2011

Assessing Conceptual and Algorithmic Knowledge in General Chemistry with ACS Exams

Thomas Holme
Iowa State University, taholme@iastate.edu

Kristen L. Murphy
University of Wisconsin - Milwaukee

Assessing Conceptual and Algorithmic Knowledge in General Chemistry with ACS Exams

Abstract
In 2005, the ACS Examinations Institute released an exam for first-term general chemistry in which items are intentionally paired with one conceptual and one traditional item. A second-term, paired-questions exam was released in 2007. This paper presents an empirical study of student performances on these two exams based on national samples of students who took the exams as part of a general chemistry course sequence. Psychometric data for student performances are presented in terms of classical difficulty and discrimination indexes, as well as item characteristic curves, as are more commonly used in item response theory. Having these data provided for all items on these two exams presents background information that researchers in chemistry education can use in studies in which these exams are part of the assessment paradigm that is used. Finally, because ACS Exams items may not be published, this manuscript presents examples of paired questions that can be used to describe the nature of the exam in subsequent publications that use these exams.

Keywords
first-year undergraduate/general, chemical education research, testing/assessment

Disciplines
Educational Assessment, Evaluation, and Research | Higher Education | Other Chemistry | Science and Mathematics Education

Comments
Reprinted (adapted) with permission from J. Chem. Educ., 2011, 88 (9), pp 1217–1222. Copyright 2011 American Chemical Society.

JOURNAL OF CHEMICAL EDUCATION

ARTICLE

pubs.acs.org/jchemeduc

Assessing Conceptual and Algorithmic Knowledge in General Chemistry with ACS Exams

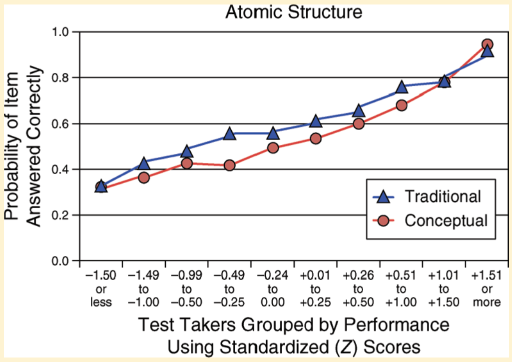
Thomas Holme*,[†] and Kristen Murphy[‡]

[†]Department of Chemistry, Iowa State University, Ames, Iowa 50011, United States

[‡]Department of Chemistry and Biochemistry, University of Wisconsin—Milwaukee, Milwaukee, Wisconsin 53211, United States

S Supporting Information

ABSTRACT: In 2005, the ACS Examinations Institute released an exam for first-term general chemistry in which items are intentionally paired with one conceptual and one traditional item. A second-term, paired-questions exam was released in 2007. This paper presents an empirical study of student performances on these two exams based on national samples of students who took the exams as part of a general chemistry course sequence. Psychometric data for student performances are presented in terms of classical difficulty and discrimination indexes, as well as item characteristic curves, as are more commonly used in item response theory. Having these data provided for all items on these two exams presents background information that researchers in chemistry education can use in studies in which these exams are part of the assessment paradigm that is used. Finally, because ACS Exams items may not be published, this manuscript presents examples of paired questions that can be used to describe the nature of the exam in subsequent publications that use these exams.



KEYWORDS: First-Year Undergraduate/General, Chemical Education Research, Testing/Assessment

FEATURE: Chemical Education Research

INTRODUCTION AND THEORY BASE

The prospect that students may learn quantitative problem-solving skills within chemistry while not understanding the conceptual basis for the content has been of interest for over 20 years. For example, Nurrenburn and Pickering found that conceptual understanding of stoichiometry lagged behind quantitative understanding.[1] Subsequently, several groups have confirmed this, as well as determining other features. Pickering established[2] that performance on conceptual questions in general chemistry was not a predictor of success in organic chemistry. Sawrey showed that difficulties with conceptual items were found for students with both high and low performance on traditional quantitative items.[3] Nakhleh and co-workers carried out a series of studies that further established the gap between conceptual understandings and algorithmic problem solving skills and sought pedagogies to mediate that gap.[4−8] A key component of all of these studies was the use of the paired-question format, in which student performance comparisons are drawn from multiple-choice item pairs that are designed to provide data about conceptual and algorithmic knowledge separately. The ACS Exams Institute provided a specific tool for this type of assessment in 1997[9] and updated the general chemistry paired questions exams in 2005 and 2007.[10,11]

The importance of conceptual misunderstandings that were uncovered via this methodology led to a wide range of studies that identified student misconceptions (or alternate conceptions) in a number of content domains of chemistry.[12−16] In addition to identifying the existence of misconceptions, it is arguable that these studies led to changes in the manner in which textbooks presented information about chemistry at the particulate level. Thus, over the past 20 years since the conceptual−algorithmic gap was first uncovered, there has been both further research and pedagogical responses.

This paper provides information about the 2005 Paired-Questions First-Semester General Chemistry Exam (GC05-PQF) and 2007 Paired-Questions Second-Semester General Chemistry Exam (GC07PQS) that were released by the ACS Exams Institute. These exams have been used nationally for several semesters, and the norm generation process of the Institute[17] has allowed for the consideration of item-level analysis of the exams over several thousand student performances for each exam. This paper provides national normative data, item statistics, and exemplars of paired questions that can be cited for research carried out using these secure exams.

While the information presented here is purely empirical, it remains important to consider theory bases that may be important for understanding the role of pair-questions in testing.

Published: July 08, 2011

C1. A student observes a temperature increase of ΔT_1 when she mixes 100 mL of a 1.0 M solution of NaOH with a 100 mL of a 1.0 M solution of HCl in a calorimeter. If she then mixes 100 mL of 1.0 M NaOH and 300 mL of 1.0 M HCl a temperature change ΔT_2 would be observed. The second temperature change, ΔT_2, is expected to be

(a) the same as the first.
(b) twice that of the first.
(c) three times that of the first.
(d) half that of the first.

T1. What can be predicted about the temperature change in the calorimeter in experiment 2?

Experiment	HBr	KOH	ΔT on Mixing
#1	100 mL 4.0 M	100 mL 2.0 M	6.2 °C
#2	100 mL 4.0 M	100 mL 3.0 M	?

(a) 3.1 °C
(b) 6.2 °C
(c) 9.3 °C
(d) Impossible to predict without further information.

Figure 1. Exemplar of paired questions. The content area for this pair is thermochemistry. C1 is classified as conceptual, while T1 is classified as traditional.

For example, it may be contended that meaningful learning, as defined by Novak and Gowin,[18] is more likely to be aligned with conceptual change as reflected in student performance on conceptual test items. In a similar way, Pushkin[19] argues that students who display conceptual understanding of chemical concepts can be categorized as having more advanced cognitive development.

Another way to view the apparent dichotomy between student performance in algorithmic versus conceptual questions is to consider dual-processing theories of cognitive processing. A recent summary of these accounts[20] refers to them as system 1 and system 2 processing. One example of this dual-processing model of cognition is to consider system 1 processing to be heuristic based and system 2 to be analytically based. In this sense, heuristic processes are fast processes that people may use without expending much cognitive effort and is therefore a frugal choice in test-taking or similar tasks. By contrast, system 2 processing involves analytical reasoning, which commonly engages more of the working memory and therefore is a more time-consuming method for achieving the cognitive task (answering the test item). Differences in student performance on the two types of exam questions implies that the "algorithmic" items are approached with heuristics more often, and conceptual items require more analytical thinking. This model does not preclude the possibility that students develop heuristics for solving conceptual problems as well.

Finally, it is important to recognize that there has been considerable effort to devise classification schemes that are more varied than just algorithmic versus conceptual. Zoller and co-workers have conducted numerous studies in this area.[21] More recently, Smith, Nakhleh, and Bretz[22] have proposed an extended system for classifying test items. In their system, the primary level of classification includes (i) definitions, (ii) algorithms, and (iii) conceptual understanding. A key development in this work lies in conceiving additional "secondary" levels of classification beyond these broad categories.

EXAM DEVELOPMENT AND STRUCTURE

The paired questions exams were prepared in a manner similar to the standard procedure for ACS Exams.[17] The key difference is that not all items in these exams were developed originally for them. A number of items were obtained from already released exams. Nonetheless, after all workable item pairs were gleaned from available items on released exams, it was determined that some content areas were not adequately covered and specific items or item pairs were developed for these exams. A trial-test phase of the development was undertaken so that student performances could provide statistical data to determine which pairs of items to include on the released exams. This process also led to the development of pairs of items that can be considered an exemplar of what the item pairs look like, while not having the security restrictions that forbid the publication of items from ACS Exams. Such a pair from the first-term exam is shown in Figure 1.

Note that this pair was not used in the exam because student performance on the conceptual item, C1, was very low (only 12%

Table 1. Content Coverage of Item Pairs

Topic	Number of Item Pairs
First-Term Exam	
Properties of matter	3 (6 items)
Atoms, elements, and compounds	1 (2 items)
Gases	3 (6 items)
Stoichiometry	5 (10 items)
Solutions and concentration	2 (4 items)
Atomic structure	1 (2 items)
Molecular structure	4 (8 items)
Thermochemistry	1 (2 items)
Second-Term Exam	
Equilibrium	3 (6 items)
Kinetics	2 (4 items)
Thermodynamics	3 (6 items)
Electrochemistry	3 (6 items)
Solutions	3 (6 items)
Acid−base chemistry	4 (8 items)
Nuclear chemistry	2 (4 items)

Table 2. Percentiles for Raw Scores on Paired-Questions Exams[a]

Score	40	39	38	37	36	35	34	33	32	31
Percentile, GC05PQF	100	100	99	98	97	95	93	90	87	84
Percentile, GC07PQS	100	100	99	99	98	97	96	94	93	91
Score	30	29	28	27	26	25	24	23	22	21
Percentile, GC05PQF	80	76	71	67	62	57	52	47	42	37
Percentile, GC07PQS	88	86	83	79	76	72	68	63	58	53
Score	20	19	18	17	16	15	14	13	12	11
Percentile, GC05PQF	32	27	23	18	15	12	9	7	5	3
Percentile, GC07PQS	48	42	36	30	25	20	15	11	8	5
Score	10	9	8	7	6	5	4	3	2	1
Percentile, GC05PQF	2	1	0	0	0	0	0	0	0	0
Percentile, GC07PQS	3	2	1	0	0	0	0	0	0	0

[a] These statistics pertain for the first-term exam (GC05PQF): Mean, 23.59; Standard Deviation, 6.93; Median, 23.1. The second-term exam (GC05PQF) statistics include: Mean, 21.11; Standard Deviation, 6.78; Median, 20.0.

Table 3. Classical Item Analysis for Paired Questions[a]

Topic	Item Pair	Item Difficulty, Conceptual Questions	Item Discrimination, Conceptual Questions	Item Difficulty, Traditional Questions	Item Discrimination, Traditional Questions
Properties of Matter	P1	0.604	0.476	0.712	0.414
	P2	0.870	0.222	0.795	0.419
	P3	0.651	0.489	0.738	0.417
Atoms	A1	0.600	0.559	0.696	0.493
Stoichiometry	ST1	0.519	0.524	0.764	0.454
	ST2	0.812	0.334	0.419	0.547
	ST3	0.851	0.334	0.696	0.455
	ST4	0.456	0.551	0.607	0.680
	ST5	0.460	0.463	0.473	0.715
Gases	G1	0.655	0.479	0.698	0.39
	G2	0.741	0.390	0.715	0.454
	G3	0.609	0.434	0.188	0.338
Solutions	SO1	0.613	0.482	0.636	0.244
	SO2	0.445	0.421	0.404	0.547
Atomic Structure	AS1	0.557	0.423	0.611	0.385
Molecular Structure	MS1	0.611	0.433	0.687	0.490
	MS2	0.611	0.441	0.447	0.562
	MS3	0.866	0.260	0.651	0.488
	MS4	0.804	0.399	0.504	0.325
Thermochemistry	T1	0.729	0.432	0.514	0.473
Equilibrium	EQ1	0.486	0.442	0.532	0.368
	EQ2	0.495	0.583	0.408	0.486
	EQ3	0.224	0.391	0.445	0.452
Kinetics	K1	0.489	0.433	0.480	0.372
	K2	0.665	0.251	0.622	0.477
Thermodynamics	T1	0.445	0.375	0.561	0.337
	T2	0.426	0.427	0.707	0.441
	T3	0.520	0.429	0.574	0.511
Electrochemistry	EC1	0.403	0.435	0.333	0.382
	EC2	0.513	0.441	0.613	0.475
	EC3	0.576	0.416	0.606	0.470
Solutions	S1	0.462	0.394	0.417	0.417
	S2	0.539	0.273	0.453	0.515
	S3	0.511	0.485	0.749	0.501
Acids—Bases	AB1	0.512	0.629	0.629	0.512
	AB2	0.766	0.412	0.422	0.330
	AB3	0.608	0.449	0.599	0.386
	AB4	0.535	0.381	0.289	0.411
Nuclear	N1	0.561	0.469	0.611	0.443
	N2	0.528	0.475	0.700	0.497

[a] Data for first-term topics (properties of matter to thermochemistry) are based on test performances from students at 12 schools, $N = 3073$. Data for second-term topics (equilibrium to nuclear chemistry) are based on test performances from students at 9 schools, $N = 3557$.

of students in the trial tests answered this item correctly). T1 was answered correctly by 54% of students in the trial test sample. It is also important to note that this conceptual item does not involve particulate level representations of chemical systems as were used in the original work.[1−8] Some conceptual items on the exams utilize diagrammatic representations, but the construction of conceptual items is broader than this construct.

Given this basic structure for item pairs, the overall released exams are constructed from 20 pairs in seven or eight content areas. Table 1 provides the overall structure of both exams in terms of content. Data returned for norm purposes, and reported here, are from students who were allowed 55 min (maximum) to complete either released exam. Instructors who purchase an exam are provided with the specific pairings. For all item pairs, the conceptual item occurs earlier in the exam than the traditional item, and there is always at least one item between the two in a specific pair.

DATA ANALYSIS

Overall norms for the exams in terms of percentiles for a particular raw score and other basic statistics are provided for both exams in Table 2. The Exams Institute also has routinely provided item statistics for normed exams based on classical test

Table 4. Bin Definitions and Sizes for ICC Construction

Z-score Increment	Students, N, for GC05PQF[a]	Students, N, for GC07PQS[b]
−1.50 or less	245	136
−1.49 to −1.00	226	495
−0.99 to −0.50	508	569
−0.49 to −0.25	300	413
−0.24 to 0.00	156	407
+0.01 to +0.25	301	164
+0.26 to +0.50	312	336
+0.51 to +1.00	434	377
+1.01 to +1.50	410	370
+1.50 or more	181	290

[a] For the first-term exam, total is $N = 3073$. [b] For the second-term exam, total is $N = 3557$.

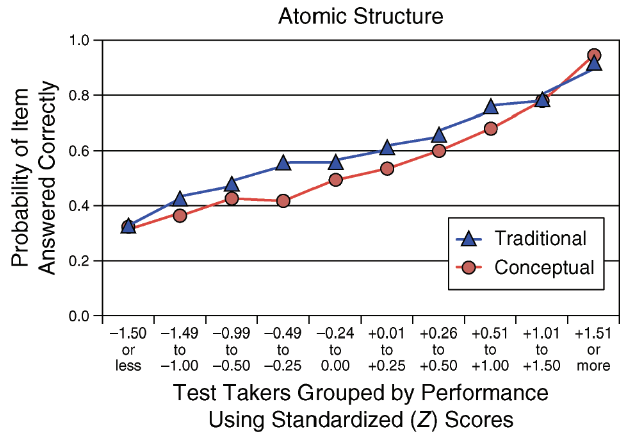


Figure 2. Item characteristic curve for the atomic structure pair.

theory. In this theory, the item difficulty is determined by calculating the fraction of students who answer it correctly. Thus, an item that is answered correctly by most students has a high value for the difficulty, an unfortunately counterintuitive scale. The second item statistic that is reported is discrimination, which is determined by calculating the fraction of correct answers among the top performing students minus the fraction correct among the bottom performing students as determined by their total score on that exam. The specific size of the sample for "top" and "bottom" is not prescribed by classical test theory, but for the data presented here, the value is 25% (i.e., top quarter and bottom quarter of students). The specific numbers obtained for discrimination are sensitive to differences in this choice, and with a large database, it would be possible to use the top and bottom 20%, for example. This choice would improve the discrimination index by an average of 9% relative to the data reported here.

The item statistics for ACS Exams are derived from voluntarily contributed data provided by 12 colleges for the first-term exam and 9 colleges for the second-term exam. Several schools participating in the data return have multiple sections of general chemistry, so there are more than 21 instructors for the students included in the data sets presented here. For the first-term exam, useable data were derived from 3073 student performances; for the second-term exam, 3557 student performances were included. Table 3 provides the item analysis from classical test theory for all 80 items contained in the two-paired-question exams.

These statistics provide some important observations about the performance of the two paired-questions exams. First, the overall pattern of difficulty between items in the pairs is evenly distributed between conceptual and traditional items. In particular, of the 40 item pairs available in the two exams, the conceptual item is less difficult for 19 pairs and the traditional is less difficult in 21 pairs. Second, in 9 out of the 12 cases where there is more than one item-pair, there is at least one pair with higher student performance on the conceptual item and one pair with higher performance on the traditional item. This characteristic suggests the exams may be useful for measuring the impact of teaching interventions both on a semester basis or for a specific content area.

It may be useful to consider the expectations for item performance based on the trial-test phase of the exams. For example, at the trial-test stage for the first-term exam, the average difficulty and discrimination of the conceptual items that were chosen for the released GC05PQF exam were 0.650 and 0.408, respectively. This same information for the traditional items that are part of the released exam was 0.625 for difficulty and 0.483 for discrimination. Thus, the design of the exam was to have the conceptual and traditional items similar in difficulty. The "predicted" difficulty was quite close for the conceptual items, but the traditional items tested slightly more difficult than when they were trial tested.

While the difficulty and discrimination statistics provide an important overview of student performance, they do not reveal the whole spectrum of student performance. A graphical means to broaden the analysis is to construct and plot the item characteristic curve (ICC). A common component of item response theory (IRT),[23] this construct graphs the performance of all levels of students, where discrimination only compares the top and bottom quarter. A number of software programs provide fitted ICCs for data, but the ICCs presented here are obtained by binning student performances and obtaining the difficulty of items for each subgroup. This methodology provides a view of item performance that is less homogenized, and therefore more illustrative for the comparisons drawn here.

To construct the binned ICC, a student's Z-score is obtained from the equation:

$$Z = \frac{\text{score} - \text{ave. score}}{\text{standard deviation}}$$

Thus, for example, a score that is one standard deviation lower than the mean has a Z-score of −1.0. To obtain sample sizes in each bin that allow for sensible determination of the difficulty for the ICC, the increment of the Z-score is adjusted to be smaller near the mean. No attempt is made to artificially equalize the number of performances in each bin; the Z-score alone determines in which bin a student's performance is counted. Thus, the increments used (along with the number of students in each bin) are provided in Table 4.

On the basis of this structure, the ICCs are constructed and plotted, with both the conceptual and traditional item in a pair on a single plot. All 40 plots are provided in the online Supporting Information, although some key examples are provided here. For many pairs, the difficulty difference is less than 0.1 (performance difference is less than 10%). Figure 2 provides an example of an ICC of such a pairing, in this case for the atomic structure question pair. While there are slight differences between performance on conceptual and traditional questions evident in this ICC, the items perform similarly for all levels of student proficiency. Five item pairs cover aspects of stoichiometry on

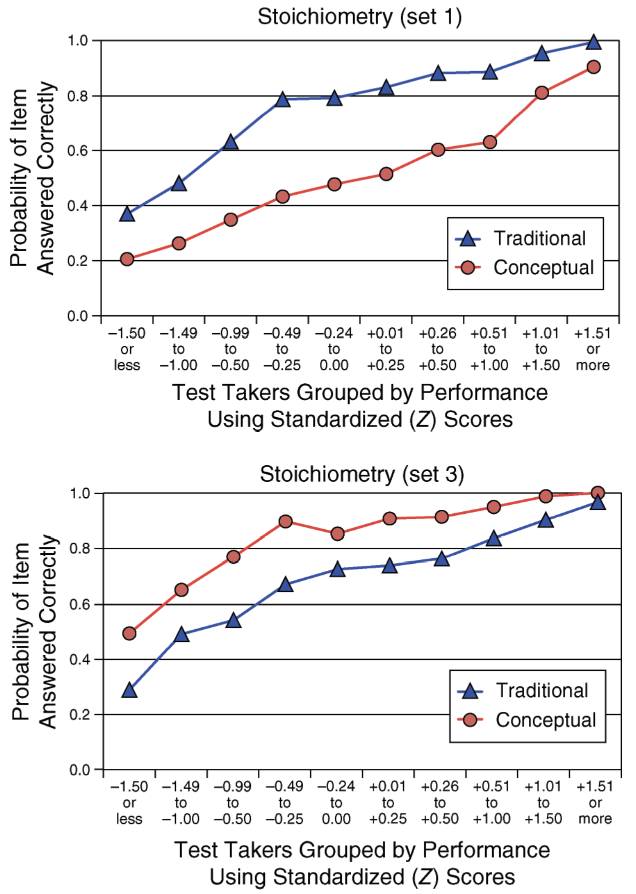


Figure 3. Item characteristic curves for two question pairs on stoichiometry.

the first-term exam. Figure 3 shows ICCs for two such pairs, one in which the performance is better on the traditional item and one in which it is better on the conceptual item.

These two plots are chosen to illustrate the increase in information provided by the ICC. The most apparent aspect in these graphs is that, in set 3, the performance on the conceptual item is clearly better than traditional item, while in set 1, the opposite is true. It is also possible to suggest more nuanced observations from these graphs. For example, in set 1, the largest differences in performance occur for students whose overall performance is closest to average: both high- and low-performing students have closer performances between traditional and conceptual items. Another key observations from set 3 is that here little difference is noted among the top-performing students. This is seen in a number of the pairs. Indeed, the average *difference* in difficulty in the top "bin" on the first-term exam is only 0.025, while the average difference among all groups is 0.060.

■ CONCLUSION

This report is designed to provide empirical data about student performance on items in the paired-question exams of the ACS Exams Institute. The structure of these exams is designed to not only assess student learning in each semester of general chemistry, but also to support research efforts about student learning in this course. In addition to classical item statistics, difficulty, and discrimination, the ability to consider the item characteristic curve allows users of these exams to compare performances on all levels of student proficiency.

The design of these two exams, benefiting from statistical analysis of trial testing, has resulted in two instruments that have roughly equivalent performances on conceptual and algorithmic items. This observation is made with national samples of more than 3000 students for each exam. Thus, the data provided here can be used for those who wish to compare how their students fare with either form of chemistry knowledge—perhaps based on specific teaching interventions.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Forty ICC plots for the first- and second-term exams. This material is available via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: taholme@iastate.edu.

■ ACKNOWLEDGMENT

The two general chemistry paired questions exams were created by a committee of excellent educators and researchers led by Diane Bunce. We gratefully acknowledge their efforts in the crafting of these exams. Use of ICCs in data analysis for ACS Exams originated in projects funded by the National Science Foundation, (DUE-0717769).

■ REFERENCES

(1) Nurrenbern, S. C.; Pickering, M. *J. Chem. Educ.* **1987**, *64*, 508–510.
(2) Pickering, M. *J. Chem. Educ.* **1990**, *67*, 254–255.
(3) Sawrey, B. A. *J. Chem. Educ.* **1990**, *67*, 253–254.
(4) Nakhleh, M. B. *J. Chem. Educ.* **1992**, *69*, 191–196.
(5) Nakhleh, M. B. *J. Chem. Educ.* **1993**, *70*, 52–55.
(6) Nakhleh, M. B.; Mitchell, R. C. *J. Chem. Educ.* **1993**, *70*, 190–192.
(7) Zoller, U.; Lubezky, A; Nakhleh, M. B.; Tessler, B.; Dori, Y. J. *J. Chem. Educ.* **1995**, *72*, 987–989.
(8) Nakhleh, M. B.; Lowrey, K. A.; Mitchell, R. C. *J. Chem. Educ.* **1996**, *73*, 758–762.
(9) Eubanks, E. D.; Eubanks, L. P. *General Chemistry Special Exam, First Term*; ACS Examinations Institute: Clemson, SC, 1997.
(10) Holme, T. A.; Murphy, K. L. *General Chemistry First Term Paired Questions Exam*; ACS Examinations Institute: Milwaukee, WI, 2005.
(11) Holme, T. A.; Murphy, K. L. *General Chemistry Second Term Paired Questions Exam*; ACS Examinations Institute: Milwaukee, WI, 2007.
(12) Niaz, M. *Int. J. Sci. Educ.* **1995**, *17*, 343–355.
(13) Smith, K. J.; Metz, P. A. *J. Chem. Educ.* **1996**, *73*, 233–235.
(14) Sanger, M. J.; Greenbowe, T. J. *J. Res. Sci. Teach.* **1997**, *34*, 377–398.
(15) Raviolo, A. *J. Chem. Educ.* **2001**, *78*, 629–631.
(16) Cracolice, M. S.; Deming, J. C.; Ehler, B. *J. Chem. Educ.* **2008**, *85*, 873–878.
(17) Holme, T. A. *J. Chem. Educ.* **2003**, *80*, 594–598.
(18) Novak, J. D.; Gowin, D. B. *Learning How To Learn*; Cambridge University Press: Cambridge, U.K., 1984.
(19) Pushkin, D. B. *J. Chem. Educ.* **1998**, *75*, 809–810.
(20) Evans, J. S. T. B. *Ann. Rev. Psych.* **2008**, *59*, 255–278.
(21) Zoller, U. *Res. Sci. Educ.* **1996**, *26*, 317–326. Zoller, U.; Tsaparlis, G. *Res. Sci. Educ.* **1997**, *27*, 117–130. Zoller, U.; Dori, Y.;

Lubezky, A. *Int. J. Sci. Educ.* **2002**, *24*, 185–203. Tsaparlis, G.; Zoller, U. *Univ. Chem. Educ.* **2003**, *7*, 50–57.

(22) Smith, K. C.; Nakhleh, M. B.; Bretz, S. L. *Chem. Educ. Res. Pract.* **2010**, *11*, 147–153.

(23) De Ayala, R. J. *The Theory and Practice of Item Response Theory*; Guilford Press: New York, 2009.