrix might list principal outcomes as row titles and courses in a curriculum as column headings. Placing check marks in the matrix to demonstrate which outcomes each course addresses will help students understand where they will learn specified knowledge and will assist instructors in spotting gaps in the curriculum.

Primary trait scoring can be used in virtually any field. For instance, at Ball State University in Indiana, sophomore competence in mathematics was tested. Students were asked to turn in their supporting work in connection with their item responses on a math test. Then instructors used a four-point scale to score responses in terms of conceptual understanding, consistent notation, logical formulation, and completeness of the solution (Emert and Parish, 1996).

At North Dakota State University faculty in sociology and anthropology developed scenarios appropriate to the discipline, then asked graduating students to respond to the scenarios in groups (Murphy and Gerst, 1997). A faculty facilitator asked questions related to outcomes faculty had identified in three areas—concepts, theory, and methods. Then two faculty observing the group work used a 0-3 scale to rate each student on each question. Looking at aggregate scores across all student groups enabled faculty working together to ascertain strengths and weaknesses of their curriculum.

Group interaction also can be assessed using primary traits and scoring rubrics. Faculty at the Purdue University College of Pharmacy in Indiana developed a five-point scale ranging from 5 = consistently excellent to 1 = inconsistent and/or inappropriate to judge the performance of students working in groups (Chalmers and Mason, 1994). The characteristics faculty were observing included the following:

- listened to others
- actively contributed to discussion
- challenged others effectively
- was willing to alter own opinion
- effectively explained concepts/insights
- summarised proposed solutions
Involving Stakeholders in Assessment

In a comprehensive assessment program, it is important to involve as many stakeholders as possible. Such groups include students, faculty, student affairs professionals, administrators, graduates, and employers.

An example of involving students may be drawn from the experience of the Department of Psychology at Montevallo University (Judith Rogers, personal communication, May 4, 1994). Students were asked to form an advisory council and to provide continuous assessment of the psychology major. Students made a number of important recommendations for improving their program. They asked that a statement of expected ethical behaviors for students be drawn up and volunteered to do this themselves. They suggested that a second research course with a laboratory be added to the curriculum because they felt they needed more research experience. They asked for more comparative psychology; additional terminals for the statistics lab; and more opportunities in all their courses to write, make oral presentations, and conduct research projects.

At Harvard University, Richard Light (1992) has involved students in teams to evaluate both courses and programs. In addition to the good information about the strengths and weaknesses of courses and curricula that the students provide to academic staff, the students experience learning benefits as they engage in the assessment process. As they interact with their peers, they strengthen their communication skills and develop competence in working in a team. Their self-confidence increases and they improve their understanding of others’ perspectives. Finally, student involvement in learning increases.

At Indiana University Purdue University Indianapolis, faculty across the institution have agreed on student outcomes in six areas, including:

(1) core communication and quantitative skills
(2) critical thinking
(3) integration and application of knowledge
(4) intellectual depth, breadth, and adaptiveness
(5) understanding society and culture
(6) values and ethics.

Efforts are underway to develop an electronic portfolio that will give students the responsibility for demonstrating their skills in connection with each of the six Principles of Undergraduate Learning. The electronic format will give students the opportunity to use multiple media to demonstrate unique individual skills and achievements. Written reflections on the material they put into the portfolio will enable students to enhance their metacognitive skills.

Involving student services professionals with faculty in the assessment of learning is illustrated in an experience at Virginia Commonwealth University (Fuhrmann, 1995). There first-year students in English composition wrote a theme each week for 15 weeks. The subject of each theme was the student’s experience with, and opinions of, some aspect of campus life, such as the admissions process, the new-student orientation program, or academic advising. Teams of faculty and student affairs professionals read and evaluated the essays. Faculty from the Department of English assessed the student work in terms of the quality of the writing, while student affairs professionals gained valuable information about students’ perceptions of the quality of various campus programs and services. As a result of this team approach to assessment, a number of changes have been made at Virginia Commonwealth, including provision of a new central advising center, a revised advising handbook, and enhanced multicultural workshops.

Kretovics and McCambridge (1999) at Colorado State University provide an example of involving employers in assessment. They have used a combination of surveys and focus groups for employers of business graduates to identify skills, knowledge, and personality attributes that are
sought by employers. The findings yielded by these assessment techniques have encouraged faculty to make curriculum changes; motivated students to develop needed skills; and strengthened ties among faculty, students, and employers. Faculty have added a credit to the business communications course that increases time for students to work in teams and make more oral presentations. In addition, business ethics and social responsibility are now discussed in introductory courses, and a new Introduction to Business course emphasizes career decision-making.

Using Assessment Findings

The time and energy devoted to outcomes assessment can only be justified if the findings are used to improve student learning and development. A few examples illustrate such uses. At Columbia College, academic staff in social work wanted students to apply critical thinking skills in their clinical practice settings (Baskind, Shank, and Ferraro, 2001). More specifically, faculty specified that students would use statistical analysis to inform practice decisions. Laboratory exercises were used to assess the skill, and faculty set as a standard that 80% of all students would earn a score of at least 3.0 on a 4-point scale on their lab exercises. Columbia College faculty found that only 64% of the social work students scored 3.0 or better. By asking students what would help them most, faculty learned that students needed more time in the laboratory—more time to practice their skills. In response, faculty made the learning laboratory available to students five days a week.

At Southern Illinois University-Edwardsville, for more than a decade faculty have required every student to undertake and pass a "senior assignment" in their discipline (Eder, 2001). In business, the senior assignment is a case study analysis. In education, students must complete a professional portfolio. In psychology, students must develop a poster presentation based on a research project. In engineering, there is a senior design project, and in nursing, students must develop a plan of care for a patient. At the end of each academic year, faculty meet to go over the senior assignments for that year and draw conclusions about strengths and weaknesses in students’ knowledge and skills. Over the years, each discipline has made changes based on the evidence provided in the student work. In business, more case studies and research projects have been required at earlier levels of the curriculum. In education, more practice in classroom management has been offered to improve student learning. In psychology the statistics requirement has been changed. In engineering, students have been given more opportunities to practice their writing and speaking skills throughout the curriculum. And in nursing, an expensive simulation lab with computerized patients has been purchased so that students may practice their clinical and diagnostic skills without harming a patient!

At Eastern New Mexico University, portfolio review and/or an audition is conducted for every fine arts student every semester by a panel that includes faculty, students, community representatives, and professional staff or faculty outside fine arts (Alec Testa, personal communication, April 27, 2004). As a result of this intensive assessment process, changes have been made in instruction and programming and now it is possible to see some results of those changes. Faculty have observed that students’ creativity, conceptualization, and technique have improved.

Characteristics of Effective Assessment

Over the years since 1980 assessment practitioners have developed a number of characteristics that define effective assessment (Palomba and Banta, 1999). First, there should be an overall plan for assessment, suggesting where and when students will be evaluated, the evaluation instruments that will be used, and what will be done to respond to the data collected. As indicated earlier, all stakeholders in higher education should be drawn into the assessment process. For instance, students and employers, as well as faculty, can help to set goals and objectives for courses and curricula, select or design assessment methods, and even collect
some of the data. Assessment findings then should be reported to all stakeholders and their use encouraged. Finally, the assessment program itself should be evaluated periodically, preferably through peer review.

Successful assessment initiatives are led by committed individuals and include collaboration between faculty and student affairs leaders. Assessment depends upon effective teamwork in planning, implementation, and utilization of findings. Assessment thrives in a supportive campus climate where faculty and administrators demonstrate their concern for students and their commitment to continuous improvement. Once assessment results are reported, campus leaders must follow up periodically to make sure that the findings are being used to improve practice.

If the experience of the last quarter-century is any guide, the need for providing evidence of accountability in higher education will not diminish, but rather will increase, as it has virtually every year since 1980. As scholarship reveals how assessment can improve instruction and student learning, more faculty will realize its benefits. Additional electronic assessment methods will be developed—already this is occurring at a rapid pace. More sharing of assessment methods will take place among groups of faculty. And ultimately, assessment will enable faculty to gain a deeper understanding of student learning and student learning will improve as assessment findings are used to effect continuous improvement of the environment for learning.

References


IUPUI Principles of Undergraduate Learning http://www.iport.iupui.edu

# PLANNING FOR LEARNING AND ASSESSMENT

1. **What general outcome are you seeking?**

2. **How would you know it (the outcome) if you saw it?** (What will the student know or be able to do?)

3. **How will you help students learn it?** (in class or out of class)

4. **How could you measure each of the desired behaviors listed in #2?**

5. **What are the assessment findings?**

6. **What improvements might be made based on assessment findings?**
The development of scientific and quantitative reasoning skills in undergraduates majoring in science, technology, engineering, and mathematics (STEM) is an objective of many courses and curricula. The Biology Department at James Madison University (JMU) assesses these essential skills in graduating biology majors by using a multiple-choice exam called the Natural World-9 (NW-9). NW-9, comprised of measures of Quantitative and Scientific Reasoning, contains items developed by faculty at JMU to assess the impact of the General Education program on the development of scientific and quantitative reasoning skills in a content-independent manner. We discuss methodology we used to involve faculty in determining the generalizability of NW-9 to assess the objectives of the biology curriculum and setting standards to interpret student achievement on NW-9. Student performance on NW-9 identified both strong and weak areas in our instruction and suggested that our biology faculty needs to reevaluate methodology for teaching students how to interpret and analyze data. More important, we can close the assessment loop by allowing faculty to participate in the assessment process and meaningfully reflect on student assessment results.

There are three options available to faculty interested in assessing the impact of undergraduate education on scientific and quantitative reasoning skills: use an existing instrument, modify an existing instrument, or develop a new instrument. Given the importance that science, technology, engineering, and mathematics (STEM) programs and national science organizations place on the development of scientific and quantitative reasoning skills, one would expect to find an endless array of reliable instruments that assess whether students graduating from undergraduate programs successfully acquired these essential skills (Howard Hughes Medical Institute 1996; NRC 2003). Many of the standardized tests, such as the Graduate Record Examination, include items that assess scientific reasoning ability, but for the most part research-based standardized tests address content knowledge (Bao et al. 2009). The Classroom Test of Scientific Reasoning developed by Lawson in 1978 is still popular among STEM educators, but this instrument addresses very broad areas of scientific reasoning and does not assess quantitative reasoning skills (Lawson 1978). Unfortunately, few readily accessible instruments are available that reliably assess both scientific and quantitative reasoning skills in undergraduates.

James Madison University (JMU) is a publicly funded, comprehensive institution of approximately 18,000 students in Harrisonburg, Virginia and has a strong emphasis on program assessment. The nationally recognized Center for Assessment and Research Studies (CARS) provides significant resources to the development of a nationally recognized assessment program (www.jmu.edu/assessment/). Building on the need for assessment of scientific and quantitative reasoning in higher education, and more specifically to inform STEM education, members of CARS in partnership with JMU faculty developed the Natural World-9 (NW-9) instrument, which contains two components: the Scientific Reasoning Test (SR-9; Sundre, 2008) and the Quantitative Reasoning Test (QR-9; Sundre, Thelk, and Wigtil 2008). All NW-9 items were written by James Madison University science and mathematics faculty to assess the objectives of the science component of the General Education program (see Table 1). Rather than investing faculty time in developing a new instrument, we decided to explore whether the NW-9 instrument developed and tested by CARS could assess scientific and quantitative reasoning skills in biology majors. We also wanted to involve faculty in this process to enhance faculty understanding and appreciation of the assessment process and results.

The Department of Biology has 56 full-time and part-time faculty, approximately 900 declared majors, and 100–125 students who graduate...
The biology curriculum is designed upon an explicit set of content, skill, and experience learning objectives developed by biology faculty. These objectives support the two major goals of the curriculum: ensuring that biology majors are literate in the scientific process and integrating research experiences into the learning environment for all our majors. Specifically, the skill objectives concentrate on scientific reasoning skills (see Table 1, skill objectives 1–10), but they also include objectives related to effective communication skills (see Table 1, skill objectives 11–14) and the ability to use quantitative reasoning skills to analyze biological phenomena (Table 1, skill objectives 7 and 14).

Assessment of the skill objectives is based on the results of two instruments, a modified version of the Academic Skills Inventory (ASI; Kruger and Zechmeister 2001) and the NW-9. The ASI differs from the NW-9 instrument in that the ASI asks students to report their experience level with a variety of academic skills, whereas the NW-9 instrument directly measures skill level. Results from the ASI indicate that students self-report behavioral gains in skills associated with written and oral communication, research methodology, and statistics (Seifert et al. 2009). Although the ASI provides insights regarding how well graduates of the biology major achieve some of the skill objectives, the NW-9 exam provides a more direct measurement of scientific and quantitative reasoning skills.

Although the NW-9 instrument was designed to assess the General Education learning objectives, there are many features of NW-9 that suggest this instrument will provide meaningful data to assess the skill objectives of the biology major. First, many of the General Education objectives are similar to the biology major skill objectives. For example, skill objectives 7, 9, and 10 and General Education objective 8 both discuss the ability of students to evaluate scientific sources, and skill objective 1 and General Education objective 6 both explore students’ ability to distinguish between association and causation. Second, CARS has extensively tested both components of NW-9 to establish two important measures of a meaningful assessment instrument: reliability and validity. The NW-9 instrument reliability and validity scores suggest that the instrument consistently measures the scientific and quantitative reasoning objectives of the General Education program (Sundre 2008; Sundre, Thelk, and Wigtil 2008). Third, NW-9 items do not test specific content knowledge. Rather, many of the items provide content necessary to determine the answer (see Figure 1a), whereas other items test concepts that do not rely on factual information (see Figure 1b). Based on these features of NW-9, we determined the generalizability of the NW-9 instrument to assess the skill objectives of the biology major. We did this by involving biology faculty in a content alignment process in which they mapped NW-9 items to the skill objectives. Results from these endeavors allow us to (1) evaluate senior biology major students’ performance on the mapped items; (2) determine whether students fell below, met, or exceeded faculty standards; and (3) discuss NW-9 assessment results at

### TABLE 1

Comparison of biology major skill objectives ($N = 14$) with General Education Cluster 3 objectives ($N = 7$).

**Biology major skill objectives**

1. Discriminate between association and causation, and identify the types of evidence used to establish causation.
2. Formulate a hypothesis and identify relevant variables necessary to test that hypothesis.
3. Design and execute experiments to test hypotheses.
4. Obtain data.
5. Organize data.
6. Analyze and interpret data.
7. Evaluate a statement, hypothesis, or claim using numerical or other evidence.
8. Locate sources of scientific information.
9. Evaluate the reliability of sources.
10. Critically evaluate a paper from the primary scientific literature.
11. Use effective professional communication in posters.
12. Use effective professional communication in lab reports.
13. Use effective professional communication in oral reports.
14. Use mathematics to understand and analyze biological phenomena.

**General Education Cluster 3 objectives**

1. Describe the methods of inquiry that lead to mathematical truth and scientific knowledge and be able to distinguish science from pseudoscience.
2. Use theories and models as unifying principles that help us understand natural phenomena and make predictions.
3. Recognize the interdependence of applied research, basic research, and technology, and how they affect society.
4. Illustrate the interdependence between developments in science and social and ethical issues.
5. Use graphical, symbolic, and numerical methods to analyze, organize, and interpret natural phenomena.
6. Discriminate between association and causation, and identify the types of evidence used to establish causation.
7. Formulate hypotheses, identify relevant variables, and design experiments to test hypotheses.
8. Evaluate the credibility, use, and misuse of scientific and mathematical information in scientific developments and public-policy issues.

Vol. 40, No. 6, 2011
departmental retreats regarding pedagogical strategies utilized by biology faculty to address the skill objectives.

**Methods**

**Content alignment of NW-9 items to skill objectives**

A critical step in determining the generalizability of the NW-9 instrument is to examine the content alignment between test items and skill objectives (D’Agostino et al. 2008). The degree of content alignment determines the ability of individual items to provide accurate information on student performance for each objective. Based on advice from the assessment experts at the CARS, we utilized item-level analysis to determine content alignment of the NW-9 instrument to the skill objectives (Martone and Sireci 2009). This was accomplished by recruiting eight faculty members, representing various subdisciplines in biology, to analyze the 66 items on the NW-9 instrument. Each faculty member provided independent judgments on whether an item successfully assessed one or more of the skill objectives. Faculty members were asked to review one stated learning objective at a time and determine whether or not each NW-9 item successfully assessed that objective. A dichotomous choice was provided for each item (yes or no). After making judgments about one objective, the faculty member proceeded to the next skill objective. No additional discussions or attempts to form consensus were attempted. This objective by objective procedure is less arduous for faculty than attempting to simultaneously make judgments about individual items across all learning objectives (D’Agostino et al. 2008). In consultation with the CARS, we developed a fairly stringent rule that an item would be deemed successfully mapped to a skill objective if six out of the eight evaluators (75%) assigned the item to a particular objective.

**Establishing faculty standards**

We used a modified Angoff method to establish a faculty standard for

---

**FIGURE 1**

Two examples of NW-9 items: (a) question that requires students to demonstrate proficiency in more than one skill, and (b) question that assesses the ability of students to interpret data.

(a) Regarding the two graphical displays given below, which of the following statements is correct?

![Average Monthly Temperature at Banebrook](image1)

![Average Monthly Temperature at Grove City](image2)

a. Banebrook has the largest changes in temperature throughout the year.
b. Banebrook and Grove City temperatures exhibit exponential behavior throughout the year.
c. Neither of the above.

(b) Suppose a researcher wants to test the hypothesis that exposure to cadmium in childhood causes neurological damage that reduces IQ. The researcher randomly selects 500 fourth graders, monitors their cadmium exposure for one year, and then tests each student’s IQ. The researcher finds that as cadmium exposure increases, IQ declines. Can the researcher conclude from the observed association between cadmium exposure and intelligence that cadmium causes reduced IQ?

a. No. The researcher did not include enough persons in the study.
b. No. There may be a third variable associated with exposure to cadmium that actually causes the lowered IQ.
c. Yes. The researcher followed the scientific method.
d. Yes. An association between the amount of cadmium exposure and lowered IQ is exactly what we would predict from the hypothesis.
each skill objective to provide greater interpretive power regarding student results (Maurer et al. 1991). The Angoff method provides a quantitative benchmark to determine whether graduating seniors are meeting faculty expectations. Biology faculty members \((n = 15)\) who had no knowledge of student test performance examined each of the NW-9 items that mapped to the skill objectives. The faculty volunteers were asked to provide a judgment of the percentage of graduating biology majors who should provide a correct response for each item. During this exercise, faculty members were asked not to discuss their ratings until after completion of the entire exercise. Following Angoff methodologies, faculty ratings for each item were grouped, on the basis of the mapping data, to the appropriate skill objectives. The mean of the scores for each skill objective represents the faculty standard for student success (see Table 2).

### Determining student performance on NW-9

We administered the NW-9 instrument to 214 graduating seniors (88 in 2008 and 126 in 2009). The mean student scores on the suite of questions corresponding to each of the seven skill objectives were calculated and transformed to the percentage correct. For each objective, the faculty standards were compared with the performance of the graduating seniors using a Mann-Whitney U nonparametric test with sequential Bonferroni post hoc analysis (see Table 2). Cohen’s \(d\) was used to determine effect size. If the mean student score for an objective was significantly higher than the faculty standard, students exceeded the faculty standard for that objective. If the mean student scores were not significantly different from the faculty standard, then students met the faculty standard. If the mean student score was significantly lower than the faculty standard, then students did not meet the faculty standards.

### Results

#### Content alignment of NW-9 items to the skill objectives

The stringent content alignment activity we utilized revealed that 25 of the 66 items strongly mapped to 7 of the 14 skill objectives. The objectives for which items were successfully aligned relate to distinguishing association from causation, formulating and evaluating hypotheses, designing experiments, analyzing and interpreting data, and using mathematics to understand biological phenomena (see Table 2). We found that multiple items were assigned to each of these seven objectives. However, using the established criteria, there were no items that mapped to skill objectives relating to obtaining data; organizing data; locating sources of scientific information; and critically evaluating a paper.

### TABLE 2

Number of NW-9 items mapped, faculty standard, and student performance for six skill objectives.

<table>
<thead>
<tr>
<th>Skill objective</th>
<th>NW-9 items mapping to objective</th>
<th>Faculty standard</th>
<th>Student performance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Student performance exceeded faculty standard</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Design and execute experiments to test hypotheses.</td>
<td>3 items (5% of test)</td>
<td>84.8%</td>
<td>91.6% ((p &lt; .0001, d = .27))</td>
</tr>
<tr>
<td>14. Use mathematics to understand and analyze biological phenomena.</td>
<td>2 items (3% of test)</td>
<td>74.8%</td>
<td>87.1% ((p &lt; .0001, d = .36))</td>
</tr>
<tr>
<td><strong>Student performance met faculty standard</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Discriminate between association and causation, and identify the types of evidence used to establish causation.</td>
<td>6 items (9% of test)</td>
<td>79.3%</td>
<td>75.5% ((p = .5920, d = .53))</td>
</tr>
<tr>
<td>2. Formulate a hypothesis and identify relevant variables necessary to test that hypothesis.</td>
<td>11 items (17% of test)</td>
<td>82.6%</td>
<td>86.5% ((p = .050, d = .21))</td>
</tr>
<tr>
<td>7. Evaluate a statement, hypothesis, or claim using numerical or other evidence.</td>
<td>15 items (23% of test)</td>
<td>78.3%</td>
<td>75.7% ((p = .9740, d = .17))</td>
</tr>
<tr>
<td><strong>Student performance fell below faculty standard</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Analyze and interpret data.</td>
<td>23 items (33% of test)</td>
<td>81.0%</td>
<td>70.4% ((p &lt; .009, d = .58))</td>
</tr>
</tbody>
</table>

*Note:* The faculty standard was derived from biology faculty predicting the percentage of graduating biology majors whom they thought would provide a correct response for each item (refer to Faculty standards in the Results section). Student performance was the percentage of correct answers that mapped to each skill objective. For each objective, the faculty standards were compared with the performance of the graduating seniors using a Mann-Whitney U nonparametric test with sequential Bonferroni post hoc analysis. Cohen’s \(d\) was used to determine the effect size \((d)\). NW-9 = Natural World-9.
primary literature; and using effective professional communication in posters, lab reports, and oral reports (skill objectives 4, 5, 8–13). This was an expected and validated finding. These learning objectives are not amenable to selected response item types. We currently use the ASI and are exploring other more direct methods to assess these skills and competencies. Some of the other NW-9 items that did not map to the skill objectives are designed to assess General Education objectives that do not align with faculty-developed curricular objectives of the biology major, such as understanding the difference between basic and applied research.

The most highly assessed objective was skill objective 6, analyzing and interpreting data, as 33% of the NW-9 items mapped to this objective (see Table 2). An example of a NW-9 item that assesses the ability of students to interpret data is shown in Figure 1a. Content in this item is not directly addressed in any biology course, which allows us to determine whether the student can transfer and generalize knowledge to interpret data in a situation in which they are not familiar with the content. Many items mapped to more than one skill objective, which reflects that many of the NW-9 items require students to demonstrate proficiency in more than one skill to achieve the correct answer. Overall, the content alignment activity provided validation for the use of NW-9 test scores to assess many of the quantitative and scientific reasoning objectives of our curriculum.

### Faculty standards

For the most part, the faculty standards for each skill objective were in the 75%–85% range (see Table 2). The highest faculty standard, 84.8%, was for designing and executing experiments to test hypotheses, whereas the lowest, 74.8%, was for using mathematics to understand biological phenomena.

### Student performance

Graduating biology majors exceeded faculty expectations for two skill objectives, met faculty expectations for three skill objectives, and fell below faculty standards for one skill objective (see Table 2). Seniors exceeded the faculty standard for designing experiments and using mathematics to understand a biological phenomena (skill objectives 3 and 14). In particular, the average score for items that map to designing and executing experiments (skill objective 3) was 91.6%, which is much higher than the faculty standard of 84.8% ($p = .0001, d = .36$). Graduating seniors met the faculty standard for formulating hypotheses, discriminating between association and causation, and evaluating a statement or claim using evidence (skill objectives 1, 2, and 7). Finally, our assessment results indicate that seniors correctly answered 70.4% of the questions that map to skill objective 6, which is significantly lower than the faculty standard for analyzing and interpreting data (81.0%, $p < .009, d = .58$).

### Discussion

As a result of this project, we have empirical evidence that the NW-9 provides meaningful measures of quantitative and scientific reasoning skills in biology majors. We found that 25 items on the NW-9 instrument map to seven of the skill objectives of the biology curriculum. The curriculum objectives assessed by NW-9 represent essential scientific and quantitative reasoning skills. Most notably, the exam scores provide insight into students’ abilities to identify and evaluate evidence that can be used to establish causation, formulate hypotheses, identify relevant variables to test hypotheses, analyze and interpret data, and use mathematics to understand biological phenomena. We recognize that the skill objectives not assessed by NW-9 are difficult to evaluate with a multiple-choice exam (e.g., effectiveness in presenting scientific research), and we will seek new direct methodologies.

Faculty had an overall prediction that on average, 79% of the seniors would answer correctly the suite of NW-9 questions that mapped to the skill objectives. Faculty expectations across the objectives showed some variability, ranging from approximately 75%–85% correct. Some items and objectives were determined to be more challenging than others. Actual student performances ranged from approximately 70%–92% correct for a suite of questions mapped to a particular objective. Faculty expectations were highest (84.8% expected to answer correctly) for questions that mapped to the skill objective related to designing and executing experiments to test hypotheses. Student performance was also highest for questions that mapped to this objective (91.6% of the students answered these questions correctly).

Faculty standards were relatively high (>82.6% of students were predicted to answer the question correctly) for NW-9 questions referring to the skill objective of formulating hypotheses and designing and executing experiments to test hypotheses (see Table 2). This may be because experimental design is emphasized in biology courses, thus faculty members have higher expectations for these skills. The faculty standard was lowest for the objective related to using mathematics to answer biological phenomena (<75% of students were predicted to answer the questions correctly; see Table 2). This may be related to the difficulty of items that assess quantitative skills, but it could also reflect that faculty members do not feel confident that courses in the biology curriculum address these skills. Likewise, faculty standards for students’ ability to analyze and interpret data were on the low end of the spectrum (81%). This is surprising given that many laboratory courses emphasize the use of statistics to analyze data.
We found that the NW-9 exam can be used to assess many of the JMU Biology Department skill objectives, which are most likely similar to the objectives other Biology Departments have for their students. Our results demonstrate that the NW-9 exam can be used to assess scientific and quantitative reasoning skills in areas outside of the General Education curriculum. Institutions interested in implementing instruments, such as NW-9, should map the items to their curriculum objectives and set faculty standards, as these will vary with student populations, curriculum, and faculty expectations. Once student performance data is collected, faculty can identify areas of strength and weakness in instruction and/or curriculum.

Overall, results from the NW-9 instrument in conjunction with the results from the ASI (Seifert 2009) suggest that the current biology major curriculum produces students who have met or exceeded faculty expectations for most of the specified curriculum skill objectives. We also noted a weak area in the curriculum regarding the skill of analyzing and interpreting data. This suggests a need for conversations to occur between laboratory instructors in regards to this essential skill objective. Lab- oratory courses should be targeted, because this is where the majority of inquiry-based learning occurs, such as analyzing and interpreting data. This study provides a baseline measure for the impact of the curriculum on skill development. We will continue to monitor our assessment results to measure the impact of changes we implement in laboratory courses to see if these changes increase student skill in data analysis.

One of the most significant outcomes we observed as we implemented our assessment design was an increase in faculty participation and interest in the assessment process and student results. By involving biology faculty in the content alignment and standard setting activities, we created a customized process that we can use, as a department, to analyze student performance in the areas of scientific and quantitative reasoning. More important, we have created a culture of assessment in our department that reflects the goals of the curriculum, the perspective of the faculty, and an awareness of student learning outcomes. This process has helped us to “close the loop” with understanding and using our assessment results. Our faculty conversations about assessment, our program, and our students’ learning have been deepened and enriched. Most important, these results provide our faculty with compelling evidence that the NW-9 instrument measures many of the biology-major student learning objectives. We were able to engage many of our faculty in the development of a community-established expectation for student performance. Finally, this set of student performance expectations gave us a new and valued interpretive framework for our assessment results.

References

Carol A. Hurney is an associate professor of biology and executive director of the Center for Faculty Innovation, Justin Brown is an assistant professor of biology, Heather Peckham Griscom (griscohp@jmu.edu) is an associate professor of biology, and Erika Kancler is an assistant professor of biology, all at James Madison University (JMU) in Harrisonburg, Virginia. Clifton J. Wigtil is graduate student in gifted education at Purdue University. Donna Sundre is a professor of psychology and the executive director of the Center for Assessment and Research Studies at JMU.
Fundamental Assessment Principles for Teachers and School Administrators

James H. McMillan
Virginia Commonwealth University

While several authors have argued that there are a number of "essential" assessment concepts, principles, techniques, and procedures that teachers and administrators need to know about (e.g. Calfee & Masuda, 1997; Cizek, 1997; Ebel, 1962; Farr & Griffin, 1973; Fleming & Chambers, 1983; Gullickson, 1985, 1986; Mayo, 1967; McMillan, 2001; Sanders & Vogel, 1993; Schafer, 1991; Stiggins & Conklin, 1992), there continues to be relatively little emphasis on assessment in the preparation of, or professional development of, teachers and administrators (Stiggins, 2000). In addition to the admonitions of many authors, there are established professional standards for assessment skills of teachers (Standards for Teacher Competence in Educational Assessment of Students (1990), a framework of assessment tasks for administrators (Impara & Plake, 1996), the Code of Professional Responsibilities in Educational Measurement (1995), the Code of Fair Testing Practices (1988), and the new edition of Standards for Educational and Psychological Testing (1999). If that isn’t enough information, a project directed by Arlen Gullickson at The Evaluation Center of Western Michigan University will publish standards for evaluations of students in the near future.

The purpose of this article is to use suggestions and guidelines from these sources, in light of current assessment demands and contemporary theories of learning and motivation, to present eleven "basic principles" to guide the assessment training and professional development of teachers and administrators. That is, what is it about assessment, whether large-scale or classroom, that is fundamental for effective understanding and application? What are the "big ideas" that, when well understood and applied, will effectively guide good assessment practices, regardless of the grade level, subject matter, developer, or user of the results? As Jerome Bruner stated it many years ago in his classic, The Process of Education: "...the curriculum of a subject should be determined by the most fundamental understanding that can be achieved of the underlying principles that give structure to that subject." (Bruner, 1960, p.31). What principles, in other words, provide the most essential, fundamental "structure" of assessment knowledge and skills that result in effective educational practices and improved student learning?
**Assessment is inherently a process of professional judgment.**

The first principle is that professional judgment is the foundation for assessment and, as such, is needed to properly understand and use all aspects of assessment. The measurement of student performance may seem "objective" with such practices as machine scoring and multiple-choice test items, but even these approaches are based on professional assumptions and values. Whether that judgment occurs in constructing test questions, scoring essays, creating rubrics, grading participation, combining scores, or interpreting standardized test scores, the essence of the process is making professional interpretations and decisions. Understanding this principle helps teachers and administrators realize the importance of their own judgments and those of others in evaluating the quality of assessment and the meaning of the results.

**Assessment is based on separate but related principles of measurement evidence and evaluation.**

It is important to understand the difference between measurement evidence (differentiating degrees of a trait by description or by assigning scores) and evaluation (interpretation of the description or scores). Essential measurement evidence skills include the ability to understand and interpret the meaning of descriptive statistical procedures, including variability, correlation, percentiles, standard scores, growth-scale scores, norming, and principles of combining scores for grading. A conceptual understanding of these techniques is needed (not necessarily knowing how to compute statistics) for such tasks as interpreting student strengths and weaknesses, reliability and validity evidence, grade determination, and making admissions decisions. Schafer (1991) has indicated that these concepts and techniques comprise part of an essential language for educators. They also provide a common basis for communication about "results," interpretation of evidence, and appropriate use of data. This is increasingly important given the pervasiveness of standards-based, high-stakes, large-scale assessments. Evaluation concerns merit and worth of the data as applied to a specific use or context. It involves what Shepard (2000) has described as the systematic analysis of evidence. Like students, teachers and administrators need analysis skills to effectively interpret evidence and make value judgments about the meaning of the results.

**Assessment decision-making is influenced by a series of tensions.**

Competing purposes, uses, and pressures result in tension for teachers and administrators as they make assessment-related decisions. For example, good teaching is characterized by assessments that motivate and engage students in ways that are consistent with their philosophies of teaching and learning and with theories of development, learning and motivation. Most teachers want to use
constructed-response assessments because they believe this kind of testing is best to ascertain student understanding. On the other hand, factors external to the classroom, such as mandated large-scale testing, promote different assessment strategies, such as using selected-response tests and providing practice in objective test-taking (McMillan & Nash, 2000). Further examples of tensions include the following.

- Learning vs auditing
- Formative (informal and ongoing) vs summative (formal and at the end)
- Criterion-referenced vs norm-referenced
- Value-added vs. absolute standards
- Traditional vs alternative
- Authentic vs contrived
- Speeded tests vs power tests
- Standardized tests vs classroom tests

These tensions suggest that decisions about assessment are best made with a full understanding of how different factors influence the nature of the assessment. Once all the alternatives understood, priorities need to be made; trade-offs are inevitable. With an appreciation of the tensions teachers and administrators will hopefully make better informed, better justified assessment decisions.

**Assessment influences student motivation and learning.**

Grant Wiggins (1998) has used the term 'educative assessment' to describe techniques and issues that educators should consider when they design and use assessments. His message is that the nature of assessment influences what is learned and the degree of meaningful engagement by students in the learning process. While Wiggins contends that assessments should be authentic, with feedback and opportunities for revision to improve rather than simply audit learning, the more general principle is understanding how different assessments affect students. Will students be more engaged if assessment tasks are problem-based? How do students study when they know the test consists of multiple-choice items? What is the nature of feedback, and when is it given to students? How does assessment affect student effort? Answers to such questions help teachers and administrators understand that assessment has powerful effects on motivation and learning. For example, recent research summarized by Black & Wiliam (1998) shows that student self-assessment skills, learned and applied as part of formative assessment, enhances student achievement.

**Assessment contains error.**

Teachers and administrators need to not only know that there is error in all classroom and standardized assessments, but also more
specifically how reliability is determined and how much error is likely. With so much emphasis today on high-stakes testing for promotion, graduation, teacher and administrator accountability, and school accreditation, it is critical that all educators understand concepts like standard error of measurement, reliability coefficients, confidence intervals, and standard setting. Two reliability principles deserve special attention. The first is that reliability refers to scores, not instruments. Second, teachers and administrators need to understand that, typically, error is underestimated. A recent paper by Rogosa (1999), effectively illustrates the concept of underestimation of error by showing in terms of percentile rank probable true score hit-rate and test-retest results.

**Good assessment enhances instruction.**

Just as assessment impacts student learning and motivation, it also influences the nature of instruction in the classroom. There has been considerable recent literature that has promoted assessment as something that is integrated with instruction, and not an activity that merely audits learning (Shepard, 2000). When assessment is integrated with instruction it informs teachers about what activities and assignments will be most useful, what level of teaching is most appropriate, and how summative assessments provide diagnostic information. For instance, during instruction activities informal, formative assessment helps teachers know when to move on, when to ask more questions, when to give more examples, and what responses to student questions are most appropriate. Standardized test scores, when used appropriately, help teachers understand student strengths and weaknesses to target further instruction.

**Good assessment is valid.**

Validity is a concept that needs to be fully understood. Like reliability, there are technical terms and issues associated with validity that are essential in helping teachers and administrators make reasonable and appropriate inferences from assessment results (e.g., types of validity evidence, validity generalization, construct underrepresentation, construct-irrelevant variance, and discriminant and convergent evidence). Of critical importance is the concept of evidence based on consequences, a new major validity category in the recently revised *Standards*. Both intended and unintended consequences of assessment need to be examined with appropriate evidence that supports particular arguments or points of view. Of equal importance is getting teachers and administrators to understand their role in gathering and interpreting validity evidence.

**Good assessment is fair and ethical.**
Arguably, the most important change in the recently published *Standards* is an entire new major section entitled "Fairness in Testing." The *Standards* presents four views of fairness: as absence of bias (e.g., offensiveness and unfair penalization), as equitable treatment, as equality in outcomes, and as opportunity to learn. It includes entire chapters on the rights and responsibilities of test takers, testing individuals of diverse linguistic backgrounds, and testing individuals with disabilities or special needs. Three additional areas are also important:

- Student knowledge of learning targets and the nature of the assessments prior to instruction (e.g., knowing what will be tested, how it will be graded, scoring criteria, anchors, exemplars, and examples of performance).
- Student prerequisite knowledge and skills, including test-taking skills.
- Avoiding stereotypes.

**Good assessments use multiple methods.**

Assessment that is fair, leading to valid inferences with a minimum of error, is a series of measures that show student understanding through multiple methods. A complete picture of what students understand and can do is put together in pieces comprised by different approaches to assessment. While testing experts and testing companies stress that important decisions should not be made on the basis of a single test score, some educators at the local level, and some (many?) politicians at the state at the national level, seem determined to violate this principle. There is a need to understand the entire range of assessment techniques and methods, with the realization that each has limitations.

**Good assessment is efficient and feasible.**

Teachers and school administrators have limited time and resources. Consideration must be given to the efficiency of different approaches to assessment, balancing needs to implement methods required to provide a full understanding with the time needed to develop and implement the methods, and score results. Teacher skills and knowledge are important to consider, as well as the level of support and resources.

**Good assessment appropriately incorporates technology.**

As technology advances and teachers become more proficient in the use of technology, there will be increased opportunities for teachers and administrators to use computer-based techniques (e.g., item banks, electronic grading, computer-adapted testing, computer-based simulations), Internet resources, and more complex, detailed ways of reporting results. There is, however, a danger that technology will
contribute to the mindless use of new resources, such as using items on-line developed by some companies without adequate evidence of reliability, validity, and fairness, and crunching numbers with software programs without sufficient thought about weighting, error, and averaging.

To summarize, what is most essential about assessment is understanding how general, fundamental assessment principles and ideas can be used to enhance student learning and teacher effectiveness. This will be achieved as teachers and administrators learn about conceptual and technical assessment concepts, methods, and procedures, for both large-scale and classroom assessments, and apply these fundamentals to instruction.

Notes:

An earlier version of this paper was presented at the Annual Meeting of the American Educational Research Association, New Orleans, April 24, 2000.

References


Contact Information:

James H. McMillan
Box 842020
Virginia Commonwealth University
Richmond, VA 23284-2020

Phone: 804 828-1332 , x553
Fax: 804-225-3554
jmcmilla@saturn.vcu.edu

Descriptors: *Standards; Professional Standards; Test Scores; Student Evaluation
RESEARCH REPORT

Development and Validation of Instruments to Measure Learning of Expert-Like Thinking

Wendy K. Adams\textsuperscript{a,b,*} and Carl E. Wieman\textsuperscript{c,d}

\textsuperscript{a}Department of Physics, University of Northern Colorado, Colorado, USA; \textsuperscript{b}Acoustical Society of America, New York, USA; \textsuperscript{c}Physics and Science Education Initiative, University of British Columbia, Vancouver, Canada; \textsuperscript{d}Physics and Science Education Initiative, University of Colorado, Boulder, USA

This paper describes the process for creating and validating an assessment test that measures the effectiveness of instruction by probing how well that instruction causes students in a class to think like experts about specific areas of science. The design principles and process are laid out and it is shown how these align with professional standards that have been established for educational and psychological testing and the elements of assessment called for in a recent National Research Council study on assessment. The importance of student interviews for creating and validating the test is emphasized, and the appropriate interview procedures are presented. The relevance and use of standard psychometric statistical tests are discussed. Additionally, techniques for effective test administration are presented.

Keywords: Assessment; Formative assessment; University; Science education; Evaluation; Assessment design

In recent years, there has been a growing effort to develop assessment tools that target students’ development of expert-like mastery of specific science topics. These involve questions that accurately probe whether students understand and apply particular concepts in the manner of a scientist in the discipline. Such assessment tools are intended to measure student learning in courses to provide Formative Assessment of Instruction (FASI). We present the methodology involved in developing and validating such assessment tools. This same methodology works equally well
for developing assessment tools to measure other aspects of student thinking, such as their perceptions of a field of science and how it is best learned.

In this paper, we will describe how a faculty member or education researcher can create a valid and reliable assessment tool of this type. We have found that a postdoctoral researcher in the field of the topic content with a few months experience in education research training can carry out such a process. Once this tool is created and validated, it provides a way to compare instruction across institutions and over time in a calibrated manner. The methodology we describe for test construction matches well with the ‘Standards for educational and psychological testing’ (American Educational Research Association [AERA], American Psychological Association [APA], & the National Council on Measurement in Education [NCME], 1999) and is closely aligned with what is called for in the National Research Council’s study of assessment (NRC, 2001). We provide a detailed discussion on how to implement this general methodology in specific domains. We will also discuss appropriate statistical analyses that are part of validity evidence and how interpretation of these statistics depends on the nature of the assessment.

One particularly important part of both the development and validation of a FASI is the use of student interviews. There is a large body of literature on the use of student interviews for the purpose of understanding student thinking (Berardi-Coletta, Buyer, Dominowski, & Rellinger, 1995; Ericsson & Simon, 1998), but student interviews are rarely used when developing educational tests, although the value of this kind of information is stressed in the 2001 NRC report. ‘The methods used in cognitive science to design tasks, observe and analyze cognition, and draw inferences about what a person knows are applicable to many of the challenges of designing effective educational assessments’ (NRC, 2001, p. 5). Here, we will discuss how to use this technique from cognitive science in developing and validating a FASI.

Beginning with the Force Concept Inventory (FCI; Hestenes, Wells, & Swackhamer, 1992), a number of FASI-type instruments have been developed to measure student learning of science at the university level in a systematic way, and these are having a growing impact on teaching and learning. While there are similarities in various groups’ approaches to the development and validation of such instruments, the procedures are not fully consistent partially due to the fact that, to our knowledge, no one has written down the process for this specific type of instrument in its entirety. As an example, while describing the steps for evaluating student misconceptions on a particular topic, Treagust (1988) writes about the value of conducting student interviews to determine misconceptions and distracters. However, after reviewing 16 FASI-type instruments for physics, chemistry, geosciences, and biology that were developed in the past 20 years only Redish, Steinberg, and Saul (1998) and Singh and Rosengrant (2003) used student interviews during both the development and validation processes. The developers of the FCI did student interviews after developing the survey; however, the results of those interviews showed that student reasoning was not consistent ‘but Newtonian choices for non-Newtonian reasons were fairly common’ (Hestenes et al., 1992, p. 148),
raising some questions about validity. For this reason, the authors go on to state that scores below 80% can only be used as an upper bound on student understanding. Another reason that development procedures are not consistent could stem from the fact that developers of FASI-type instruments are necessarily content experts so may not be familiar with work in cognitive science or the field of assessment design. For example, the FCI authors discuss grouping of questions that address different aspects of the content but were later criticized for not using statistical measures such as a factor analysis to support these groupings (Heller & Huffman, 1995; Huffman & Heller, 1995) and since have retracted the suggestion of using groups of statements and recommend looking at only the total score (Hestenes & Halloun, 1995).

In most of the other examples listed, the appropriate psychometric statistical tests have either not been carried out or the results of such tests were misinterpreted, often due to confusion as to the distinctions between single-construct and multiple-construct assessments. Finally, the validity and reliability of any results obtained with FASI-type instruments depend on how it is administered, but test administration options and tradeoffs are seldom if ever discussed in the literature presenting the instruments.

This paper is intended to describe the complete process of developing a FASI based on previous work and our own experience in such a way that a content expert can create a valid and reliable instrument for their discipline. Individually, and with collaborators, we have now developed nine assessments that test expert-like mastery of concepts; four have been published (Chasteen & Pollock, 2010; Goldhaber, Pollock, Dubson, Beale, & Perkins, 2010; McKagan, Perkins, & Wieman, 2010; Smith, Wood, & Knight, 2008) and several that measure expert-like perceptions about the learning and application of various science subjects with two that are published (Adams et al., 2006; Barbera, Perkins, Adams, & Wieman, 2008). This process has now become relatively refined and straightforward.

Development Framework

Our process, as discussed in detail below, follows the four phases outlined in the Standards for Psychological and Educational testing (AERA, APA, & NCME, 1999, p. 37):

- Phase 1. Delineation of the purpose of the test and the scope of the construct or the extent of the domain to be measured;
- Phase 2. Development and evaluation of the test specifications;
- Phase 3. Development, field testing, evaluation, and selection of the items and scoring guides and procedures; and
- Phase 4. Assembly and evaluation of the test for operational use.

Phase 1

The basic theoretical idea behind all these instruments is that there are certain ways of thinking that experts within a particular subject share. ‘Studies of expert-novice
differences in subject domains illuminate critical features of proficiency that should be the targets for assessment’ (NRC, 2001, p. 4). Ericsson and others have identified how experts have particular mental structures by which they organize and apply information (Ericsson, Charness, Feltovich, & Hoffman, 2006). Expert chess players see patterns in the arrangement of the pieces that tell them how the game is progressing and identify optimum strategies. Physicists organize knowledge around fundamental concepts. When faced with a problem, they recognize patterns in the problem that cue concepts that will be productive for working out a solution. Identifying the unique characteristics that make up specific areas of expertise in different disciplines is an active field of research that is extensively reviewed in Ericsson et al. (2006). Expert thinking, however, goes beyond how information is organized and applied to include discipline-specific heuristics for monitoring problem-solving and other aspects of thinking such as perceptions of how the subject is best learned and where it applies. For example, physicists perceive physics as describing the real world and that it is best learned in terms of broadly applicable concepts. When learning something new, they believe that it should be carefully examined as to how it makes sense in terms of prior knowledge of physics and the behaviour of the world.

A suitable educational goal is to have students thinking more like experts and approaching the mastery of the subject like an expert, and so it is desirable to have test instruments that measure student thinking on a scale that distinguishes between novice and expert thinking (Shavelson & Ruiz-Primo, 2000). This requires a process for first identifying consistent expert thinking and then creating a valid test for measuring the extent to which students learn to think like experts during their instruction in any particular course.

The NRC calls for having three elements in the foundation to all assessments: cognition, observation, and interpretation (NRC, 2001). The expert–novice differences of value to teachers define the cognition that is being probed by the FASI; the questions themselves are the tasks that elicit this thinking (observations); and the validation process—demonstrating that the FASI measures what is intended—determines the interpretation by delineating what inferences about student thinking can be drawn from the results of the FASI.

There are several practical requirements for such an instrument if it is to be used on a widespread basis so it can impact instruction. (1) It must measure value added by the instruction, and hence it must be possible to administer it on both pre- and post-instruction basis. (2) It must be easy to administer and grade in the context of a normal course schedule without any training. It is typical to have to make some tradeoffs between the breadth and depth of what is measured and ease of administration, but this is possible to do without seriously compromising the validity. (3) The instrument must test ‘expert thinking’ of obvious value to teachers. (4) It needs to be demonstrated that it measures what it claims (evidence of validity).

It is unrealistic to have a test that meets goals 1–3 above and tests everything of importance for students to learn in any given course, particularly in a format that can be easily given and graded in a small amount of time. So a more practical goal is to
Development and Validation of a FASI

test mastery of a limited set of important, hard-to-learn concepts. Usually these will serve as an effective proxy for measuring how effectively concepts are being learned in the course as a whole. This is supported by the evidence that certain pedagogical approaches are superior in supporting conceptual learning, independent of the particular concept being taught (Bransford, Brown, & Cocking, 2000 and references therein). In selecting which concepts to include, consideration should be given to maximizing the range of institutions where the test can be used. The FCI is a good example of a test that is designed according to this principle. It covers a relatively small subset of the concepts of force and motion that are covered in a typical first-term college physics course. However, this particular set of concepts is taught and valued by nearly everyone, students have particular difficulty with them, and do worse than many teachers expect. The results from this test have been sufficiently compelling to cause many teachers to adopt different instructional approaches for all of their introductory physics course material.

Phase 2

Development and evaluation of the test specifications include: item format, desired psychometric properties, time restrictions, characteristics of the population, and test procedures.

According to classical and modern test theory: ‘In test construction, a general goal is to arrive at a test of minimum length that will yield scores with the necessary degree of reliability and validity for the intended uses’ (Crocker & Algina, 1986, p. 311). Although the intended use of a FASI is to measure how well students are thinking like experts, the primary goal is not to obtain a summative assessment of student learning; rather it is to provide formative assessment of teaching. Thus the results from the students as a group are more important than ranking individual students, which is fundamentally different from many other assessments. This makes this a low-stakes assessment and the testing approach must be tailored to this use. It also relaxes a number of constraints on test design (NRC, 2001). In addition to not needing to cover all of the course material, the instrument will also often be probing many different facets of thinking and learning, rather than a particular block of content (a single construct), such as how well the student can describe the molecular anatomy of genes and genomes. This makes the test design and statistical tests of the instrument rather different.

Psychometricians typically use either item analysis or Item Response Theory (IRT) to determine which items in the pool of potential test items will be the best to construct the most efficient, reliable, and valid test. However, the standard acceptable range of values for these statistical tests was determined for single construct, and summative tests intended to provide maximum discrimination among individual students. Because a FASI serves a much different purpose, statistical-item analysis takes a secondary role to student interviews, which provide much more information about the validity of the questions, as well as often indicating why a question may have statistics that fall outside of the ‘standard’ psychometric range. That makes it particularly
essential that the students interviewed for both the development and validation of the test represent the breadth of the population for which the test is to be used.

When creating a FASI, it’s desirable to use a forced answer (multiple-choice or likert-scale) test. This format is easy to administer and to grade. Also, unlike open-ended questions that are graded with a rubric, it easily provides reliably consistent grading across instructors and institutions. To be useful, FASIs usually need to be administered in a pre- and post-format to normalize for initial level of knowledge. We have found that surveys of perception, which can be completed in less than 10 minutes, can be given online, while conceptual tests (20–50 minutes) need to be given during class time, with a careful introduction, so that students take the test seriously in order to obtain a useful result. See Test Administration Tips at the end of this paper for more details.

**Phases 3 and 4**

These two aspects of test construction comprise the bulk of the work needed to develop a new FASI. Here, we will list and then describe in detail the six steps that are required to develop and validate the test. These six steps are undertaken in a general order; however, it is usually an iterative process. Often part of the validation procedure reveals items that need modification, and then it is necessary to go back and create and validate these new or modified items. The entire process takes very roughly a half person year of a good PhD level person’s time to carry out. That effort will likely need to be spread out over one to two years of calendar time, due to constraints of the academic schedule and availability of student test subjects. Key elements of the process are the student interviews that are carried out at Steps 2 and 5. Although both sets of interviews largely rely on a think-aloud format, they are distinctly different. The validation interviews in Step 5 are far more limited in scope than the interviews done in Step 2. Step 5 interviews follow a much stricter protocol as discussed below.

1. Establish topics that are important to teachers (in our case, college or university faculty members).
2. Through selected interviews and observations, identify student thinking about these topics and the various ways it can deviate from expert thinking.
3. Create open-ended survey questions to probe student thinking more broadly in test form.
4. Create a forced answer test that measures student thinking.
5. Carry out validation interviews with both novices and subject experts on the test questions.
6. Administer to classes and run statistical tests on the results.

Modify items as necessary.

*Establish aspects of thinking about the topic that are important to faculty members.*

‘Educational assessment does not exist in isolation, but must be aligned with curriculum and instruction if it is to support learning’ (NRC, 2001, p. 3). Establishing
important aspects of thinking on the topic is usually the easiest step for a test designer who knows a subject well. These can be done through a combination of the following four methods: (1) reflect on your own goals for the students; (2) have casual conversations or more formal interviews with experienced faculty members on the subject; (3) listen to experienced faculty members discussing students in the course; and (4) interview subject matter experts. One can pick up many aspects of the topic that are important to faculty members who are experienced teaching the subject simply from casual conversations. Merely ask them to tell you what things students have done and said that indicated they were not thinking as the faculty member desired. Normally what is desired is to have the student thinking like an expert in the discipline. An even better source of information is to listen to a group of thoughtful teachers discussing teaching the subject and lamenting where students are not learning. This automatically picks out topics and thinking where there is a consensus among teachers as to what is appropriate for students to learn; otherwise they would not be discussing it. Such discussions also identify where there are conspicuous student difficulties and shortcomings in achieving the desired mastery. This information can be collected more systematically by interviewing faculty members who are experienced teachers to elicit similar lists of expert thinking that students often fail to achieve. As an example, a number of faculty members across multiple domains have mentioned domain-specific versions of ‘When the student has found an answer to a problem, be able to evaluate whether or not it is reasonable’.

Interviewing content experts, even if they have not taught extensively, such as research scientists, in a systematic fashion about the topics under consideration can also be valuable. The ideas that come up consistently in such interviews represent the core ideas that are considered important in the discipline. In our studies in physics, we see significant evolution of thinking between graduate students and faculty members about many fundamental topics in physics, and so for physics and likely other subjects, graduate students should not be considered to be subject experts (Singh & Mason, 2009).

Usually interviewing 6–10 faculty members/content experts is quite adequate to determine the important relevant expert thinking. FASIs are useful only if they focus on ideas that are shared by an overwhelming proportion of experts. So if something does not emerge consistently from interviewing such a small number of experts, it is best not to include it. In this step and the following, one is carrying out the steps called for in the NRC report: ‘Understanding the contents of long term memory is especially critical for determining what people know; how they know it; and how they are able to use that knowledge to answer questions, solve problems, and engage in additional learning’ (NRC, 2001, p. 3).

Selected interviews and observations to understand student thinking about these topics and where and how it deviates from expert thinking. Interviewing and observing students to understand thinking about the topics determined in Step 1 above is in accordance
to the call for developing additional cognitive analyses of domain-specific knowledge and expertise (NRC, 2001). As mentioned above, much of the non-expert-like student thinking that is of concern will already be revealed by thoughtful experienced teachers, which in many cases, will include the test developers. In some fields, there is also a significant literature on student misconceptions that is highly relevant. However, it is important to also do direct student observations such as observing and participating in course help sessions and conducting systematic student interviews. These are likely to reveal quite unexpected student thinking and perspectives if one listens carefully and avoids ‘filling in blanks’ with one’s expert knowledge and assumptions.

Course help sessions (a session, often optional, where students come in to work together on homework, often with some form of assistance by TAs) provide a wealth of information on student knowledge, how it is connected, and in what context students apply it. Help sessions are a very useful setting to start understanding student thinking because they have a lot of students working on the course material in one public location, allowing for efficient collection of information. Using the help session to specifically look for topics where student thinking deviates from the experts’ often reveals information that may have gone unnoticed previously. Begin by simply observing, listening to student discussions, and taking notes—particularly noting what content students think relates to particular homework questions, and how they use this content. After collecting observational data on how students respond to questions without researcher interference, it is often useful to work with a few students and carefully probe more deeply into what they are thinking both during informal help session interviews, or more formal arranged interviews.

To carry out interviews, one starts by soliciting student volunteers that span the full range and beyond (age, gender, ethnicity, background preparation, grade point average, career aspirations, etc.) of the student population to be tested. It is important to interview as broad a range of students as possible. Differences in thinking, both between different students and between students and experts become much more obvious when one looks at extreme cases. As there can be aspects of student thinking that an expert would never guess, because the perspective is so different, the enhanced ‘signal’ provided by considering extreme cases is almost always very helpful. For example, when discussing the idea of an electric or magnetic field with first-year physics or other science majors, they typically have an idea of some sort of invisible force; however, when discussing this idea with non-science majors we found that many visualized a sort of field similar to a farm field or field of grass. Another example came from a general science course for elementary education majors. The first author was extremely surprised to discover that many of these students believe that the continents float on the Earth’s oceans.

Student interviews can take a variety of forms, but ideally one will pose the student a question or problem on a subject where there is a clear expert consensus, and have them simply explain it or solve it, using a ‘think-aloud’ protocol (Ericsson & Simon, 1998, p. 182). A think-aloud protocol is where the student is told to think aloud while working a particular problem. The interviewer is restricted to prompting
the students to think aloud when they become quiet but must not ask the students to explain their reasoning or ask them how or why they did something. Once questions of this nature are asked, the students' thinking is changed when they attempt to answer—usually it is improved (Berardi-Coletta et al., 1995; Ericsson & Simon, 1998). However, for a think-aloud interview to be successful, the material that the student explores or solves must be very carefully chosen. Since Step 2 is one of the initial steps of FASI development and the goal here is to find the material that is appropriate for your test, it’s typically necessary to deviate at times from the think-aloud protocol to ask questions that probe certain areas more directly. These interviews may require the interviewer to ask students to explain their answer or to say more about what a particular term or concept means to them. Specific sorts of questions that can be useful are to ask students to explain how they think about or answer exam or homework questions, particularly where students in a course do worse than the teacher expected. Topics that teachers have seen student difficulties or misconceptions on are also useful to ask about. If creating a test about perceptions, it is useful to explore things that students say about how they study or learn the subject, or what they see as constituting mastery in the subject that are in conflict with what experts believe. For example, ‘I am well prepared for this physics test, I made up flashcards with all the formulas and definitions used in this course and have spent the last week memorizing them’.

During the interview, it is important to get the students to say what they think, rather than what they think the interviewer wants to hear, so one needs to be careful to avoid cuing students to think or respond in a certain way. An interview should never be a conversation. Deviating from the strict think-aloud protocol is more challenging because the interviewer still must spend most of his/her time simply listening quietly and resisting any urge to interject; however, occasional prompts are necessary. For example, to ask, once a student feels that s/he has finished an answer, ‘why did you choose to use this process’ or ‘what does this term mean to you?’. These are very minimal probes but enough to flesh out the details of why students chose the concepts/strategies that they used and what those mean to them. Because these sorts of probing questions do alter student thinking and could likely help students think of connections they may not have in an actual testing situation, strict think-aloud interviews must be performed for validation once the test is constructed. See Section ‘Carry out validation interviews on test questions’ for details.

It is often very useful to have an independent source listen to the recording or watch the video of an interview, particularly with an inexperienced interviewer, to see if the interviewer participation was kept to a minimum and the interpretation of the student responses was accurate. When possible, it is even better to have an experienced interviewer sit in with the new interviewer for the first one or two interviews. Students respond quite positively to being interviewed with two interviewers when one is in training. We have seen that in addition to having good interview skills, an interviewer must also have a high level of subject expertise in order to properly interpret what students are saying about the content, particularly to detect when it may deviate from expert-like thinking in subtle ways.
All interviews should be recorded either via audio or video with audio. Immediately following each interview and before the next interview, the interviewer should spend roughly half an hour summarizing the interview. Some interviewers find it useful to listen to or watch the recordings to check their summaries, but we find this becomes redundant with experienced interviewers who listen closely. As one is simply trying to get consistent general ideas of student thinking from such interviews, it is seldom worth the time and expense of transcribing and individually coding such interviews.

Create open-ended survey questions to probe student thinking more broadly in test form. Once patterns in student thinking begin to appear, then the data from help sessions and interviews can be coded more systematically to identify the type and frequency of student thinking about any particular topic. One should pay particular attention to student thinking that is most conspicuously different from that of experts in ways that faculty members may not expect. FASI questions that demonstrate such unexpected thinking are likely to have the most impact on instruction because they surprise and inform the teacher. These questions are also often the most sensitive to improved methods of instruction, because the faculty members did not realize the problem and hence were doing little to address it. As an example, chemistry faculty members at the University of Colorado created a short FASI-type instrument for physical chemistry. One of the questions addressed the notoriously difficult concept area of gas laws and isotherms. After seeing poor results one semester on this question (only a 5% increase in score), the faculty member developed a series of clicker questions to help the students work through the concepts and had a 44% gain the following semester.

It is difficult to give a simple number for how many students should be interviewed, as this depends so much on the topic and the student responses. If there is a great range of student thinking on a particular topic it will require more interviews. Twenty is probably a reasonable upper limit however. If certain common patterns of thinking have not emerged with 20 interviews, it probably means that the thinking on the topic is too diverse to be measured, and hence the topic is not suitable for such an instrument. More typically, within a dozen or so interviews, one will have a sufficiently clear impression to know which topics should be written into FASI questions.

Guided by the expert and student interviews, one then creates open-ended questions that probe student thinking on topics where there are apparent discrepancies between expert thinking and that of students. These open-ended questions are then given to an entire class. This tests, with a larger sample, the issues that arose in the student interviews. When possible, phrasing the question in terms of actual student wording that emerged during interviews is most likely to be interpreted as desired by the students. Examples of productive questions are ‘Describe what an electric field is and how you picture it’ or ‘How long did this feature take to form?’ (see Figure 1).
The responses from the class(es) to these questions are then systematically coded according to the patterns of student thinking, similar to what is done for the interviews. These open-ended responses from students provide the best source of possible answer choices to be used in the FASI multiple-choice test.

Create forced answer test that measures student thinking. As discussed above, there are major practical benefits to a FASI composed of multiple-choice questions. When appropriate distracters have been chosen, the test not only easily provides the teacher with information about how many students get the answers correct but also provides information on the incorrect thinking. This matches the NRC guidelines: ‘assessments … should focus on making students’ thinking visible to both their teachers and themselves so that instructional strategies can be selected to support an appropriate course of future learning’ (2001, p. 4).

Classroom diagnostic tests that are developed to assess individual students on a particular topic sometimes use other question formats including two-tier. These assessments are focusing on representative coverage of concepts within a topic area so that the researcher/instructor can characterize in detail student learning on a particular topic. Two-tier questions first ask a multiple-choice question—typically a factual type question with only two possible answers. This is followed by asking a second multiple-choice question about the ‘reason’. Such two-tiered questions, while valuable for guiding instruction, are not ideal for the goals of a FASI-type instrument, because FASI instruments aim to have a minimum number of questions, all of which focus on student reasoning, and so would be compromised predominantly of the second portion of the two-tier questions. Brevity and ease of scoring and interpretation are more important for a FASI than detailed characterization of student learning. For these reasons, two-tier questions are outside the scope of this paper.

The primary challenge in creating good multiple-choice questions is to have incorrect options (distracters) that match student thinking. Typically three to five distracters are offered, although there are exceptions. Actual student responses during interviews or to open-ended questions are always the most suitable choices for the multiple-choice question responses, both incorrect and correct (Figure 1). This is the language students are most likely to understand as intended. If one cannot find suitable distracters from among the student responses, then probably the question is unsuitable for use in a multiple-choice form. Care should be taken that wording of the distracters does not inadvertently steer the students toward or away from other answers if they are using common test-taking strategies (Figure 2). For example students avoid options if the answer does not seem ‘scientific’ or involves a statement of absolutes such as ‘never’ or ‘always’. They also look to see which choices have different length or grammatical form.

Some teachers are bothered by providing distracters that are so inviting to students that they can score lower than if they simply selected answers at random. However, that emphasizes the fundamentally different purpose between these tests and standard summative tests used in classes. Here the purpose is to determine...
student thinking, and so all options should be very inviting to at least some students. The incorrect options should match established student thinking, which should result in student answers that are as non-random as possible.

When creating possible answers for multiple-choice questions, a danger to avoid is the creation of answers that have multiple parts. A particularly common failing is to ask questions of the form, ‘A is true because of B’. There are two problems with this form of question. First, it takes much more time for a student to read and interpret because it requires combining multiple ideas that may be tightly linked for an expert in the subject but are far less closely linked for the student. Second, the student may well believe A is true, but not for the reason given in B, or s/he may believe the reason B but not that A is true. So the student will then need to struggle over how to interpret the question and what answer s/he should give, and it makes interpretation of his/her responses problematic. We have consistently seen difficulties with these types of multiple-part answers in our validation interviews.

In creating a test that measures perceptions, rather than concepts, it is typical to have the item format be statements rather than questions. Students respond on a likert-scale ranging from ‘strongly agree’ to ‘strongly disagree’ (Crocker & Algina, 1986; Kachigan, 1986). Because these take students much less time to answer than conceptual questions, the upper limit for a perceptions survey is about 50 simple clear statements. The limitations on how FASIs can be administered argue strongly against any concept test requiring more than 30 minutes for all but the slowest students to complete. This typically means 20–30 clear multiple-choice questions.

*Carry out validation interviews on test questions.* Once multiple-choice questions have been created, they need to be tested for correct interpretation by both experts
(teachers of the subject) and students. In the case of experts, one needs to verify that all agree on the correct response and that the alternative answers are incorrect. All experts also need to agree that the question is testing an idea that they expect their students to learn. Typically only 6–10 experts might be interviewed on the test itself, as there is normally a high level of consensus; however, several times that number need to take the test to ensure consistent expert answers. Teachers will often have suggestions on making the wording more accurate. It is not unusual to have situations where teachers want to word questions to be more technical and precise than is actually necessary or suitable for optimum probing of student thinking. It is necessary in those cases to try to find suitable compromises that teachers will accept, but are still clear to students.

It is considerably more work to produce appropriate wording for students than for teachers, often involving multiple iterations and testing with student interviews. Student interviews are necessary to verify that students interpret the question consistently and as intended. It is also necessary to verify that students choose the correct answer for the right reasons and that each wrong answer is chosen for consistent reasons. Figure 2 is an example of a question, which all experts felt had clear appropriate wording and answer choices; however, when students were interviewed they were able to choose the correct answer without any reasoning about genetics. Twenty to forty student interviews are typically required to establish validity with good statistical confidence. We stress how essential these interviews are. Unfortunately, it is not unusual to encounter test questions, even those used in very high-stakes contexts, that multiple experts have examined carefully and on that basis they are considered valid. However, in interviews on test questions, we have consistently seen that multiple experts can conclude that a question is perfect, but then students can have a completely reasonable but quite different interpretation of the question from what was intended. Ding, Reay, Lee, and Bao (2008) have also observed differing student and expert interpretation of questions.

Figure 2. Multiple-choice question from the GCA (Genetics Concept Assessment) which had to be reworded due to the correct answer being chosen for the wrong reasons during student interviews (Smith et al., 2008)

A single DNA nucleotide change of an A to a T occurs and is copied during replication; is this change in DNA sequence necessarily a mutation?

- (a) Yes, it is a change in the DNA sequence.
- (b) Yes, but only if the nucleotide change occurs in a sex cell (sperm or egg).
- (c) Yes, but only if the nucleotide change occurs in the coding part of a gene.
- (d) Yes, but only if the nucleotide change occurs in the coding part of a gene and alters the amino acid sequence of a protein.
- (e) No, because A and T are similar enough, they can substitute for each other.

Student who earned a D in genetics: 'I don’t like to see the word only in answers. Answers with only are never true. There are four yes answers and one no, so I will go with answer (a)'.

Figure 2. Multiple-choice question from the GCA (Genetics Concept Assessment) which had to be reworded due to the correct answer being chosen for the wrong reasons during student interviews (Smith et al., 2008)
Student interviews are much more sensitive than interviews of teachers. With teachers, you want their opinions; with students you’re trying to probe their thinking. This is where work in cognitive science can provide methods that allow the interviewer to learn about student thinking without altering it. These interviews differ from the interviews used to determine student thinking about concepts as described in Section ‘Selected interviews and observations to understand student thinking about these topics and where and how it deviates from expert thinking’. Interviews on the test items must follow a strict think-aloud protocol (Ericsson & Simon, 1998). To get accurate information from the interviews it is important that the interviewer does not alter the student thought processes by asking questions. The student should be put in an environment very similar to that in which the test will be administered. Thus, the students sit down and are given the test to fill out just as if they were doing it in a classroom setting. While they are filling out the test, they are asked to think out loud. The interviewer should only say things like ‘please tell me your thoughts as you go’; but, never ask them to explain how they interpreted each question and why they chose the answer they did as this is likely to alter the thought processes (Berardi-Coletta et al., 1995):

When participants are thinking aloud, their sequences of thoughts have not been found to be systematically altered by verbalization. However, when participants are asked to describe and explain their thinking, their performance is often changed—mostly it is improved. (Ericsson & Simon, 1998, p. 182)

It requires some preparation to put people into a comfortable think-aloud mode. We always start interviews with 5 or 10 minutes of ‘icebreaker’-type questions to get the student comfortable with the interviewer. For example, asking them about their major, year in school, classes they like or dislike, future career plans, etc. Often the interviewer will follow the icebreakers with some practice think-aloud exercises. This provides practice for the game of thinking aloud so that interview information from the very first question on the FASI will be valuable. It is sometimes difficult to interpret student thinking from think-aloud interviews since ‘inner speech appears disconnected and incomplete’ (Vygotsky, 1962, p. 139), but asking for clarification must be avoided. For example, some students use the strategy of picking an answer as soon as they see one they like without reading all the choices. If the interviewer asks for an explanation of each possible choice, they are no longer seeing the student in authentic test-taking mode.

After the interviewer and student have used the think-aloud protocol to go through the entire test, then the interviewer can go back and ask for some further explanation on each item. But these explanations must be considered with care as they do not represent the student thinking while taking the test. The responses will include some reflection and new thoughts that are generated by the interviewers’ questions and the need to turn the students’ thoughts into intelligible explanations. However, these follow-up explanations can provide some useful information such as why students skipped over certain options (e.g. it did not look scientific enough) or identify options that for the student have multiple ideas contained in a single answer.
By definition, these interviews only provide validation evidence for the population that is interviewed. Therefore the broader the range of students used in the validation interviews, the more broadly the FASI can be safely used. Consider interviewing students of both genders, various ethnic backgrounds, academic specializations, and academic performance. It is typical to have to go through two or three and sometimes more iterations of the questions and possible answers to find wording that is consistently interpreted as desired, so that when the correct expert answer is chosen, it is because the students were thinking in an expert-like manner, and when the students select an incorrect answer it is because they have the non-expert-like thinking the choice was intended to probe.

**Administer to classes and run statistical tests on the results.** The final step of the development is to administer the test to several different classes and then perform statistical analyses of the responses to establish reliability and to collect further evidence for validity. Class sizes of a few hundred or more are desirable but fewer will suffice in many cases if the statistics are handled carefully. There are many psychometric tests that will provide useful information; however, many of the commonly used statistical tests are specifically designed for assessments that measure a single construct or factor. One characteristic of FASIs is that they usually measure thinking about multiple concepts, so the results of statistical measures must be interpreted accordingly.

**Reliability.** Traditionally, three broad categories of reliability coefficients have been recognized (AERA et al., 1999):

1. Administer parallel forms of the test in independent testing sessions. Then calculate an alternate forms coefficient;
2. Administer the test to two equivalent populations and obtain a test–retest stability coefficient; and
3. Calculate internal consistency coefficients that measure the relationship of individual items or subsets of items from a single administration of the test.

These three types of sampling variability are considered simultaneously in the more modern *generalizability theory* to create a standard error of measurement (Cronbach, Gleser, Nanada, & Rajaratnam, 1972).

Internal consistency coefficients, (3) above, can also be described as measures of task variability. Because the goals of a FASI include probing multiple concepts with a minimum number of questions and it is not designed to accurately measure the mastery of an individual student, task variability is not a good reflection of reliability of the instrument. The time required to create a parallel validated form of a FASI, as in (1), vastly exceeds the benefits. This makes (2), administering the test to two equivalent populations and obtaining a test–retest stability coefficient, also described as sampling variability due to occasions, the primary method for measuring reliability of a FASI. Note that all three forms of reliability listed above apply to
the reliability of the instrument when used on a group and not for individuals. Individuals may have fluctuations that will average out when a group is evaluated as a whole.

A test–retest stability coefficient measures the consistency of test results if the same test could be given to the same population again under identical circumstances. Of course this is impossible because it would require that giving it the first time does not have any impact on the test takers or that they have not changed in any other way between the first and second administrations. However, when administering tests to large university courses (enrolment over 150), one has the ideal situation. The test can be administered again the following year to the same course. The population of students who enrol in a course is very similar from year to year if the university maintains constant admissions criteria. Each year’s students have similar preparation for the course, similar experience in college and are of similar demographic composition from year to year. The FASI should be given at the very beginning of each course and then a Pearson Correlation Coefficient can be calculated between the two sets of results. For the FASIs we have been involved with creating, we consistently see coefficients over 0.90 when they are administered in this way; but, there is no agreed upon accepted value.

It is quite common to see the statistic Cronbach’s $\alpha$ or the Kuder–Richardson reliability index (KR-20) quoted as a measure of reliability (Cronbach, 1951; Kuder & Richardson, 1937). These indices would fall under (3) above, internal consistency coefficients. They are primarily useful for a single-construct test. Both indices depend on both the correlation between questions and the number of questions (Crocker & Algina, 1986; Field, 2009). However, in the words of Cronbach: ‘Coefficients are a crude device that do not bring to the surface many subtleties implied by variance components. In particular, the interpretations being made in current assessments are best evaluated through use of a standard error of measurement’ (Cronbach & Shavelson, 2004, p. 394). In fact, having a high correlation between items, which results in a higher value for $\alpha$ or KR-20, means that these items are repetitive. The way a FASI is typically administered puts a premium on minimizing the time required to complete the assessment and hence the number of questions. So a low Cronbach’s $\alpha$ or KR-20 on a FASI would be quite reasonable, and a high Cronbach’s $\alpha$ or KR-20 on a FASI does not guarantee that the test will be more reliable for its intended use and may be an indication that there are redundant questions that should be removed.

**Item analysis.** Item analysis is a term broadly used to describe any statistical property of examinees’ responses to an individual test item. For FASIs we have found that *item difficulty, item discrimination, and point biserial correlation* provide useful information. Knowing these various statistics for each question helps describe how the questions on the test relate to each other and the test as a whole. It is also useful to provide this information when the test is published to inform test users about what sorts of values are likely to be observed for each question. However, as
previously mentioned, these statistics are often used for single factor summative assessments of individual students, so the more common interpretation of their values and acceptable cut off ranges do not apply with a FASI.

Item difficulty is simply the percentage of students who got the item correct. It is valuable to have a range of difficulty levels. A wider range means the assessment can provide feedback on student learning for a range of student backgrounds and levels of mastery (majors and non-majors courses for example). If all items are very difficult for someone who has not had the course, it is possible that the FASI will only be valuable when given as a post-test (McKagan et al., 2010)—more detail is given in the Test Administration Tips section at the end of this paper.

The discrimination index is a measure of how well an item differentiates between strong and weak students, as defined by their overall test score. \( D = p_u - p_l \) where \( p_u \) is the portion of students in the upper group who answered the item correctly and \( p_l \) is the portion of students in the lower group who answered the item correctly. There is no agreed upon percentage that should be used to determine these groups. We have seen between 27% and 50% used in the literature, but when working with large numbers of students, the results of the discrimination index will not fluctuate substantially based on your choice (Crocker & Algina, 1986). Again, the standard criteria often stated for the ‘desirable’ range for the discrimination index are not useful, as these are established based on single-construct tests designed to achieve maximum discrimination between students.

The discrimination index can be useful for identifying some important different types of questions. A question that nearly every student gets wrong before instruction but nearly all get correct after good instruction would have a very low discrimination index. However, such a question would be quite valuable on a FASI, as it would discriminate effective from ineffective teaching (formative assessment) by showing that most students were mastering the item in question when taught well. From a completely different perspective, namely teacher psychology, it is good to have at least a few questions on the test where most students do badly at the beginning of the course and well at the end for all reasonable forms of instruction. They show the teacher that students are learning something and the test is measuring student learning. If a teacher saw that there was no improvement on any of the questions, the natural reaction may be to conclude the students were hopelessly incompetent or the test was meaningless. Showing that students are learning some concepts but not others will be more likely to get the teacher to think about how they can make changes to improve their teaching. Another desirable type of question with very low discrimination index is one that probes understanding of a vital concept and the results show that no student learned it. The results of this type of question tend to have a large impact on teachers, particularly if the concept being measured is one that the teacher had erroneously believed students were learning based on results from typical course exams. Physics education researchers have found a number of such psychologically powerful examples (Mazur, 1997).
The point biserial coefficient ($\rho_{\text{bis}}$) provides an additional measure of how consistent each question is with the whole assessment, and so the same general caveats to using standard ranges of desirable values apply as have already been discussed for Cronbach’s $\alpha$. This statistic is simply the Pearson Correlation Coefficient between a dichotomous variable (one that can only take on the values 0 or 1) and a continuous variable (test score). If there is a particular concept that students learn much better or worse than the other concepts, you want to know that, particularly if you see that those relationships vary considerably with different student populations and courses. For a question that revealed such variation, the standard point biserial coefficient would be much lower than is considered acceptable in traditional assessments. But for a FASI, that would be a fine question. It identifies certain concepts that are much more difficult for students to learn, and/or it identifies that this learning is quite dependent on the specifics of how the concept is taught in a way that is not true for the course material and other concepts as a whole. This provides very valuable feedback for improving how that topic is taught. So a low biserial coefficient can often be considered more desirable, because that shows that the question is probing a particular concept more specifically, rather than testing general knowledge.

Item Response Theory (IRT) is commonly used to create a response curve (probability of a student with a particular ability to answer the question correctly) for each item and/or to create a scaled score for the whole test based on what is known about each item. Practical uses of IRT are the construction of equivalent test forms or development of tests that discriminate at a particular level of ability (Crocker & Algina, 1986). Neither of which are goals for a FASI. In addition, the underlying assumption of IRT is that student scores on a test are based on a single latent trait—some general ability that may or may not be what the test was designed to measure. There are ways to handle multiple traits but these become much more complicated and are not as well understood. However, the idea of knowing more about which students answer an item correctly is of potential interest to developers of a FASI. It is also interesting to create an item response curve for each distracter to learn which students choose each distracter (Ding & Beichner, 2009).

An alternative to IRT is to use total scores as a substitute for the latent trait abilities to create an item response curve for each item (and its distracter if desired) on the FASI. This can be done by graphing the proportion of students who choose the correct response versus the total test score. This is a much easier method that will render useful results for a FASI (Ding & Beichner, 2009; Morris et al., 2006). Full on IRT requires specialized statistical software and large sample sizes from 200 (1 parameter Rausch—arguably not appropriate for a FASI) to 1,000 students (Crocker & Algina, 1986).

Test analysis. There are a few statistics that characterize the test as a whole, for example reliability measures. Another useful whole test statistic is Ferguson’s delta ($\delta$) or the coefficient of test discrimination (Ferguson, 1949). This statistic measures the
Development and Validation of a FASI

Discriminatory power of an entire test by investigating how broadly the total scores are distributed over the possible range. Ferguson’s delta ranges from 0 to 1. A test with $\delta > 0.9$ is considered to be a test with high discriminatory power. Note that this statistic also depends on the population of students used to calculate $\delta$ so it can be interesting data when comparing different courses.

Correlations with other measures such as course outcomes. Validity evidence can be added to the FASI by calculating the Pearson Correlation Coefficient between the FASI results and other measures that are valued in the course such as exams, homework, course questionnaires, or other standard tests. These correlations can be used to establish that the FASI is measuring something that teachers care about. It can also inform the teachers about the incoming thinking of their students. We see this as adding to the evidence of validity in four different forms: (1) Evidence based on relationships to other variables, which includes correlations with course outcomes or other standards; (2) Predictive and concurrent evidence—does it predict and show what it claims; (3) Convergent evidence—the FASI results relate closely to course grades; and finally (4) discriminate evidence—a FASI that measures understanding of physics concepts should relate closer to a test of logical reasoning skills than it does to a test of the students’ writing ability for example.

With many FASIs only some of the above correlations were studied before introduction of the FASI. A few examples can be found in Adams et al. (2006), Hestenes and Halloun (1995) and Smith et al. (2008). However, many of these sorts of correlations were done in later studies. All of them are by no means required before a new FASI is considered to have enough evidence of validity and reliability to be published. As this research is conducted for a variety of courses and universities, it strengthens the validity generalization of the FASI and each new set of data and correlations makes for useful research studies of their own.

Factor analysis. A factor analysis uses student responses to determine groups of questions that are answered in a correlated manner by students, thus indicating aspects of student thinking that are closely linked. The usefulness of such analysis depends on the design and purposes of the FASI. Some assessments have presented categories based on the author’s expert view of the questions; but were criticized because these expert categories did not hold up under a factor analysis (Heller & Huffman, 1995; Hestenes & Halloun, 1995; Huffman & Heller, 1995). The factor analysis revealed that students did not answer questions within an expert category consistently, demonstrating that these categories do not represent single consistent ideas in students’ minds. This is relevant when using the scores of individual categories to inform teaching about the particular concept the category may be identifying. With perception surveys, it is especially useful to perform a factor analysis because perceptions can be broad and novices may organize these ideas quite differently from experts.
Factor analysis is data-intensive, requiring a minimum of 10 times the number of responses as questions to carry it out in a statistically meaningful way. There are also multiple-factor analysis methods, a discussion of which is beyond the scope of this paper. See Ding and Beichner (2009), Heller and Huffman (1995), Hestenes and Halloun (1995), Huffman and Heller (1995), and Ramlo (2008) for discussion of factor analysis performed on concept inventories and Adams et al. (2006) for discussion of factor analysis for perception surveys.

**Test Administration Tips**

Because there are many subtle variations to how a FASI can be administered, when the test is published, it is important to include details of how the developer administered the test when validating it. This is both in keeping with AERA, APS, and NCME guidelines for educational and psychological testing and ensuring that results will be appropriately comparable. Here, we will describe the methods we have found to be effective and discuss how variations on these could affect the validity of a test result.

It is important to get as high a response rate as possible and have the students take the test seriously. However, there are obvious problems with having a test like this count directly toward a student’s grade, because then the students will be motivated to learn about the test questions and memorize the answers without understanding them, potentially making the test worthless.

Here is the process that we have found to be effective at achieving the goals of a conceptual FASI. The exam is given on the first day of class, in class. We make it clear to the students that it will not count in their course grade, what is important is that they complete it, and that it will benefit them to do so, because it will improve instruction in the course by establishing their level of prior knowledge. We tell them that they are not expected to know the correct answers, what we are interested in is their current thinking about these topics, and they should put down their best guess but not agonize over the question if they do not know the answer. Nearly all students take the exam seriously when given under these circumstances. Students should be required to turn in both their answer sheets and the test itself to reduce the chances of test questions being circulated.

We have found we can get a serious effort from nearly all students in a course by giving the ‘post’ version of the test in class on the next to last day of class, with the guidance to the students that they will not be graded on their answers, but that these are all questions on important topics for the course and hence will provide guidance in studying for the final exam and that the FASI results will be reviewed and questions that students do particularly poorly on will be reviewed as part of the exam review. Again, students should be required to turn in both their answer sheet and the test itself.

In some courses, such as introductory quantum mechanics, students have no previous experience with the material, and so their pre-instruction scores are simply random guessing. Once a teacher has given the FASI and verified this lack of
incoming knowledge on the subject, it is adequate in subsequent offerings of the course at that institution to only measure post-instruction results. However, it is never safe to assume knowledge of incoming student thinking about a subject without measuring it. Unfortunately, it is not adequate to assume preparation based on past courses and grades they have received.

On tests of expert-like perceptions, students always have opinions and so it is important to measure both pre and post. Pre-information is a way to learn about the population; however, to learn about the impact of instruction it is particularly important to only consider results that have matched pre- and post-tests for individual students, because selection effects are so important in such tests. It is reasonable to expect a strong correlation between a student’s perceptions about the subject and whether or not they will be inclined to fill out a survey. Selection effects exist for tests of conceptual mastery as well, but they are not as tightly coupled to performance on the test.

Tests of perceptions can be given online with less risk of students consulting other sources to try and find a ‘right’ answer, and so for simplicity of data compilation and to avoid using class time, they are commonly administered online during the first and last weeks of a course. It is good to give the pre-test as early as possible, as our studies have demonstrated that students’ perceptions are affected by the course very early in the semester. There are particular challenges to getting students to put serious thought into filling out online assessments, and we have developed some strategies to achieve this (Adams et al., 2006). A small number of bonus points are given to the students to provide incentive to fill out the test, reminder emails are sent out a couple of times during the week. Some instructors also find it effective to add the survey to the first homework assignment where students get the credit for simply completing the survey. Criteria such as dummy questions and time used to fill out the test are used to filter out tests that were not answered with adequate care.

If the test is altered—removing question(s), adding question(s), changing the wording of question(s), etc.—results cannot then be compared with the published results, and in accordance with AERA, APS, and NCME guidelines this fact should be noted when the test is published. The reliability and validity were collected for the FASI as it was constructed only, and the value of comparisons with the results of modified tests is quite limited.

Some authors publish their instrument with the development and validation article while others only distribute the actual test when requests are received. Opinions are divided on this topic; however, the FCI and CLASS appear to still be quite effective tools and are publically available. While there was fear that students would get answers, invalidating the tests, there is no indication that this has happened, even for the FCI, the oldest and most widely used of such tests. Scores on the FCI have remained stable over decades in the absence of pedagogical changes. This is not surprising. Because the results of these tests typically do not count toward the students’ grade, they have no incentive to take the time required to search out the answers and memorize them. If there is concern about compromising the questions,
Maloney, O’Kuma, Hieggelke, and Van Heuvelen (2001) suggest not using the real name of the FASI when administering to students. They also name the file differently so that web searches using the official test name do not pull up the test.

Another common concern is that it is possible that since students have seen the FASI on the pre-course test it will inflate their post-course results. The empirical evidence is that such an effect is very small, as there have been many examples of pre–post improvements (gain) that are actually zero, most notably the extensive results obtained over many years with the FCI (Hake, 1998) by many instructors. In our experience, a surprisingly large fraction of the time, students will indicate they do not even remember taking the pre-test, let alone remember any questions from it. With many years of giving multiple FASI tests, we have yet to encounter a student who brought up one of the FASI pre-test questions at some point during the course. This lack of retention is not so surprising when one considers that these are questions students answer with no stakes involved, usually the student does not know how to answer most of them, they get no feedback on the answers, and they usually see each question for well under a minute and then not again until several months later.

Summary

In 2001, the National Research Council called for research on new forms of assessment that can be made practical to measure the effectiveness of instruction. They noted that the most serious barrier in traditional assessment design is the requirement of a team of experts (scientists, educators, task designers, and psychometricians) for creating a new test. They also recognize that further work is needed to integrate what is known from cognitive science into assessment design.

Many tests of instruction, both conceptual understanding and perceptions about a discipline and how it is best learned, have been created in the sciences. These tests were developed and validated using many of the same general procedures. However to our knowledge, these procedures have not been written down. Here, we described how a formative assessment of instruction can be created without a team of interdisciplinary experts by using student and expert interviews with methodology from cognitive science. The assessments that have been created using these procedures have been validated and published in peer-reviewed journals and widely adapted on an international scale. This provides evidence that this method is successful at creating a useful instrument for formative assessment of instruction.

Acknowledgements

We would like to thank the members of the Carl Wieman Science Education Initiative (CW-SEI) at the University of British Columbia and the Science Education Initiative (SEI) at the University of Colorado, especially Sarah Gilbert, Laurie Langdon, and Jennifer Duis for their helpful comments. This work was supported by
the University of Colorado through the SEI and the University of British Columbia through the CW-SEI.

**Note**

1. We use this statement to discard the survey of people who are not reading the questions: Please select agree-option four (not strongly agree) for this question to preserve your answers.

**References**


What Research Tells Us About Good Assessment

As our experience with alternative measures grows so does our knowledge, but we still have far to go to verify that these promising new approaches promote quality education.

JOAN L. HERMAN

Educational assessment is in a process of invention. Old models are being seriously questioned; new models are in development. Open-ended questions, exhibits, demonstrations, hands-on experiments, computer simulations, and portfolios are a few examples. The promise of the new approaches is alluring and is being effectively advanced at the national, state, and local levels all over the country.

Although the new designs offer potential, what we know about them is relatively small compared to what we have yet to discover. How close are we to having the assessments required? Here I summarize the research supporting current beliefs in testing, identify qualities of good assessment, and review the current knowledge on how to produce such measures.

Does Assessment Support Change? Interestingly, much of the research supporting the power of testing to influence schooling is based on traditional standardized tests and concludes that such tests have a negative impact on program quality. A number of researchers — using teacher surveys, interview studies, and extended case studies — have found that accountability pressures encourage teachers and administrators to focus planning and instructional effort on test content and to devote more and more time to preparing students to do well on the tests (Dorr-Bremme and Herman 1983, Herman and Golan 1991, Kellaghan and Madaus 1991, Shepard 1991, Smith and Rottenberg 1991). Insofar as standardized tests assess only part of the curriculum, many of these researchers conclude that the time focused on test content has narrowed the curriculum by overemphasizing basic-skill subjects and neglecting higher-order thinking skills. Herman and Golan (1991), among others, have noted that such narrowing is likely to be greatest in schools serving at-risk and disadvantaged students, where there is the most pressure to improve test scores.

Cheerier pictures emerge, however, when assessments model authentic skills. Studies of the effects of California's 8th grade writing assessment program, for example, indicate that it encourages teachers both to require students to write more and to give them experience in practicing writing in a wider variety of genres.

Beyond impact on instruction, studies of some states and districts have found improved student performance over time associated with new assessment programs (Chapman 1991, Quellimatz and Burry 1983). One district in southern California, for instance, involved its teachers in the development of an analytic scoring scheme for assessing students’ writing and trained a cadre of teachers from each school in its use. Over the next several years, the district witnessed an improvement in students’ writing, which it attributed to the districtwide standard, the focus it provided for teachers’ instruction, and the district’s attention to writing instruction.

This latter point is important in interpreting both the district and the state stories: change in assessment practices was one of several factors that potentially influenced teachers’ and students’ performance. The California Writing Project and a number of statewide training efforts occurring at the same time gave teachers effective models of writing instruction and stressed the importance of giving students ample opportunities to write.

Pressure to improve tests scores, in the absence of serious, parallel supports for instructional improvement, however, is likely to produce serious distortions. In 1987, John Cannell, at that time a pediatrician in West Virginia, was surprised to read that the students in his state had performed above the national average on the statewide assessment program. If the largely disadvantaged students in West Virginia were scoring above the national average, who, he wondered, might be scoring below it? When he contacted the states and many large districts, almost all reported scoring above the national norm sample — a finding that was essentially replicated by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) using more rigorous
methods (Linn et al. 1990).

How can all students be performing "above average," a clear contradiction in the meaning of performance? Shepard concludes that the answer in large part lies in teachers' directly teaching to the test, often providing daily skill instruction in formats that closely resemble tests. Shepard and her colleagues have found that such improvements in test scores do not generalize to other measures of achievement (Koretz et al. 1991). In other words, superficial changes in instruction to improve test performance are not likely to result in meaningful learning. As a result, scores no longer represent broader student achievement, but only the content and formats included on the tests.

Ellwein and Glass (1987), looking at the effects of minimum competency testing and other assessment-based reforms, illuminate other potential distortions when serious consequences follow from test results (Glass and Ellwein 1986). When policymakers and others try to raise standards based on test results, "safety nets are strung up (in the form of exemptions, repeated trials, softening cut-scores, tutoring for retests) to catch those who fail"; and, further, "standards are determined by consideration of politically and economically acceptable pass rates, symbolic messages and appearances, and scarcely at all by a behavioral analysis of necessary skills and competencies" (Glass and Ellwein 1986). Shaped by political realities, as well as concerns for equity and future consequences, test-based standards often become diluted and have little or no influence on teachers and their instruction or on students and their learning.

**What Is Good Assessment?**

These findings aside, a number of current policy initiatives show continuing optimism in the power of good assessment, finding the problem with the assessments used, not with the basic model of accountability.

**Good assessment is built on current theories of learning and cognition and grounded in views of what skills and capacities students will need for future success.** To many, good assessment is also defined by what it is not: standard, traditional multiple-choice items.

According to cognitive researchers, meaningful learning is reflective, constructive, and self-regulated (Bransford and Vye 1989, Davis and Maher 1990, Marzano et al. 1988, Wittrock 1991). To know something is not just to have received information but to have interpreted it and related it to other knowledge one already has.

Recent studies of the integration of learning and motivation also highlight the importance of affective and metacognitive skills (McCombs 1991, Weinstein and Meyer 1991). For example, research suggests that poor thinkers and problem solvers differ from good ones not so much in the skills they possess as in their failure to use them in certain tasks. Competent thinkers or problem solvers also possess the disposition to use the skills and strategies as well as the knowledge of when to apply them.

The role of the social context of learning in shaping cognitive ability also has received recent attention. It has been noted that real-life problems often require that people work together as a group. Further, groups may facilitate learning by modeling effective thinking strategies, scaffolding complicated performances, providing mutual constructive feedback, and valuing the elements of critical thought (Resnick and Klopfer 1989).

**Can We Ensure Quality?**

Our new understandings of the nature and context of learning have supported the movement toward alternative assessments. A CRESST project illustrates the current enthusiasm; it has located more than 171 examples, representing the active efforts, conservatively, of 19 state departments, more than 30 school districts, and a dozen other groups.

Assuring the quality of the new assessments poses significant R&D problems. Face validity, that an assessment appears to be assessing complex thinking, is not sufficient. Essential is the notion that students' performance represents something of importance, something beyond the specific task assessed.

At the simplest level, validity indicates whether test scores accurately reflect the knowledge, skills, and abilities they are intended to measure. For traditional multiple-choice measures, concerns for validity have focused on issues of reliability (stability and consistency of performance) and patterns of relationships that may suggest whether the assessment is tapping the intended construct.

While these traditional notions are still applicable, Linn and colleagues (1991a) call for additional criteria for judging the quality of an assessment:
• Consequences. The consequences of an assessment influence how people respond to its results and, as Cannell's (1987) findings suggest, can rebound to influence the validity of the results themselves. This overarching criterion requires that we plan from the outset to appraise the actual use and consequences of an assessment.

• Fairness. Does the assessment equitably consider the cultural background of those students taking the test? Winfield and Woodard (in press) warn that standardized performance assessments are at least as likely to disadvantage students of color as traditional measures. With Winfield and Woodard, Linn and colleagues (1991a) point to additional equity problems stemming from students' "opportunity to learn" that which is assessed: Have all students had equal occasions to comprehend the complex thinking and problem-solving skills that are the targets of these new assessments?

• Transfer and generalizability. Do the results of an assessment support accurate generalizations about student capability? Are they reliable across raters, consistent in meaning across locales? Research on these issues raises perplexing questions about feasibility.

• Cognitive complexity. We cannot tell from looking at an assessment whether it actually assesses higher-level thinking. Schoenfeld (in press) cites a telling example: An award-winning teacher, whose reputation was based on his students' Regents exam performance, had drilled his students on the geometry proofs likely to appear on the exam.

• Content quality. The tasks selected to measure a given content domain should themselves be worthy of the time and efforts of students and raters. The content must reflect the best current understanding of the field and important aspects of a discipline that will stand the test of time.

• Content coverage. Coverage raises issues of curriculum match and whether the assessment tasks represent a full curriculum. As Collins and colleagues (1990) have noted, if there are gaps in coverage, teachers and students may underemphasize those topics and concepts excluded from assessment.

• Meaningfulness. One rationale for more contextualized assessments is that they will result in worthwhile educational experiences and in greater motivation for performance. However, additional evidence is needed to support this theory, as is investigation into the relationship between alternative assessments and student motivation to do well on them.

• Cost/efficiency. With more labor-intensive, performance-based assessments, greater attention will need to be given to efficient data collection designs and scoring procedures.

How Far Along Are We?
Currently, most developers of the new alternatives (with the exception of writing assessments) are at the design and prototyping stages, some distance from having validated assessments. The CRESTT database project, for example, indicates that few have yet collected data on the technical quality of their assessments or about their integrity as measures of significant student learning.

Knowing how to reliably score essays and other open-ended responses is one area of relative technical strength. Research on writing assessment informs us that: (1) raters can be trained to score open-ended responses reliably and validly; (2) validity and reliability can be maintained through systematic procedures — including specified scoring schemes, sound training procedures, and ongoing reliability checks throughout the rating process; and (3) rater training reduces the number of required ratings and costs of large-scale assessment (Baker 1991, p. 3). Studies Baker reviews from the literature in the military further support the feasibility of large-scale performance assessments and the feasibility of assessing complex problem solving and team or group performance.

Trials in progress in various states, districts, and schools provide similar data. Vermont's experiments with portfolios, Connecticut's and California's pilots of hands-on math and science assessment, and Maryland's integrated assessment also indicate that it is logistically possible to administer these assessments on a large scale, schemes can be devised to score them, and teachers can be trained to reliably score them.

The generalizability of these scores — however reliable the scoring process — remains a challenge as indicated by, for example, the research of Shavelson and his colleagues on hands-on assessments in math and science (1990a,b; 1991). They essentially asked, "How many tasks does one need to get a stable estimate of a student's problem-solving capability in a given topic area?" Their answer varied from one data set to another, but the range is telling: from approximately 8 to 20 tasks were needed to obtain reliable individual level estimates. Further, they (1991) found great variability across content or topic areas within a given discipline: at least 10 different topic areas may be needed to provide dependable measures of one subject. Given the time required for administering a single hands-on experiment, these findings give pause for thought.

Also giving pause for thought are findings from Shavelson and others
which suggest that the context in which you ask students to perform influences the results. Shavelson looked at how students’ performance on science experiments compared with that on simulations and on journals, all intended to measure the same aspects of problem solving. Similarly, Gearhart et al. (1992) compared how students’ performance in writing was judged when based on their writing portfolios, classroom narrative assignments, and responses to a standard narrative prompt. Both studies showed substantial individual variation across the various tasks.

A study by Linn and colleagues (1991b) of the comparability of writing results across different state assessments addresses similarly thorny issues, and ones particularly germane to discussions about a national system of tests to assess progress toward national standards. Under current proposals, national standards are to be articulated and states would develop tests tied to their state curriculums, to assess students’ progress toward those standards. The results might be used for student certification, college admissions, and/or job applications, as well as to evaluate the quality of schooling at the state, district, and school levels. Because of the high stakes potentially associated with students’ performance, concerns for equity demand concern for comparability of results from the different assessments.

Linn and colleagues (1991b) used the results of statewide writing assessments to examine the comparability of results from 10 states. When trained raters used their state’s scoring schemes to score student papers from a different state, the results showed relatively high correlations between students’ scores on the different scoring schemes. The student essays rated as the best, average, and poorest tended to be the same regardless of the specific scheme used. Such a high level of agreement of the relative ordering of student performance, according to Linn, is necessary but not sufficient for any system intending to compare results within a state to a common national standard. Also required is agreement on the absolute standard of mastery; in this area, Linn found rather substantial differences in the level of scores assigned to the same papers by different states, meaning variations in leniency and in absolute standards for performance. Assuring comparability of results, in short, will require more work.

**What Remains to Be Done?**

The following example illustrates both the exciting progress being made across the country and internationally and the problems that will need to be addressed if assessment is to meet its promise.

Building on past experiences with assessment in the service of accountability and on an expanded set of criteria for productive assessment, researchers at CRESST are developing new approaches to assessment, generating appropriate psychometric theory to undergird them, and exploring the process and impact of new alternatives in educational practice. For example, the center’s content assessment project has produced a prototype for assessing the depth of student understanding in specific subjects (Baker et al. 1991).

Starting with students’ understanding of American history, the project developed an approach that asks students to read primary source materials (for example, the Lincoln-Douglas debates) and then write an essay explaining the issues raised in the reading (explain the causes of the Civil War). Essays are then rated for quality of understanding using a scoring scheme that provides holistic and analytic ratings.

On the positive side, this project has demonstrated that it is possible to:

- design comparable, parallel tasks, based on prespecified design characteristics (the same scheme can be used to assess, for example, Civil War history, immigration history);
- use uniform scoring schemes across disciplines;
- use the same assessment to derive holistic information for large-scale assessment and diagnostic information for improving classroom practice.

But importantly, these studies also indicated that student performance on the new kinds of measures is dismally low, a finding shared by most states and districts that have tried such assessments; and that teachers need substantial training and follow-up support in both suitable assessment techniques and appropriate instructional strategies.

In conclusion, progress is being made to clarify the potential of the new alternatives, but substantial challenges remain. We must ensure that assessment supports, and does not detract from, quality education. Assessment practices themselves must be accountable to criteria that define quality assessments. These criteria force attention not only to technical issues but also to consequences of an assessment and to students’ opportunity to learn that which is assessed.

Finally, changes in assessment are only part of the answer to improved instruction and learning. Schools need support to implement new instructional strategies and to institute other changes to assure that all students can achieve the complex skills that these new assessments strive to represent.

**References**


Author’s Note: The work reported herein was supported under the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed herein do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

Joan L. Herman is Associate Director, University of California—Los Angeles National Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation, UCLA Graduate School of Education, 405 Hilgard Ave., Los Angeles, CA 90024-1522.