Journal of Learning Design

IMPROVING INSTRUCTIONAL DESIGN WITH BETTER ANALYSIS OF ASSESSMENT DATA

Kristen L. Murphy
Department of Chemistry and Biochemistry
University of Wisconsin, Milwaukee, WI, USA
kmurphy@uwm.edu

Thomas A. Holme
Department of Chemistry
Iowa State University, Ames, IA, USA
taholme@iastate.edu

Abstract

As more instructors articulate learning objectives for their students within one course, or academic staff collaborate to articulate learning outcomes for programs, a robust means to assess student performance within these becomes increasingly important. The Examinations Institute of the American Chemical Society (ACS), Division of Chemical Education, has recently published content maps that utilise a structure of subdiscipline-independent fundamental concepts narrowing down to content details that are specific to subdisciplines. This structure has then been utilised to align items and can be used to assess student performance throughout a program. Learning objectives that are designed for a course can then be aligned to the framework and used to gauge student learning within a course or across a program. One key to making well-informed instructional decisions is to obtain as much valid information from such assessment work as possible. This paper describes the combination of using a rubric for assigning complexity with student performance to gauge achieving learning objectives that are aligned to fundamental concepts in the content maps in general chemistry and organic chemistry. It can be argued that information in these forms can provide useful guidance for designing improved instruction.

Keywords

Testing and assessment, general chemistry, organic chemistry, chemical education research

Introduction

Classroom instruction, although unique to the instructor, has many common or unifying features. These include aspects of the course itself from the use of assessment to the order and depth of content. Other features can include the use of learning objectives to communicate to the students what they will expect to learn in the course and how they will be assessed. These assessments can be used by the course instructor to assign the level of content knowledge that individual students exhibit as well as providing information about the class as a whole. Such class-

wide assessment data can be used to inform future changes in instructional design, such as changing teaching methodologies or altering the order or depth of content. Additionally, because many instructors may be facing increased pressure to provide assessment data for programmatic assessment, these data can also be used to examine the development of students' knowledge in the domain over the course of the program. In order to connect assessment data between tests within one course or between courses within one program, a method is required to align both the content and the difficulty of the individual assessment items in order to provide information collectively about what students know.

Literature Review

There are a number of concerns associated with educational measurement and testing and a wide array of papers (Aubrecht & Aubrecht, 1983; Barbera & VandenPlas, 2011; Bates & Galloway, 2010; Englehardt, 2009; Hattie & Bond, 1999) and textbooks (Crocker & Algina, 1986; Haladyna, 1994; Kline, 2005; Nitko, 1983) have appeared that are related to this subject. At the most fundamental level, tests and other educational measures must be both *valid*, i.e. they measure what is intended to be measured, and *reliable*, i.e. that repeated measures would yield the same result if they were to be taken. Beyond this minimal level of quality, however, there are any number of demands educators might make of their assessment efforts. Within this array of work, the concept of item statistics that identify relative student performance on different tasks provides an important emphasis (Nitko, 1983). One key development of more modern test theories (Wilson, 2008) such as Item Response Theory (IRT) lies in a more robust, probabilistic treatment of student performance. Ultimately, all methods of item analysis seek to identify the chance that students will answer a given item correctly.

One key factor that influences the potential error associated with measurement lies in the complexity of the cognitive tasks that students are expected to perform. Any domain with a relatively complex cognitive structure is likely to impose challenges on the development of meaningful assessments (Charalambous, Kyriadkides, & Philippou, 2012). Within chemistry, Johnstone's (2006) Information Processing Model noted the centrality of task complexity in the learning and assessment of content. Bernholt and Parchmann (2011) specifically noted the role of complexity within the cognitive development of students and how this interplay can lead to missed opportunities for learning. In light of the importance of assessing the role of complexity, rubrics have been proposed (Knaus, Murphy, Blecking, & Holme, 2011) and refined (Raker, Trate, Holme, & Murphy, 2013) to harness expert assessment of test item complexity.

In addition to the technical components of measurement of student learning, the ultimate objective of measurement relative to curricular demands has also been an area of active effort in the recent past. Approaches to identifying the relative importance of content and skills within a discipline have varied among different countries. For example, the establishment of Threshold Learning Outcomes

(TLOs) in Australia (Hay, 2012) and the response of chemistry instruction to the TLOs (Schultz, Mitchell Crow, & O'Brien, 2013) has included significant countrywide coordination. By contrast, within the United States the adoption of learning outcomes based approaches, particularly in chemistry, has more often been driven by forces within individual universities, or at least by individual university responses to a rather amorphous national trend (Towns, 2010). The trends within chemistry in the US, however, benefit from an organised effort in assessment through the American Chemical Society (ACS) and its Examinations Institute. Thus, a number of chemistry educators have participated in the development of an *anchoring concept content map* (ACCM) (Holme & Murphy, 2012; Murphy, Holme, Zenisky, Caruthers, & Knaus, 2012; Raker, Holme, & Murphy, 2013; Zenisky & Murphy, 2013) without any policy level demands from governmental entities. The development and structure of the ACCM are similar to other efforts to design curriculum with assessment in mind from the outset (Holme, 2014; Huff, Steinberg, & Matts, 2010). Within the ACCM, a total of four levels of increasing detail describing the domain of chemistry for a standard undergraduate curriculum in the US are specified. An example of four levels in general chemistry and organic chemistry is provided in Table 1. As shown, Levels 1 (anchoring concept) and 2 (enduring understanding) are common to all subdisciplines and Levels 3 (subdisciplinary articulation) and 4 (content details) are subdiscipline specific and thus given separately for both general and organic chemistry.

Table 1

Example of the four levels in the ACCM for general and organic chemistry

Level 1 Anchoring Concept (AC)	**Level 2** Enduring Understanding	**Level 3** Subdisciplinary Articulation	**Level 4** Content Details
Atoms: Matter consists of atoms that have internal structures that dictate their chemical and physical behaviour.	Electrons play the key role for atoms to bond with other atoms	*General:* For a neutral atom there are as many electrons as there are protons, but the electrons can be categorised as core (inner) and valence (outer) electrons.	*General:* Valence electrons, which determine the properties of elements, are correlated with the groups in the periodic table.
		Organic: Electrons play a role in understanding the relative stability of resonance structures.	*Organic:* Stabilisation of anions helps to explain pK_a values and relative acidities of protons.

Level 1 is the broadest of the four levels described in Table 1 and labeled as the *Anchoring Concept* (AC). These ten anchoring concepts are listed in Table 2.

Table 2

Ten anchoring concepts with descriptions from the ACCM

Anchoring Concept (AC)		Description
AC1	Atoms	Matter consists of atoms that have internal structures that dictate their chemical and physical behavior.
AC2	Bonding	Atoms interact via electrostatic forces to form chemical bonds.
AC3	Structure/Function	Chemical compounds have geometric structures that influence their chemical and physical behaviors.
AC4	Inter-molecular Interactions	Intermolecular forces, electrostatic forces between molecules, dictate the physical behavior of matter.
AC5	Chemical Reactions	Matter changes, forming products that have new chemical and physical properties.
AC6	Energy and Thermo-dynamics	Energy is the key currency of chemical reactions in molecular scale systems as well as macroscopic systems.
AC7	Kinetics	Chemical changes have a time scale over which they occur.
AC8	Equilibrium	All chemical changes are, in principle, reversible and chemical processes often reach a state of dynamic equilibrium.
AC9	Experiments, Measurement and Data	Chemistry is generally advanced via empirical observation.
AC10	Visualisation	Chemistry constructs meaning interchangeably at the particulate and macroscopic levels.

Level 2, *Enduring Understandings*, still spans content that crosses all areas of chemistry but the statements are more detailed than the first level. The third and fourth levels, *Subdisciplinary Articulations* and *Content Details* respectively, are specific to a subdiscipline, often a particular course within the full undergraduate curriculum, and continue to narrow the content to a Level (4) appropriate for writing a single test item. Content maps for general (Holme & Murphy, 2012) and organic chemistry (Raker, Holme, & Murphy, 2013) have been published as well as the process by which these maps were developed (Murphy, Holme, Zenisky, Caruthers, & Knaus, 2012; Zenisky & Murphy, 2013).

This paper seeks to establish a means to analyse exam performance data that can provide comparisons between students, courses and institutions despite the differences that might arise due to variability in implementation strategies of learning outcomes at different institutions.

Methods

All ACS Exams are secure tests that are administered in a secure manner and delivered either via a paper and pencil exam or electronically. All exams (both printed and electronic) have instructions for administering the exam including specifying additional materials (such as a non-programmable calculator) that are allowed for use by the student and the time allowed to take the test. All exam items analysed as a part of this study are forced-response multiple choice with four response options of which only one is correct. Scoring is based on total number correct with no partial credit or penalties for incorrect responses.

Users of ACS Exams are encouraged to submit their students' scores for inclusion in the construction of the national norms. Users are also encouraged to submit individual student responses to all items for the construction of the item statistics for each released test. The item statistics reported here are based on voluntary submission of student performances, and participation levels vary by test. The number of performances, institutions contributing to the item statistics, total number of items and average performance by test are shown in Table 3. Item statistics that are routinely provided for ACS Exams include difficulty and discrimination. Difficulty is reported as the fraction of students who got the item correct and thus ranges from 0 (none of the students got the question correct) to 1 (all students got the question correct). Typically, difficulty values range from 0.30 to approximately 0.85 for ACS Exam items. Discrimination is reported as the fraction of higher performing students who got the question correct minus the fraction of lower performing students who got the question correct. The high and low groups are determined by the overall score on the test and the fraction of high vs. low students can range from 25% to 33%. Therefore, discrimination values can range from −1 (all lower performing and no higher performing students got the question correct) to 1 (all higher performing and no lower performing students got the question correct). Ideally, values of 0.25 or higher are expected for ACS Exam items.

Table 3

Number of institutions and performances by test type

Test		# institutions	# performances	# items	Average aggregate performance
General Chemistry	Full Year	15	580	70	37.6
Organic Chemistry	Full Year	31	1060	70	41.5
Organic Chemistry	First Term	18	1115	70	36.8

Note to Table 3. This test has a format of 64+6 items from one of two content areas for a total of 70, but there is an option for which both content areas are chosen (for a total of 64+12 or 76) based on course content.

Alignment and complexity analyses were conducted via focus groups with instructors or postdoctoral students who have taught the courses targeted by the tests. The types and locations of these focus groups are listed in Table 4. Participants were sought through general announcement of workshops at the Biennial Conference on Chemical Education, email or direct contact invitation to current or past exam development committee members with experience in the targeted subdisciplines. There were no common raters between the general chemistry and organic chemistry tasks. There were common raters between the organic chemistry first term and organic chemistry full year tasks.

Table 4

Location, type of focus group, and task/test organised by year

Year	Location	Task	Test
2011	241st National Meeting of the American Chemical Society, Anaheim, CA	Alignment and complexity	General Chemistry, Full Year
2012	22nd Biennial Conference on Chemical Education, Pennsylvania State University, PA	Alignment and complexity	Organic Chemistry, Full Year
2013	245th National Meeting of the American Chemical Society, New Orleans, LA	Alignment and complexity	Organic Chemistry, Full Year
2013	246th National Meeting of the American Chemical Society, Indianapolis, IN	Alignment and complexity	Organic Chemistry, First Term

The complexity analysis was conducted using two published complexity rubrics (Knaus, Murphy, Blecking, & Holme, 2011; Raker, Trate, Holme, & Murphy, 2013). Participants in the focus groups were provided with background information on what complexity is and how a rubric might be designed to capture this objectively from an expert perspective. The participants were then instructed on the different components of the complexity rubric and how to use the rubric to arrive at a complexity value. The participants were provided with a worksheet with instructions and a grid to complete for each item (shown in Figure 1). The training concluded with collectively working through an example item (not from the targeted test for analysis) using the rubric. Participants worked alone to complete the ratings. Training took approximately 15 minutes and, due to time constraints, not all participants completed complexity ratings for all items on the exam they were analysing. The process of assigning complexity is described in detail elsewhere (Knaus, Murphy, Blecking, & Holme, 2011; Raker, Trate, Holme, & Murphy, 2013).

In general, the process involved a series of 10 steps for each item. For the first step (not shown in Figure 1), the participant reads the item. Following this (Step 2), the participant had the option of identifying the key factors in the item, or what the item generally tested (often a broad content area such as stoichiometry or stereochemistry). The participant then determined the elements involved in correctly solving the item (Step 3), which can include what students must know or recognise to do and the difficulty of each of these elements (Step 4). These were summed by difficulty (Step 5), given a rating from the rubric (Knaus, Murphy, Blecking, & Holme, 2011; Raker, Trate, Holme, & Murphy, 2013) (Step 6), and summed (Step 7). The relationship between these elements was evaluated in solving the item correctly in Step 8. Finally, the role of the distractors was evaluated from how a student would solve the item in Step 9. The complexity value of the item was then determined from the sum of the ratings, the interactivity and the role of distractors (Step 10). There is no minimum or maximum expected number of elements or a corresponding rating. The rubric was intentionally designed to be flexible for the individual experience of ratings (e.g. where one rater may identify two difficult elements and another may identify 8 easy elements).

2. Key Factor(s) *optional*:					
3. Elements	**4. Difficulty of each element**	**5. Total number of each category**	**6. Rating of elements**	**7. Sum of ratings**	**8. Interactivity rating**
		# Easy: # Medium: # Hard:	Easy = ____ Med = ____ Hard = ____		Easy = 1 Med = 2 Diff = 3 *(circle your value)*
9. Role of the distractors rating: Select = 0 Eliminate = 1 Evaluate = 2 *(circle your value)*					
10. Final Complexity rating: [Sum, #7] + [Interactivity, #8] +[Distractor, #9] = _____ + _____ + _____ = _____					

Figure 1. Complexity analysis grid (one used for each item for each participant)

The alignment analysis was conducted using the published content maps for general chemistry (Holme & Murphy, 2012) and organic chemistry (Raker, Holme, & Murphy, 2013). Participants in the focus groups were provided with background information on criterion referencing and the development of the content maps. The participants were also instructed on the general tenets of complexity and the components that are considered when assigning an item as "easy, medium or hard" in difficulty from an expert perspective. The participants were then provided with the content map and the general layout or format of the content map was discussed. Finally, the participants worked in small groups of 2 or 3 and assigned a content location to the exam items as well as a complexity value on a 1-3 scale (with 1 as "easy"). Participants were instructed that exam items could be placed into more than one content location. All items were able to be placed into the content map. Participants in the focus groups did not have access to student performance data so they did not know *a priori* what items students actually found "hard" or "easy." Training took approximately 15 minutes and due to time constraints of the workshop format, not all participants completed aligning all items on the exam they were analysing.

Alignment and complexity data was entered into a spreadsheet for analysis. Complexity ratings were assigned either using the rubric, through the alignment process or through a sorting process (into easy, medium and hard categories). Complexity ratings were analysed for inter-rater reliability using Cronbach's alpha. A single complexity value for each item was determined from the average of all raters. Alignment data were entered through all four levels of the content map for all items (where available) for each rater. No hierarchy in multiple ratings by one single group was requested and, therefore, no hierarchy in multiple ratings for any single item was assigned. The alignment values for content (i.e. what content was tested by the item) were assigned by using a majority rule so that an item was placed within the content area assigned by the largest number of raters. When agreement between raters was lacking, a secondary rater with expertise in

the domain was used to determine the rating(s). Reliability of the content assignment was based on matching through either two or four levels of the content map. Because the assignment values are nominal, a match is assigned as 1 and non-match as 0 for each level. Matching percentages are determined as averages of these.

Results

General Chemistry

Item statistics for 70 items on a standard full year general chemistry exam were based on 580 submitted student performances. The plot of discrimination vs. difficulty is shown in Figure 2. The typically expected value cutoffs for difficulty and discrimination are shown with the shaded box, with the "accepted" values in the shaded region. Because exams are developed using a trial-testing phase (Holme, 2003), the decisions about which items are selected for a released exam are based on item statistics of difficulty and discrimination from trial testing. However, it is also expected that some items may perform differently on released exams, and in addition, it is up to the discretion of the exam developing committee to include items for content coverage that may be outside of "accepted" values. Therefore, although the majority of the exam items are within what is nominally considered the accepted range, several items are not.

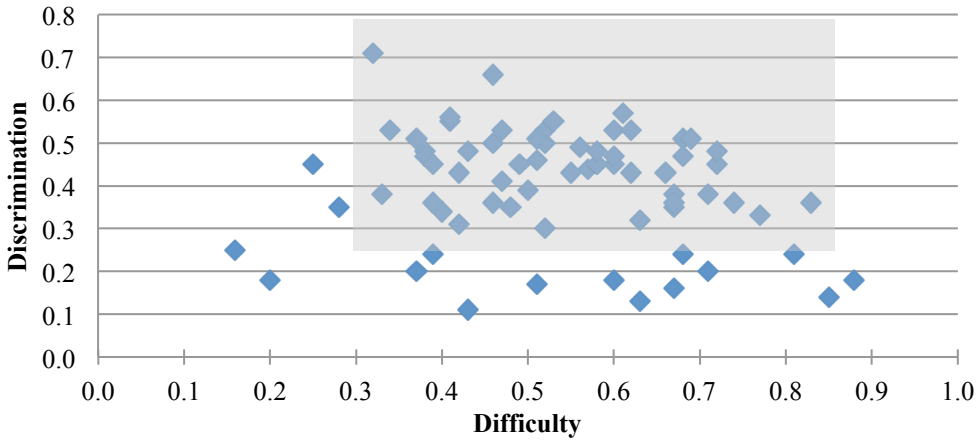

Figure 2. Discrimination vs. Difficulty by Item
General Chemistry Exam (Full year); n (items) = 70; n (performances) = 580

All 70 items from the released exam were examined for complexity and alignment. The complexity rating was assigned using the complexity rubric (Knaus, Murphy, Blecking, & Holme, 2011) with 5 raters. The descriptive statistics for these ratings are provided in Table 5. The key point of this table is to establish the boundaries of ratings for complexity using this process. The agreement between the five raters was analysed using Cronbach's alpha and determined to be 0.85, which is above the standard limit of 0.8. All five raters contributed ratings for all 70 items.

Table 5

Descriptive statistics for complexity ratings for general Chemistry, full year

Statistical measure	Rating
Mean	4.3
Median	4
Mode	4
Standard Deviation	1.3
High	11
Low	3

When considering the complexity values, unless the range of item complexity is small and skews towards low values, as the complexity value increases the performance or difficulty index should decrease. This is evident in Figure 3, where the average complexity values are plotted against the difficulty values. The Pearson-product moment correlation between the complexity values and difficulty was negative but was not significant, $r(68) = -0.21$, $p = 0.081$.

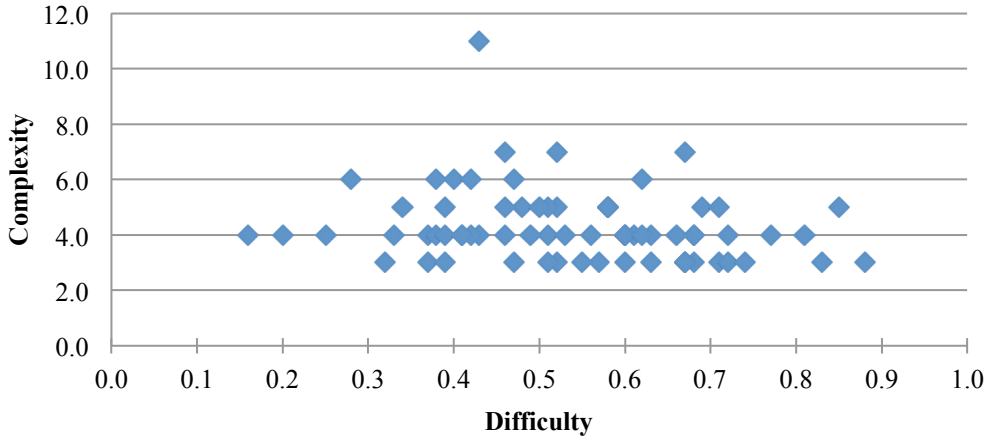

Figure 3. Complexity from rubric analysis vs. Difficulty by Item
General Chemistry Exam (Full year); n (items) = 70; n (performances)= 580

The process of assigning complexity using the rubric was valuable but time consuming. Because complexity assignment on a three-point scale was included in the alignment process, with an explanation of the fundamental tenets of complexity presented prior to the alignment process, the complexity assigned during the alignment process was also analysed. The correlation between the two assignments of complexity was positive and significant, $r(68) = 0.64$, $p < 0.001$. The correlation between the complexity on the three-point scale and difficulty was negative and significant, $r(68) = -0.28$, $p = 0.021$. Because collecting complexity during the alignment process was possible and the validity of the complexity assignments collected via this process was consistent with the complexity

assigned via the more time-consuming rubric analysis, the complexity values on the three-point scale were used in all further analyses.

The alignment of each of the items was conducted using the published content map for general chemistry (Holme & Murphy, 2012), without an assignment of complexity because this had been made using the rubric. As shown in Table 1, the first two levels of the content map are subdiscipline independent and the most broad. The agreement between the raters for this alignment was 97% through these first two levels. As the content map enters into the next two subdiscipline specific levels, the specificity of the content narrows and the agreement between the raters reduced to 83% through all four levels.

Performance on any test is comprised of the aggregate scoring of all exam items. This type of aggregate performance information may be valuable in knowing what fraction of the students are successfully completing the exam tasks, but the aggregation masks more nuanced comparisons. For example, the aggregate method may not be the best means to compare between learning objectives or broad content areas. In principle, performance in one specific content area or learning objective could be estimated by aggregate scoring of a subset of items. This assessment design is likely prone to error, however, because exam items are not equivalently difficult. Therefore, deducing that student performance lags in an area known to be more cognitively complex than others would be inaccurate. The objective, expert-based assignment of complexity, however, can be used to scale the performance data and account for inherent differences in content complexity. This process is demonstrated here for two content areas, general chemistry and organic chemistry.

When scaling performance, the complexity values were used as a multiplier on a 1-point scale, such that low complexity was scaled by 0.33; medium was scaled by 0.67 and high complexity was unscaled (essentially multiplied by 1). This choice for scaling factors was designed to retain difficulty values that scale from 0 to 1. Other choices for scaling would show the same trends, but this choice means that scaled aggregate difficulty values will inevitably be lower than unscaled difficulty, as is reflected in the data presented in the figures to follow. The scaled difficulty and difficulty are plotted for all 70 general chemistry items in Figure 4.

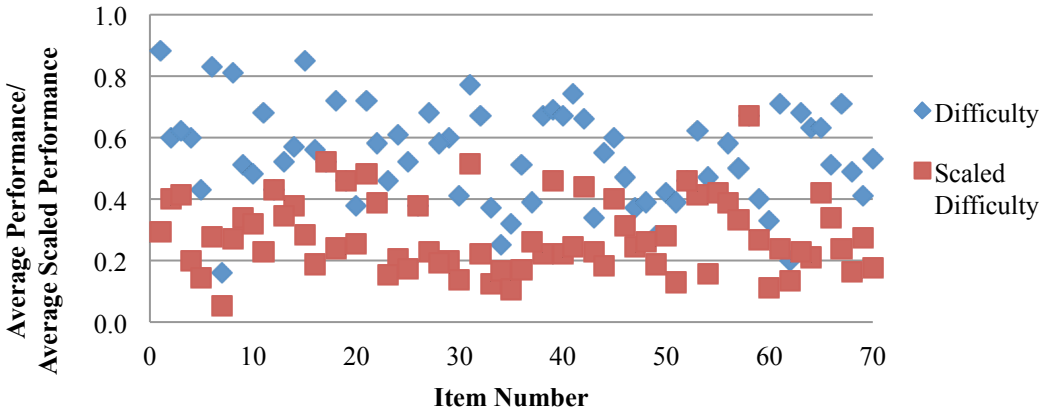

Figure 4. Difficulty and Scaled Difficulty by Item
General Chemistry Exam (Full year); n (items) = 70; n (performances) = 580

As expected, the scaled difficulty values are lower than the actual difficulty values. Scaled difficulty certainly presents a more abstracted meaning, but the results become valuable when comparing between content areas or learning objectives, thus adjusting for "harder" or "easier" content areas (at least in terms of items on the exam being analysed). The aggregate difficulty and aggregate scaled difficulty are shown by anchoring concept in Figure 5 (only 9 out of the 10 anchoring concepts are shown as there was only one item in the tenth anchoring concept). When considering the aggregate difficulty, the three anchoring concepts in which students performed the best were kinetics (AC7), experimental (AC9) and atoms (AC1). However, it is possible that the items within these concepts were easier than the items in other concepts. Considering the aggregate scaled difficulty, the three anchoring concepts in which the students performed the best were bonding (AC2), structure and function (AC3) and energy and thermodynamics (AC6). This analysis of performance in common topics in first-year chemistry courses could assist course design by emphasising in new ways what students know, categorised either by anchoring concept or learning objective, using both difficulty and scaled difficulty.

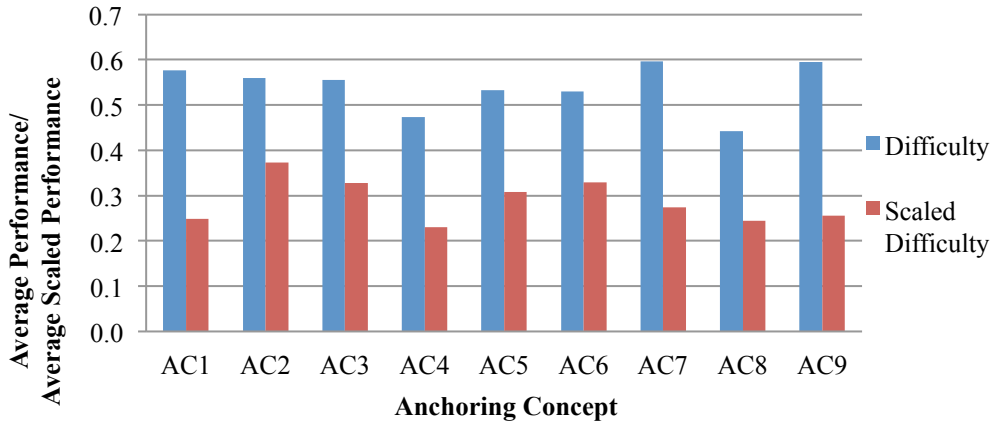

Figure 5. Difficulty and Scaled Difficulty by Anchoring Concept
General Chemistry Exam (Full year); n (items) = 70; n (performances) = 580

Organic Chemistry

Item statistics for 70 items from a standard full-year organic chemistry exam as well as 76 items from a standard first-term organic chemistry exam were based on 1060 submitted student performances for the full-year exam and 1115 for the first-term exam. The plot of discrimination vs. difficulty is shown in Figure 6 for the full-year exam and Figure 7 for first-term exam. The shaded box is included again as a reference for typical cutoffs, with several items outside of this region for both tests.

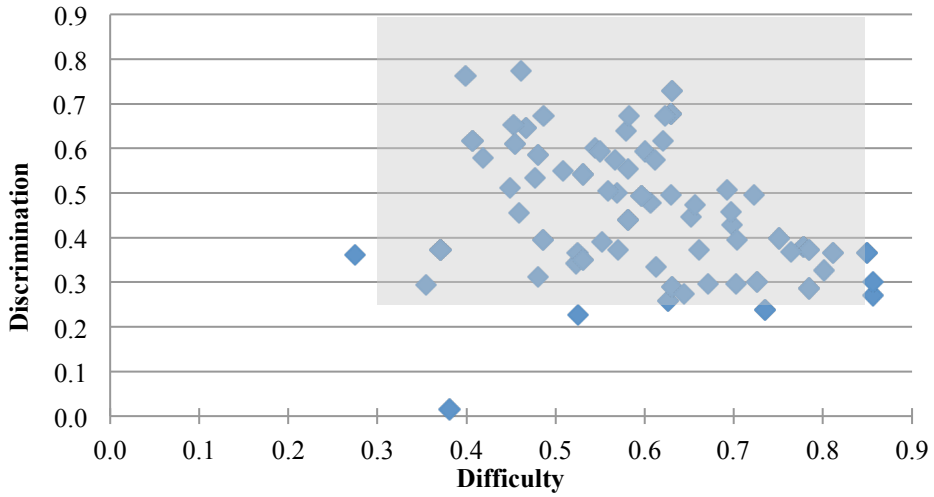

Figure 6: Discrimination vs. Difficulty by Item
Organic Chemistry; n (items) = 70; n (performances) = 1060

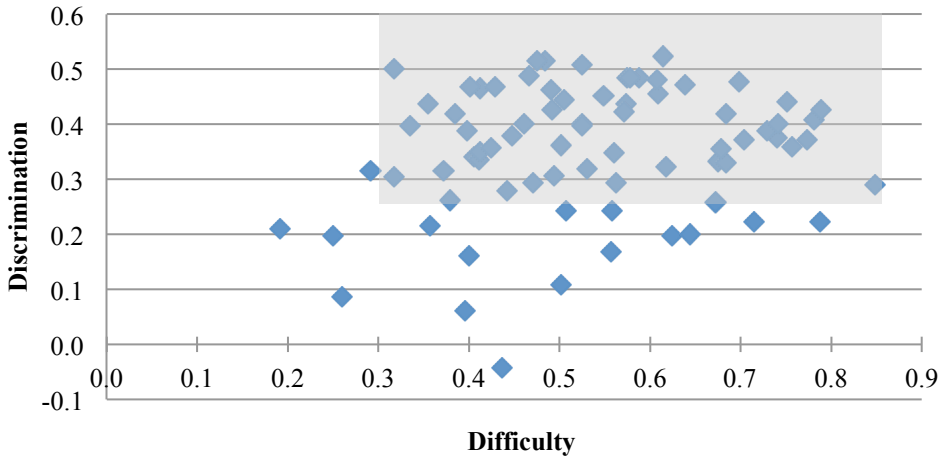

Figure 7. Discrimination vs. Difficulty by Item
Organic Chemistry (first term); n (items) = 76; n (performances) = 1115

The items from both released exams were examined for complexity and alignment. The complexity rating was assigned as described above within the alignment process using a three-point scale (Raker, Trate, Holme, & Murphy, 2013), with 6 raters for the full-year exam and 4 raters for the first-term exam.

The agreement between the 6 raters for the full-year exam was 0.74 with ratings available for 41 items. After removing one rater that had the lowest completion rate, the agreement rose to 0.87 with ratings available for 62 items. The agreement between the 4 raters for the first-term exam was 0.77 with ratings available for 53 items. After removing one rater that had the lowest completion rate, the agreement dropped to 0.69 with ratings available for 73 items.

Examining the relationship between the complexity ratings and difficulty values, both the full-year and first-term exams show similar trends to the full-year general chemistry exam. The Pearson-product moment correlation between the complexity values and difficulty was negative but was not significant for the full year, 68) = –0.17, $p = 0.16$. However, the correlation for the first term was both negative and significant, $r(74) = -0.30$, $p = 0.009$.

The alignment of each of the items was conducted using the published content map for organic chemistry (Raker, Holme, & Murphy, 2013) for both the full year and first-term exams. The agreement between the raters for the full-year alignment was 81% through all four levels and 97% through two levels. The agreement between the raters for the first-term alignment was 69% through all four levels and 77% through two levels. Once again, raters were not aware of item performance statistics when these tasks were completed.

The scaled difficulty and difficulty are plotted for all items in Figures 8 (full year) and 9 (first term). As seen previously, the scaled difficulty values are lower than the actual difficulty values and again adjust for "harder" or "easier" questions.

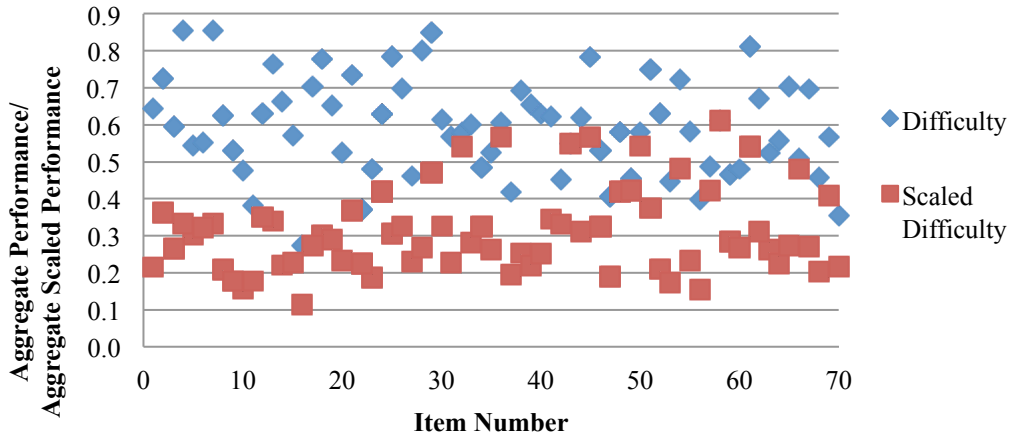

Figure 8. Difficulty and Scaled Difficulty by Item
Organic Chemistry; n (items) = 70; n (performances) = 1060

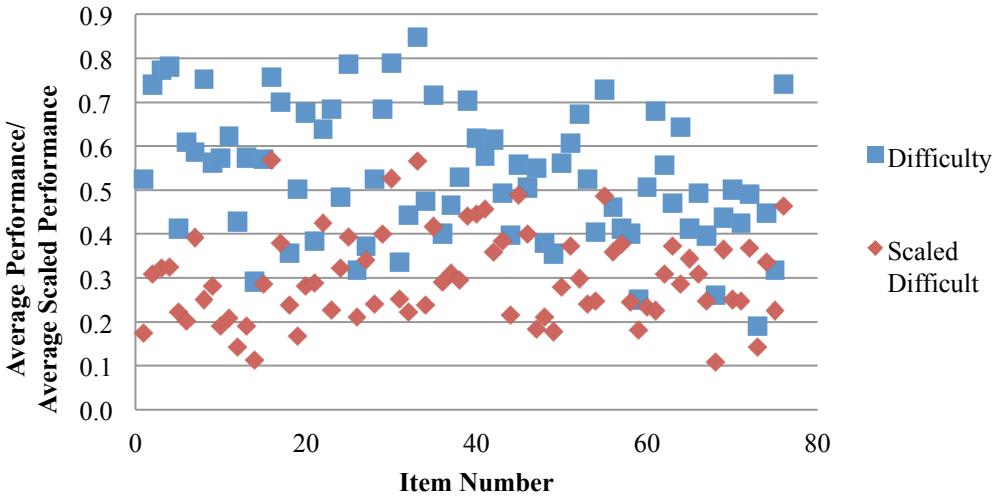

Figure 9. Difficulty and Scaled Difficulty by Item
Organic Chemistry (first term); n (items) = 76; n (performances) = 1115

The aggregate difficulty and aggregate scaled difficulty are shown by anchoring concept in Figures 10 (full year) and 11 (first term). When considering the aggregate difficulty, the two anchoring concepts in which the students performed the best for the full year were equilibrium (AC6) and kinetics (AC7). Considering the aggregate scaled difficulty, the anchoring concepts in which the students performed the best were still kinetics (AC7) now joined by atoms (AC1). When examining the first term results, the two anchoring concepts in which the students performed the best were also equilibrium (AC6) and kinetics (AC7) using aggregate difficulty, and these remained the highest when considering scaled difficulty.

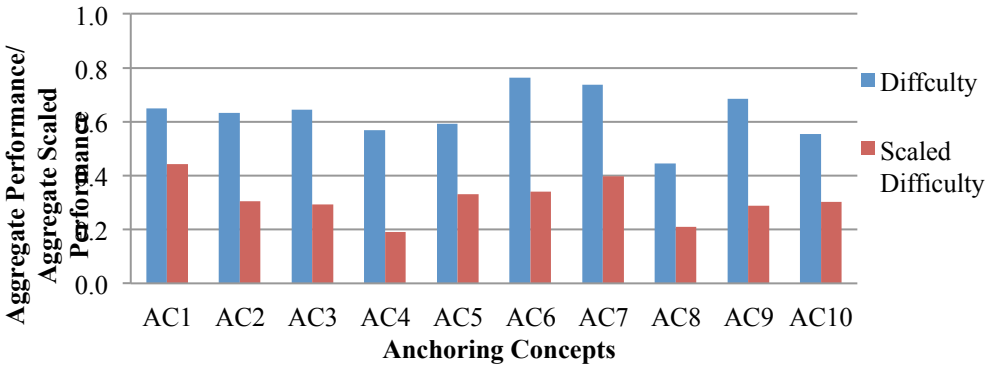

Figure 10. Difficulty and Scaled Difficulty by Anchoring Concept
Organic Chemistry; n (items) = 70; n (performances) = 1060

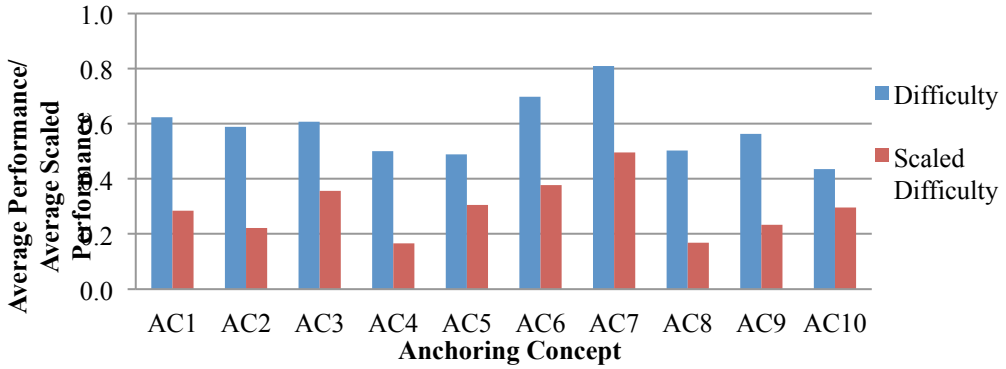

Figure 11. Difficulty and Scaled Difficulty by Anchoring Concept
Organic Chemistry (first term); n (items) = 76; n (performances) = 1115

However, unlike general chemistry, where the number of items in each anchoring concept was reasonably well spread, the nature of instruction in organic chemistry results in the observation that the highest number of items fall within anchoring concept 5 (reactions) with much smaller numbers in the other anchoring concepts. For this content area therefore, examining the enduring understandings (level 2) within the "reactions" anchoring concept may also be valuable when making decisions about what students know and in curriculum design. This analysis is shown in Figures 12 (full year) and 13 (first term).

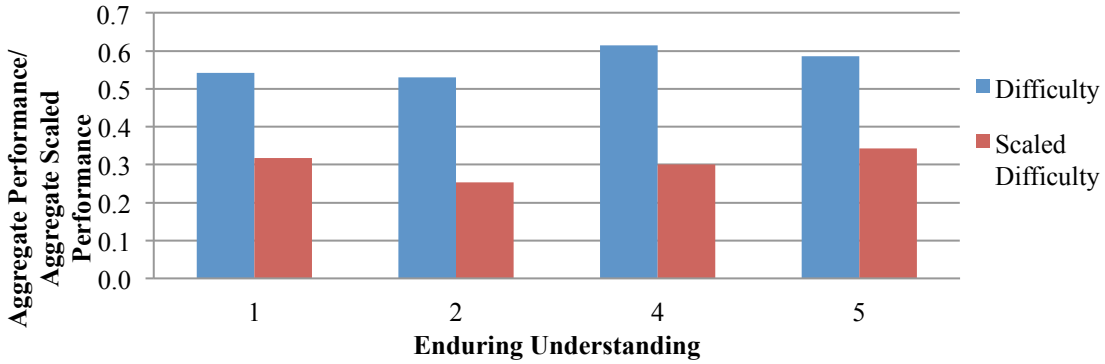

Figure 12. Difficulty and Scaled Difficulty by Reactions (AC5)
Organic Chemistry; n (items) = 70; n (performances) = 1060

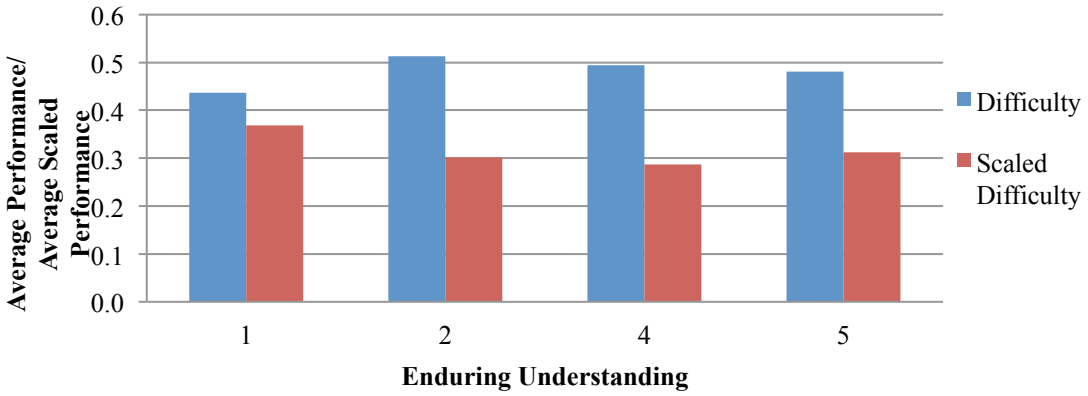

Figure 13. Difficulty and Scaled Difficulty by Reactions (AC5)
Organic Chemistry (first term); n (items) = 76; n (performances) = 1115

One of the enduring understandings (labeled #3 in the ACCM) has no items that mapped to it for either test, so it is omitted in these graphs. The content area in which the students performed the best for the full year by difficulty is 4 (which assesses different reaction types within organic chemistry) but changes to 5 (which describes how chemists control reaction outcomes in organic chemistry) when considering scaled difficulty. When considering the first-term exam, the content area in which the students performed the lowest when considering aggregate performance only was the enduring understanding #1 (which describes the ways reactions are represented), however, when considering scaled difficulty, this became the highest performing category. It is important to remember that the process by which ACS Exams are created (Holme, 2003) has a winnowing effect on items present in the exam. Thus, items which are included in the released version of an exam tend to have similar student performance data. Differences seen between enduring understandings here, therefore, may be less than might be seen for instructor written exams, where an extended editing process is generally not included in the test development process.

Conclusion and Implications for Practice

Utilising assessment data to make informed decisions about what students know is a critical component of classroom instructional design and extends into judgments about programs through programmatic assessment. Overall test scores can be useful in assigning grades for a course, but may not be the best measures when evaluating student performance on specific content areas or learning objectives. Additionally, the use of aggregate scoring on a collection of items within one content area or learning objective may be artificially skewed by the presence of more or less complex items. When considering a set of learning objectives within one course or examining student performance within one learning objective over a series of courses, having a more robust means of comparison can assist in making better decisions about what students know.

The combined use of complexity ratings and content placement of exam items allows for the calculation of a scaled difficulty value that can be aggregated for a

collection of items within a common content placement. This process illustrated here for multiple-choice items on ACS Exams can also be utilised for other exams or question types. Additionally, the content placement or alignment process can be utilised with other content maps for courses or programs. This process is reliant on a group of experts within the domain who can reliably rate the complexity of the item and identify the content that is being tested by the item. The availability of a widely vetted content map for chemistry, the ACCM in this case, is also very helpful, though any agreed upon content rubric could be used in this role. Using these ratings with student performance data then allows for standard comparisons to be made (such as the aggregate performance in a content area) as well as scaled comparisons between content areas.

Examining the performance on a national general chemistry exam in the US, the content areas in which the students performed the best were bonding, structure and function and energy and thermodynamics. The content areas in which the students performed the worst were in intermolecular forces and equilibrium. An instructor could use this information to re-evaluate the instruction in these content areas and consider how to best design instruction to increase student learning. Using scaled performance on subsequent exams would provide insight into the efficacy of these changes.

When considering organic chemistry in the US, it could be argued that greater value can be found not by examining the broad content areas of level 1 (or anchoring concepts) of the content map, but rather within the single anchoring concept of reactions, examining scaled performance on more specific content areas on the level 2 or enduring understandings. This distinction arises from the nature of content coverage in this sub-discipline, an artifact that has arisen historically within the teaching of chemistry. Importantly, from the perspective of instructional design, the example of organic chemistry illustrates that the concept of scaled difficulty measures can be applied at different granular levels of course content. Thus instructional design concerns can be informed with either relatively broad considerations, or for more specific content goals.

There are many potential implications for instruction when interpreting assessment results. The ability to gauge performance that includes a mechanism to adjust for the difficulty of the items is valuable in making decisions both about individual students and their knowledge and for an entire class. Integrating scaled performance into an advising portfolio for identifying strong and weak content areas for a student can assist when counselling the student on his or her preparedness for course work. Class-wide data can be used locally to inform instructional practices and more globally to inform programmatic areas of strength and weakness. Ultimately, a better understanding of what students know coincides with efforts to improve practice.

Acknowledgements
The authors gratefully acknowledge the work of all participants in the focus groups as well as the many users of ACS Exams who contribute student data for

generating national norms and item statistics. Partial funding for the development of the ACCM was provided by the National Science Foundation (DUE 071779, 0943783). Any opinions, findings, and conclusions communicated in this material are those of the authors and do not necessarily reflect the views of NSF.

References

Aubrecht, G. J., & Aubrecht, J. D. (1983). Constructing objective tests. *American Journal of Phys*ics, *3*, 613-620. doi: 10.1119/1.13186

Barbera, J., & VandenPlas, J. R. (2011). All assessment materials are not created equal: The myths about instrument development, validity and reliability. In D. M. Bunce (Ed.), *Investigating classroom myths through research on teaching and learning* (pp. 177-193). ACS Symposium Series.

Bates, S., & Galloway, R. (2010). Diagnostic tests for the physical sciences: A brief review. *New Directions*, *6*, 10-20.

Bernholt, S., & Parchmann, I. (2011). Assessing the complexity of students' knowledge in chemistry. *Chemical Education Research and Practice*, *12*, 167-173. doi: 10.1039/C1RP90021H

Charalambous, C. Y., Kyriakides, L., & Philippou, G. N. (2012). Developing a test for exploring student performance in a complex domain: Challenges faced, decisions made and implications drawn. *Studies in Education Eva*luation, *38*, 93-106. doi: 10.1016/j.stueduc.2012.08.001

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.

Engelhardt, P. V. (2009) An introduction to classical test theory as applied to conceptual multiple-choice tests. In C.R. Henderson & K.A. Harper (Ed.), *Getting Started in Physics Education Research* (pp. 1-40). American Association of Physics Teachers, College Park, MD. Retrieved from http://www.per-central.org/items/detail.cfm?ID=8807

Haladyna, T. M. (1994). *Developing and validating multiple-choice test items*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hattie, J., Jaeger, R. M., & Bond, L. (1999). Persistent methodological questions in educational testing. *Reviews of Research in Education*, *24*, 393-446. doi: 10.3102/0091732X024001393

Hay, I. (2012). Over the threshold – Setting minimum learning outcomes (benchmarks) for undergraduate geography majors in Australian Universities. *Journal of Geography in Higher Education, 36*, 481-498. doi: 10.1080/03098265.2012.691467

Holme, T. A. (2003). Assessment and quality control in chemistry education. *Journal of Chemical Education*, *80*, 594-597. doi: 10.1021/ed080p594

Holme, T. A. (2014). Comparing recent organizing templates for test content between ACS exams in general chemistry and AP chemistry. *Journal of Chemical Education*, Article ASAP, doi: 10.1021/ed400856r

Holme, T. A., & Murphy, K. L. (2012). The ACS Exams Institute undergraduate chemistry anchoring concepts content map I: General Chemistry. *Journal of Chemical Education*, *89*, 721-723. doi: 10.1021/ed300050q

Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education*, *23*, 310-324. doi: 10.1080/08957347.2010.510956

Johnstone, A.H. (2006). Chemical education research in Glasgow in perspective. *Chemical Education Research and Practice*, 7, 49-63. doi: 10.1039/B5RP90021B

Kline, T. J. B. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks, CA: Sage.

Knaus, K. J., Murphy, K. L., Blecking, A., & Holme. T. A. (2011). A valid and reliable instrument for cognitive complexity rating assignment of chemistry exam items. *Journal of Chemical Education*, 88, 554-560. doi: 10.1021/ed900070y

Murphy, K. L, Holme, T. A., Zenisky, A. L., Caruthers, H., & Knaus, K. J. (2012). Building the ACS Exams anchoring concept content map for undergraduate chemistry. *Journal of Chemical Education, 89,* 715-720. doi: 10.1021/ed300049w

Nitko, A. J. (1983). *Educational tests and measurement: An introduction*. San Diego, CA: Harcourt Brace Jovanovich.

Raker, J. R., Holme, T. A., & Murphy, K. L. (2013). The ACS Exams Institute undergraduate chemistry anchoring concepts content map II: Organic Chemistry. *Journal of Chemical Education*, *90*, 1443-1445. doi: 10.1021/ed400175w

Raker, J. R., Trate, J. M., Holme, T. A., & Murphy, K. L. (2013). Adaptation of an instrument for measuring the cognitive complexity of organic chemistry exam items. *Journal of Chemical Education*, 90, 1310-1315. doi: 10.1021/ed400373c

Schultz, M., Mitchell Crow, J., & O'Brien, G. (2013). Outcomes of the Chemistry Discipline Network mapping exercises: Are the Threshold Learning Outcomes met? *International Journal of Innovation in Science and Mathematics Education*, *21*, 81-91.

Towns, M. H. (2010). Developing learning objectives and assessment plans at a variety of institutions: Examples and case studies. *Journal of Chemical Education*, *87*, 91-96. doi: 10.1021/ed100066c

Wilson, M. (2008). Cognitive diagnosis using item response models. *Zeitschrift für Psychologie, 216*, 74-88. doi: 10.1027/0044-3409.216.2.74

Zenisky, A. L., & Murphy, K. L. (2013). Developing a content map and alignment process for the undergraduate curriculum in chemistry. In T. A. Holme, M. M. Cooper, & P. Varma-Nelson (Ed.), *Trajectories of chemistry education innovation and reform* (pp. 79-91). ACS Symposium Series.