Assessment of Student Learning

by Laura Slocum and Thomas A. Holme

Thomas Holme *received his Ph.D. from Rice University and is a professor of chemistry at the Iowa State University. He has been the Director of the Examinations Institute of the Division of Chemical Education of American Chemical Society since 2002 and conducts research on how to measure student learning in chemistry as a key component of his scholarship. Contact e-mail: taholme@iastate.edu*

Laura Slocum *graduated from Western Connecticut State University with a B.A. in chemistry and from Ball State University with an M.S. in chemistry. She has taught high school chemistry for the past 16 years in both Connecticut and Indiana. Presently, Laura teaches at University High School of Indiana. She serves as the Secondary School Associate Editor for the* Journal of Chemical Education. *She also serves on the Board for Trustees for the Examinations Institute and as an Alternate Councilor for the Division of Chemical Education. Contact e-mail: lslocum@universityhighschool.org*

Introduction

The importance of assessment in any discipline at any level is nearly self-evident. Ultimately, for pragmatic reasons alone, the nature of assessments plays a major role in the practical formulation of what occurs in the classroom; if the assessments for a course are aligned to well-defined state or national standards and if these same assessments require students to acquire deep understanding, then both students and teachers will respond. More recently, the emphasis on assessment within the No Child Left Behind (NCLB) Act has provided even more motivation for teachers to understand how to better assess their students, particularly, in line with state and national standards.

Nurrenbern and Pickering (1987) made a rather startling observation that "the last 30 years of chemistry can be characterized by the 100 most popular questions." The two decades since this statement have seen important developments in assessment (Pellegrino et al., 2001) to be sure, but a glance at any textbook test bank reveals many of the same questions that constituted chemistry tests in the 1980s are still present in today's exams. Although this longevity may simply speak to the veracity of the fundamental knowledge of chemistry, it also suggests a more careful inspection is warranted regarding the assessment of learning in chemistry. What are the possibilities for constructing chemistry assessments? How do these assessment techniques play a critical role in the development of the teaching of chemistry at all levels?

Significant effort has resulted in the development of assessments related to chemistry in many states. For example, the state of Indiana used a document that detailed student learning objectives known as "Chemistry Proficiencies" prior to the adoption of state academic standards in 2000. On the basis of these proficiencies, the Indiana Department of Education (DOE) and a small group of teachers prepared a test called a Core 40 Exam. This exam was available for chemistry teachers to use, voluntarily, at the end of the school year to test student proficiencies in chemistry. Once the state standards were established in 2000, the DOE started working with various groups of teachers and the Center for Innovation in Assessment at Indiana University to prepare a chemistry test that mapped directly to state standards. This process went through several iterations and in May 2005, the Core 40 Classroom Assessment–Chemistry I was published. This exam has two sections, containing mostly multiple-choice questions with some open-ended response questions at the end of each section. Administration of this exam is not required, but the exam is available online at a password-protected site for teachers to use. Because this exam is not required, no statistical information is collected on student performance. There is presently no required exam in chemistry in Indiana; however, starting in the 2007/2008 school year, there is a required Biology I–End-of-Course Assessment that all students must take. Other states have undertaken similar processes.

Chemistry, as a discipline, also has the distinction of a long-standing effort for nationally normed exams as produced by the Examinations Institute of the American Chemical Society (ACS). The first high school exam was released in 1954, and new exams are produced every two years, as is described more fully below under "Available Resources." Therefore, there exists a range of concepts and guidelines that can be utilized to enhance assessment efforts in high school chemistry. In the remainder of this chapter, we describe important terminology for assessment and survey resources for assessment that are constructed with national standards in mind. We describe in detail how ACS exams are constructed and how teacher involvement in the process leads to important professional development activities. Lastly, we conclude by discussing how the principles of assessment can be incorporated into teacher-designed test items.

Constructing Assessments

Individuals who study to become chemistry teachers typically spend most of their time developing their content knowledge, while comparatively little time is spent studying measurement of student learning via tests. The scholarship associated with developing robust tests (and the questions on these tests) remains an area of active research, but some guiding principles have been established over the years. Regardless of the identity of the test generator—a national society, the state department of education, or an individual classroom teacher—there are two key ideas that provide the fundamental framework needed to assess educational measurement efforts, namely validity and reliability (Popham, 2003).

Table 1. Face Validity of Chemistry Questions

Less face validity	More face validity
1. How long is 4.7 yards in mm? 2. Which process is exothermic, one that has $\Delta H = -315$ kJ or one with $\Delta H = 356$ kJ?	1. A chemical bond is 124 pm long; express this distance in meters. 2. Which energy change could represent a combustion reaction, one with $\Delta H = -315$ kJ or one with $\Delta H = 356$ kJ?

Validity is best considered as a measure of whether or not a test measures what we intend it to measure. There are several categories of validity. The first, and perhaps most apparent form, is referred to as *face validity*. At this level, the question that one might ask of a chemist or chemistry instructor is, "Is the item being considered a chemistry question?" This question may seem almost too obvious to ask, but consider the role of background knowledge, particularly in mathematics, for a chemistry course. It is common for a chemistry teacher to ask questions related to dimensional analysis that are not related to chemistry, such as the conversion of a distance in an obscure unit system (furlongs, for example) into metric distances. Thus, considering face validity, Table 1 suggests two ways that teachers consider asking similar questions, but one way has more face validity and the other less.

The next important category for validity is called *content validity*, and it is a measure of whether the item being considered actually measures the content it is expected to answer. Accuracy of content is incorporated in this concept, but it is richer than accuracy alone. Content validity addresses the question of whether or not a test assesses the skills and knowledge of the course for which it is intended. In Table 2, we present two sample questions related to NSES Standard B1: "*All students should develop an understanding of structure of atoms, structure, and properties of matter… .*" Each sample question is worded in two different ways to illustrate how a question can have more content validity and less content validity.

Table 2: Content Validity of Chemistry Questions Related to NSES Standard B1

Less face validity	More face validity
1. What is the atomic number of Bromine, Br? 2. What is the atom with an electron configuration of $1s^2 2s^2 2p^6 3s^1$?	1. Describe how the atomic number and group location of an element on the periodic table provides information about the element? 2. Describe how the electron configurations of sodium and potassium are similar.

Finally, a more technical component of test writing is called *construct validity*, which measures the extent to which the format, language, and other factors of the question influence the measurement of knowledge. Many questions provide unintentional hints about correct answers (or ways to eliminate some of the incorrect answers, i.e., distractors.) In many cases, a form of statistical analysis, called item statistics, is capable of providing insight into the construct validity, particularly for multiple-choice items.

The two most important concepts related to item statistics are the difficulty and discrimination associated with each question. The difficulty of a question is traditionally defined as the fraction of students who answer it correctly. A high-difficulty index means a large fraction of the students who take the exam answer the question correctly, so it is a relatively "easy" question. Discrimination measures the relative performance on a question between high-proficiency students and low-proficiency students. It is traditionally calculated by looking at the top 25% of students (defined by total score on the overall exam) and the bottom 25% of students. Discrimination is the fraction correct in the top 25% minus the fraction correct in the bottom 25%.

For norm-referenced exams, the difficulty of an individual item will typically vary from 0.4 to 0.7, so that there is an ability to spread out overall scores in a sample of students. Discrimination is also important for norm-referenced exams because the premise of a norm-referenced exam is that student scores are spread out to discriminate those who know the content from those who do not.

This discussion has focused so far on validity at the item level. Another key idea is reliability, which is a measure of whether a test would measure the same performance if a student were to retake it. Because students seldom take the same test twice, statistical estimates are used for reliability. These estimates often are akin to comparing how students perform on the first half of the test versus their performance on the second half of the test.

For teachers who face new expectations for assessment based on NCLB or other factors, these definitions can help them decide what must be considered in creating or choosing assessment materials. We will now look at some resources that are available and then consider teacher-developed assessment materials.

Available Resources

Plenty of potential resources are available to high school chemistry teachers in terms of assessment. Quantity is not a problem. Rather, the key challenge lies in determining how well these resources align with national standards. The Examinations Institute of the American Chemical Society produces nationally normed exams for both first-year high school chemistry and advanced high school chemistry. The first year exam is released every other year (in odd years), and, since 2003, has been designed with the goal of matching the content standards for chemistry present in the *National Science Education Standards* (NSES). Table 3 shows the distribution of items associated with the pertinent standards for the three most recently released exams.

Table 3. Alignment of Items From ACS First-Year High School Chemistry Exams with NSES Standards

NSES Standard	Standard Descriptions: "Students should develop …"	Number of questions (HS03)*	(HS05)	(HS07)
A1	… abilities to do scientific inquiry.	11	6	6
A2	… an understanding about scientific reasoning.	33	1	1
B1	… an understanding of the structure of atoms.	5	7	8
B2	… an understanding of the structure and properties of matter.	31	44	38
B3	… an understanding of chemical reactions.	28	15	23
B5	… an understanding of the conservation of energy and increase in disorder.	10	4	4
B6	… an understanding of interactions of energy and matter.	0	3	2
D1	… an understanding of energy in the earth system.	1	0	0
F5	… an understanding of natural and human-induced hazards.	1	0	0

***Note that HS03 refers to the 2003 First Year High School Chemistry Exam, HS05 to 2005, and HS07 to 2007.**

Table 3 would appear to suggest that there were more questions on the 2003 exam, than in 2005 or 2007. In 2003, however, questions that the test-writers believed matched to multiple standards were double or even triple counted in some cases. Beginning in 2005, the directions given to test-writers were changed so that the questions they submitted would map to only one national standard.

With regard to national assessment programs, perhaps the College Board and the Advanced Placement (AP) tests represent the most widely utilized resource for chemistry teachers. The College Board offers a Subject Test (formerly SAT II: Subject Test) in Chemistry. This test is designed for students that have taken a one-year college preparatory chemistry course. It is also important to note that the College Board is currently engaged in an extensive, NSF-funded process to design a large-scale change in all AP science courses, including chemistry.

Alternative forms of assessment have also been devised for the chemistry classroom. The use of portfolios, for example, has steadily developed over the past decade (Phelps, 1997). At the high school level, the International Baccalaureate (IB) program uses a portfolio approach, particularly related to laboratory work (IBWeb, 2007).

Assessment and Professional Development

Few teachers have access to professional development opportunities related to assessment from within their content-based specialty. For example, few workshops in writing good test questions in chemistry are offered at national conferences related to chemical education.

The ACS Division of Chemical Education (DivCHED) has sponsored the Examinations Institute for many years and provides opportunities for assessment-based professional development. For example, workshops at summer conferences allow teachers to experience, in an abbreviated format, the process by which an ACS exam is written by a test committee. When an ACS Exam is constructed, the process involves several important steps that help teachers at all levels become more versed in matters related to the construction of test questions.

Construction of the first-year high school exam begins two years prior to the release of the exam by a committee of chemistry teachers. Each exam committee is composed of approximately 20 members—4 or 5 members have served on two or more previous committees, 5–10 members have served on one prior test committee, and there are typically 4 or 5 new committee members. Each committee member is assigned specific topics and provided writing style directions for the test questions they are asked to submit for possible use on the final exam. Each question is mapped to an NSES indicator by the initial author of the question. Each member submits at least five questions for each assigned topic area, so that the committee begins its work with about 450 potential questions.


39th IChO

The exam committee then meets a couple of days prior to that summer's chemical education conference (namely, the Biennial Conference on Chemical Education, which meets in even years or ChemED, which meets in the odd years), to select the 120 questions that will be field-tested during the following spring. Submitted questions are not identified by writer, meaning that each test writer has an opportunity to share their questions in an environment that is educational, but not judgmental. Each item is verified for accuracy and also goes through several editing cycles. The questions are then sent to the Examinations Institute to compile trial exams for use by students in the classrooms of volunteer teachers.

In the spring of the second year of each exam development cycle, the 120 questions are field-tested, and the Examinations Institute calculates the item statistics for each question. The committee members use these statistics to identify items that perform well (per the criteria discussed earlier in this chapter, such as difficulty and discrimination). During the summer chemical education conference of the second year, the committee selects the final 80 questions for the newest first-year high school exam.

Test committee members have consistently noted that serving on an ACS exam committee is one of their most rewarding and professionally enriching experiences. They find it not only educationally informative, but also personally rewarding. Committee members become great support colleagues for each other and continue to learn from each other throughout their careers.

Teacher–Designed Assessment

Although nationally normed tests provide important forms of assessment in the current educational climate, most assessment remains teacher designed. Several key concepts that tend to enhance the quality of test questions are worth noting here.

First, assessment related to standards must reflect the emphases of the standards. Within most systems, this requirement will mean questions that move away from easily measured, discrete knowledge, toward items that align with the expectations of the national (and state) standards. In particular, teachers have expressed a greater desire and need for inquiry-based or reasoning questions. This does not necessarily require open-ended questions or questions about specific laboratory experiments. For example, adding a laboratory context (as shown in Fig. 1*B*), will elicit greater reasoning skills within a multiple-choice item.

Figure 1. Examples of Items

What is the pH of a 0.00001 M HNO_3 solution?	
(A)	1
(B)	3
(C)	5
(D)	9

If an experiment requires a pH of 5, which solution could be used?	
(A)	0.5 M HNO_3
(B)	0.5 M CH_3COOH
(C)	0.00001 M HNO_3
(D)	0.00001 M CH_3COOH

Second, the objectives for a specific test should be formulated in terms of words that require student action. Rather than starting a question that begins with "What is …" terms such as demonstrate, describe, distinguish, or assess provide a richer environment for devising test questions. This difference is particularly important with open-ended questions. Fig. 2 shows how such wording changes can produce a more robust level of student response for similar chemical content.

Figure 2. Examples of Items

What is the precipitate when $AgNO_3$ and HCl are mixed?	**Describe how to distinguish between a colorless solution that contains silver ions and one that does not.**

Third, when writing items, every effort should be made to ensure that each item (1) be written concisely and with minimal complications; (2) avoid adding irrelevant material, recognizing that if the objective of the item is to assess student analysis of data, it is possible to have more data than needed, but this would not qualify as irrelevant information; (3) the problem being asked should be clear enough that the student is not confused about what is expected; (4) avoid long descriptions about what is expected; and (5) avoid phrasing the question in negative terms. Multiple-choice items add additional constraints, in particular, the structure of the responses becomes important. Distractors should all be plausible, at least to a student who doesn't have the requisite knowledge or skills to complete the item.

Consider the examples shown here in Fig. 3. Both items reflect the science standard that calls for an understanding of atomic structure, but Fig. 3*A*, asks a student to provide an answer about an electronic configuration that can be readily memorized without much understanding of actual atomic structure, while the item shown in Fig. 3*B* also utilizes the construct of electron configuration but is more likely to show deep understanding of the concept. For teacher-designed assessments, this change in emphasis provides important additional information about student knowledge relative to more traditional test questions.