

# ACS Exams As An Example Of Scholarship-based Assessment In A Discipline

Thomas Holme and Megan Grunert

*ACS Exams Institute, Department of Chemistry; 0213 Gilman Hall, Iowa State University, Ames, IA 50011*

**Abstract.** The Examinations Institute of the American Chemical Society has been producing norm-referenced exams for over 75 years and these efforts are reviewed here. The process by which exam-writing committees produce these exams involves both the setting of the content and trial testing of items prior to establishing the released exam. Beyond this process, the Institute has engaged in research based on data derived from various tests.

**Keywords:** assessment, cognitive complexity.

**PACS:**.01.10.Fv, 01.40.gf

## INTRODUCTION

A fundamental premise about education is that no matter what the level or the specific style employed in the classroom, the act of teaching for any instructor is personal. This premise then dictates an unavoidable tension in education because the broader institution (college or high school) also has an unmistakable corporate interest in the outcomes of teaching. In recent years, the institutional interests in quality education have increasingly been expressed via an emphasis on assessment. As the role of assessment is debated, examples of long-standing testing programs may provide interesting data for decision-making. In chemistry, the Examinations Institute of the American Chemical Society (ACS) has been producing nationally normed exams for over 75 years. In the past decade it has also undertaken an increase in research efforts about the nature of measurement evident in a standardized test for chemistry. Both the historical development of the Institute and a sample of research findings related to ACS Exams are presented here.

By 1984, the scope of the Exams Committee had grown significantly and DivCHED formalized the operational aspects of these efforts by transitioning from an "Exams Committee" to an "Exams Institute". Since this time, the Institute has had a Director, and is hosted by the university at which the Director teaches. The Director reports to a Board of Trustees, which in turn reports to the Executive Committee of DivCHED. The Institute funds all of its activities through sales of exams and study materials related to exams. There is no subsidy received from either the broader ACS or DivCHED. Currently, the Institute is hosted by Iowa State University.

## CURRENT SCHOLARSHIP OF ACS EXAMS

The Exams Institute produces exams for use at all high school and undergraduate levels where chemistry is customarily taught. While data return for the calculation of norms is voluntary, an interactive web site has increased the participation of exam users significantly relative to previous decades. There is, therefore, a relatively large pool of student performance data upon which to carry out analysis of the psychometrics related to exams in chemistry. The largest data sets are typically in General Chemistry, due to its relatively large enrollment, so all studies reported on here are derived from tests used in this course.

### Exam Development And Test Content

One important aspect of ACS Exams is that all tests are "grass roots". In other words, there is no sense that the Institute is an authority that dictates what must be tested. Rather, the members of the exam writing committee begin the process by discussing the current state of content coverage in the course of interest and devising a template of how many items for each content topic should be written. This step provides the main warrant for content validity for the exam as a whole when it is complete. This method for determining exam coverage is also a long-standing tradition of ACS Exams. Indeed, Ted Ashford noted in the first decade of the program that, "It was often pointed out that when teachers, a group of leaders in their field, get together to argue what is important to test they are really arguing what is important to teach."

With a putative content coverage in hand, individual members of the committee write and submit

items for consideration. Typically about four times the number of items as will be needed are initially constructed. A second meeting of the committee is used to edit items and choose roughly twice as many as will be needed for the released exam. Two trial tests are created and used in classrooms of volunteers around the US. Item statistics are generated from these student performances and the committee meets a third time to select the items that will be on the final version of the exam, using the item statistics for guidance.

The primary purpose of the trial test phase of the process is to assure high quality items that have desirable psychometric properties. Nonetheless, there are observations that can be made as a result of this process. Consider, for example, a recently developed exam for Instrumental Analysis (typically a senior level course that focuses on how instruments and experiments using them are designed to provide high quality data and fidelity of interpretation of that data). The committee included several items on the nature of NMR (such as how signals are derived and manipulated in an NMR experiment) but all these items tested at or below guessing. Thus, the trial testing phase of ACS Exam development, with a national sample of student performances, strongly suggests that instructors are choosing to drop content coverage of NMR in the Instrumental Analysis course. This is not the only example of content on trial tests showing poor student performance. Usually, because the exam is norm-referenced, the committee does not include an item that does little to discriminate between high and low proficiency students. On occasion, however, a committee will decide that a specific content objective is so important that it will include an item with performance near guessing.

### **Item Construct Versus Item Content**

It is critical to recognize that performance on a given exam item is an example of a student carrying out a cognitive task. A useful model to consider this was developed within chemistry education by Alex Johnstone and his colleagues [1,2]. The model is referred to as the Information Processing Model, and it is derived from long-standing cognitive models of working memory [3] and cognitive load [4]. The critical feature for this model as it relates to test item performance is that as the complexity of the cognitive task required to answer an item goes beyond some threshold set by the working memory capacity of the student, the item becomes a measure of the working memory capacity rather than the student understanding of the course content.

With this model in mind, the Institute has established a way to devise reliable expert-ratings of

item complexity [5], and a methodology for obtaining the student perception of complexity for exam items [6]. Studies using these tools find that the combination of expert-rated complexity and student-rated complexity explains more than half of the variability of student performance on a set of items in a low-stakes student practice exam.

This result suggests that attention to complexity when writing exam items can be quite important. Working with several groups of faculty volunteers has provided preliminary evidence that deciding the complexity of a quantitative cognitive task tends to be easier for experts than similar conceptual questions. As a result, the Institute carried out an analysis of all the conceptual items present in a series of general chemistry exams to determine whether or not the construct of the item itself plays a significant role in the overall complexity.

The study analyzed 120 test items that explicitly targeted conceptual understanding in some way. Both the item stem and item answers were broken down into a total of 30 categories. Scoring of an item within any such category is dichotomous; if that feature is present the item has a 1, if not a 0. Thus, an item that includes a stem with a sentence and an illustration of the particulate nature of matter (PNOM) along with a key to describe that figure would have three categories present.

At this point, this study has focused on items that are part of released exams — items that have survived the trial test portion of the exam development and therefore have shown reasonable psychometric properties at the outset. Perhaps not coincidentally, a key conclusion of this study is that exam items on released exams are not dictated by the elements of the construct itself. In other words, the majority of the variability in student performance is related to the content the items are testing rather than their constructs, a certainly positive result.

There are some hints of possible effects that warrant future investigation, however. One exam had several items that used the construct where students essentially choose “true or false” for two related concepts. For example, a student could be asked to identify if a process increases or decreases both enthalpy and entropy. Such constructs appear to lead to lower student performance, and because they inherently include an interaction component [7], they are often more complex items. Another conceptual construct used often in chemistry, the PNOM illustration, also is found to occasionally to lead to items with statistically significant lower performance when the illustration appears in the stem. This result suggests that PNOM illustrations may have inherent complexity that contributes to that lower performance.

## Item And Answer Order Effects

Another place where the existence of a large data pool of student performances on exam items allows for research is the role of item order or answer order in student performance. This study is predicated on the fact that for large enrollment classes, particularly in general chemistry, the Institute produces two forms of most exams to assist instructors in limiting cheating opportunities for students. A fairly comprehensive report of work in this area is appearing elsewhere soon [8], but some major observations can be noted here.

The first noteworthy observation is that item and answer order effects are fairly common. Exams that have been analyzed typically have about 20% of their items that show significant effects related to item position or answer positions. Fortunately, from the perspective of an overall norm-referenced exam, the differences usually cancel out over an entire exam.

It is also possible to identify trends that are often present when item or answer order effects are observed. First, a common form of item order effect arises when an item that normally has a large fraction of correct answers follows several very challenging items in a row. Not all item order effects observed can be explained by this trend, but it is the single most commonly observed pattern. Second, priming, both positive and negative, can affect performance on a given item. In positive priming, a recently encountered item cues students to access content knowledge relevant to the item at hand. Negative priming can occur when an earlier item cues a common misconception, or induces the student to use non-analytical strategies like the recognition heuristic [9] on a given item. Finally, the most important observation relating to answer order effects appears to be that for conceptual items, if the most common misconception appears before the correct answer, students may be more inclined to select that incorrect response.

## Testing and program assessment

Ultimately, a key driver of faculty behavior related to assessment is the increase of demands for the measurement of learning outcomes. Thus, while norm-referenced exams provide an externally calibrated reference about the nature of student learning, they don't necessarily afford ready comparison to desired learning outcomes. In response to this emerging need, the Institute has undertaken a process to establish a map that organizes chemistry content in terms of anchoring concepts, or "big ideas". Once this map is established, it will be possible to align items from ACS Exams, and departments interested in program

assessment will have a tool for measuring student learning within the overall norm-referenced test suite of the Institute.

### *Organizing To Build A Content Map For Chemistry*

The creation of a content map of an entire undergraduate curriculum is inherently daunting. Moreover, the Exams Institute has from its inception been viewed as an instrument to provide a service (testing) that responds to classroom trends rather than an authority that leads content coverage. The methodology utilized to build a content map has drawn on extensive input from groups of volunteers.

The task accomplished via this process is an emerging map of content at four levels. Level 1 consists of 10 anchoring concepts. Borrowing language from Backwards Design [10], Level 2 is referred to as "enduring understandings." This is a grain size smaller than big ideas, but large enough to span chemistry in any sub-discipline that a student encounters as an undergraduate. In working with many instructors from various disciplines, it became clear that different courses place substantially different emphases on the various components of the subject. Thus, the third level of the map allows the individual sub-disciplines within chemistry to articulate how the enduring understandings are approached in the specific undergraduate courses they teach. Finally, at the finest grain size, the typical specific course content details one finds in an undergraduate course are organized within the level 3 articulations.

**TABLE 1.** Timeline for content map development

Date	Sub Discipline	Activity
Mar 2008	Any	Level 1 & 2
July 2008	General	Level 2 & 3
Aug. 2008	Organic	Level 2 & 3
March 2009	General	Level 3 & 4
Aug. 2009	Organic	Level 3
March 2010	General	Item Alignment
March 2010	Physical	Level 3
July 2010	Organic	Level 3 & Complexity
Dec. 2010	Organic	Complexity
	Analytical,	
March 2011	Biochemistry, Physical	Level 2 & 3

The process of building this map has been time consuming, and the map itself is far from complete. Table 1 provides a timeline of the various events that have been undertaken thus far. The table elaborates what level within the map was addressed by faculty from which sub-discipline in chemistry. In general, the initial efforts were undertaken at the early levels of the undergraduate curriculum. Finally, it is also worth noting that in order to conduct meaningful alignment

of test items at a later stage, a process to have chemists rate complexity of items has also been conducted and is included in this timeline.

### *Description Of The Emerging Content Map For Chemistry*

While the process to build the content map has begun in most sub-disciplines of chemistry, it has neared a steady state only for General Chemistry. Even within this area, the map must be dynamic and is regularly updated. Specifically, because of the way in which the discipline content is divided, some anchoring concepts receive significantly greater attention in some sub-disciplines than others. When a new sub-discipline takes on their initial attempt to articulate level-3 for the enduring understandings, new statements are sometimes added, and others are edited. This process ultimately improves the enduring understandings, but effectively requires a "retro-fit" for sub-disciplines that are further along in the process.

For example, Anchoring Concept #9 refers to the experimental basis of chemical knowledge. Many of the nuances of experimental methodologies utilized in chemistry are taught in Analytical Chemistry, so the work conducted in March 2011 improved the enduring understanding level of description within this anchoring concept and both General Chemistry and Organic Chemistry must be updated within this big idea.

The Institute is preparing to release the first installment of the map for General Chemistry near the beginning of the Fall 2011 semester. The entire map is well beyond the scope of this summary, as implied by the numbers provided in Table 2 for the extent of information at each level.

**TABLE 2.** Approximate structure of emerging content map.

Level	Title	Number
I	Anchoring Concept	10
II	Enduring Understanding	~60
III	Sub-disciplinary Articulation	~550
IV	Course Content Details	~1200

### *Facilitating Program Assessment*

A crucial aspect of the establishment of this content map lies in how Chemistry Departments will be able to use it for program assessment. This need is ultimately what dictated the structure that has emerged. While the current ACS Exams are all essentially developed at the third and fourth levels of the map, all sub-disciplines share the top two levels. Thus, a department that uses ACS Exams throughout the undergraduate curriculum

will be able to map how their students' knowledge about key aspects of chemistry changes over time.

It is also apparent that this organization leads to several other observations. First, as chemistry instructors work on the map, organizing along anchoring concepts often causes them to view their own courses differently. Second, ACS Exams are typically 60 or 70 items in length. Even with students taking 6-10 exams over an undergraduate career, there will be areas that are not fully assessed. Of course, it may be possible for the Institute to address holes in test coverage in the future, but the intent of the map, by design, is for it to be rather exhaustive. It is not designed to be a description of learning objectives. Rather, it is meant to be able to incorporate such objectives that departments choose to elaborate. The intent is to make it possible for a department to identify its own priorities for learning objectives and then see how using ACS Exams will help them assess their success at having students achieve those objectives.

## **ACKNOWLEDGMENTS**

Several aspects of the work summarized here were supported by the National Science Foundation under grants (EHR-0817409 and EHR-0943783). The number of volunteers that make the work of the Exams Institute possible is staggering, and the efforts of these dedicated educators are gratefully acknowledged.

## **REFERENCES**

1. Johnstone, A. H., *Chem. Educ. Res. Pract.* **7**, 49-63 (2006).
2. El-Banna, H., and Johnstone, A. H., *Educ. in Chem.* **23**, 80-84 (1986).
3. Baddeley, A. D., *Essentials of Human Memory*, London: Psychology Press, 1999.
4. van Merriënboer, J. J. G., and Sweller, J., *Educ. Psych. Rev.* **17**, 147-177 (2005).
5. Knaus, K., Murphy, K., Blecking, A. and Holme, T., *J. Chem. Educ.* **88**, 554-560 (2011).  
student perceptions of item complexity as well. An
6. Knaus, K., Murphy, K. and Holme, T., *J. Chem. Educ.* **87**, 827-832 (2009).
7. Ayers, P., *Learn. Instr.* **16**, 389-5400 (2006).
8. Schroeder, J., Murphy, K., and Holme, T., *J. Chem. Educ.* submitted (2011).
9. Goldstein, G. G., and Gigerenzer, G., *Psych. Rev.* **109**, 75-90 (2002).
10. Wiggins, G. P. and McTighe, J., *Understanding by Design (2<sup>nd</sup> Ed)*, Alexandria, VA: Association for Supervision and Curriculum Development, 2005, pp. 13-34.