2012

Investigating Factors That Influence Item Performance on ACS Exams

Jacob Schroeder
Clemson University

Kristen L. Murphy
University of Wisconsin - Milwaukee

Thomas Holme
Iowa State University, taholme@iastate.edu

Investigating Factors That Influence Item Performance on ACS Exams

Abstract

General chemistry tests from the Examinations Institute of the Division of Chemical Education of the American Chemical Society have been analyzed to identify factors that may influence how individual test items perform. In this paper, issues of item order (position within a set of items that comprise a test) and answer order (position of correct answer relative to incorrect distractors) are discussed. Answer order is identified as potentially important, particularly for conceptually based items. When the correct answer appears earlier among the answer choices, there is some greater propensity for student performance to be better. Item-order effects are also possible, particularly when students encounter several challenging items consecutively. Performance on the next item may be lower than expected, possibly because of cognitive-load effects.

Keywords

first-year undergraduate/general, testing/assessment

Comments

Reprinted (adapted) with permission from J. Chem. Educ., 2012, 89 (3), pp 346–350. Copyright 2012 American Chemical Society.

JOURNAL OF CHEMICAL EDUCATION

Article

pubs.acs.org/jchemeduc

Investigating Factors That Influence Item Performance on ACS Exams

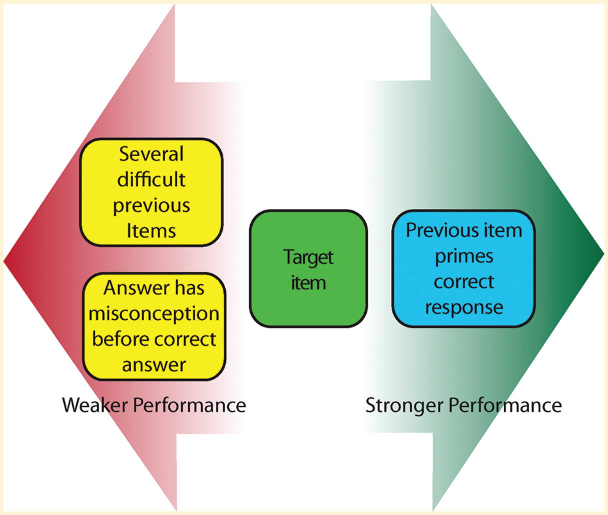
Jacob Schroeder,[†] Kristen L. Murphy,[‡] and Thomas A. Holme*[,§]

[†]Department of Chemistry, Clemson University, Clemson, South Carolina 29634, United States

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

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

ABSTRACT: General chemistry tests from the Examinations Institute of the Division of Chemical Education of the American Chemical Society have been analyzed to identify factors that may influence how individual test items perform. In this paper, issues of item order (position within a set of items that comprise a test) and answer order (position of correct answer relative to incorrect distractors) are discussed. Answer order is identified as potentially important, particularly for conceptually based items. When the correct answer appears earlier among the answer choices, there is some greater propensity for student performance to be better. Item-order effects are also possible, particularly when students encounter several challenging items consecutively. Performance on the next item may be lower than expected, possibly because of cognitive-load effects.


Several difficult previous Items
Answer has misconception before correct answer
Target item
Previous item primes correct response
Weaker Performance
Stronger Performance

KEYWORDS: *First-Year Undergraduate/General, Testing/Assessment*

INTRODUCTION

Testing student knowledge in any course is so fundamental to the educational endeavor that understanding factors that affect the fidelity of measurement in testing carries inherent importance. For over 75 years, the American Chemical Society's Division of Chemical Education (ACS DivCHED) has been developing standardized multiple-choice exams for assessing chemistry content knowledge in a variety of venues and grade levels. This program is currently administered by the Examinations Institute (EI), which provides norm-referenced exams in 12 areas: (i) high school chemistry; (ii) college placement exams; (iii) general chemistry; (iv) organic chemistry; (v) analytical chemistry; (vi) physical chemistry; (vii) inorganic chemistry; (viii) instrumental analysis; (ix) biochemistry, (x) diagnostic of undergraduate chemistry knowledge; (xi) polymer chemistry; and (xii) chemical safety. Statistical analysis of these exams generally finds them to be reliable measures in the content domain for which they are designed.

The manner in which a collection of items is used to construct an exam represents a critical factor for the development of any test, particularly for national, norm-referenced exams. The procedure used by the EI to devise exams places emphasis both on content and on the construction of the items themselves.[1] In terms of content, the key factor lies in the "grass roots" development of the content specifications—the choices of what topics to be included on an exam are determined by each committee at the outset of the exam writing process. There is no preascribed template of content determined by the ACS or the EI for what topics are tested. Practitioners who teach a course in which the exam is meant to be used determine the content coverage of each exam.

The role of item construct is closely monitored by the EI. Guidelines for item writing, along with extensive editing and trial testing of items, ensures high-quality items on each exam. Because of this rather extensive process, a typical ACS Exam usually requires an average of two−three years to develop before entering the field in its final form as an assessment instrument. It is not practical for individual instructors to spend this amount of time on exams for their course, so it may seem alarming that there are factors that influence what an ACS Exam measures. While not exhaustive, this paper presents some critical considerations for item performance within a multiple-choice exam and discusses possible ways to reduce construction-related variability of such an exam.

The EI constructs parallel versions of exams for general chemistry, where test environments are more likely to necessitate students sitting close to each other. These parallel versions are composed of identical items, but the order of these items is mixed (typically within sections of contextually related material). Additionally, the answer order for many of these

Published: January 19, 2012

dx.doi.org/10.1021/ed101175f | J. Chem. Educ. 2012, 89, 346−350

items is changed. The investigations reported and discussed in this paper are derived from analyzing the item statistics for these parallel versions from general chemistry.

LITERATURE REVIEW

The EI is by no means alone in its interest in developing high-quality assessment items. Recent reviews of the literature have sought to build a consensus in an effort to develop guidelines to address areas of concern for test validity.[2,3] Such concerns may include ways in which test-wise examinees can exploit the instructor's practice of key balancing.[4] These test strategies are learned behaviors that lead to answers that may not require cognitive engagement with the item content. They can include "edge aversion" and "edge attraction", *namely, avoidance of the response "A"*, depending on the choices or answers presented.[6] This leads examinees to often seek out correct answers in the middle of a set of choices, a tendency mirrored by some examiners by where they place correct answers.[7] Such strategies can be minimized by using *key randomization* instead of key balancing and *number right scoring* instead of formula scoring.[5] McLeod, Zhang, and Yu[8] found no inherent disadvantages to students when exam items or the order of the answer choices were randomized. Finally, there is evidence to suggest some benefit to limiting the number of answer choices, with three being considered ideal.[9]

The concept that item order plays a role in achievement tests is well established empirically. Much of the research associated with item-order effects[10] has been conducted to investigate these effects on test-equating algorithms used in large (typically >10,000 students in multiple locations) test programs. Test equating is required when students who are to be compared with an achievement test actually take different versions of the exam.

Several studies have suggested that changing the position of an item on an operational exam relative to its position during trial testing development leads to a change in the difficulty of the item.[11−14] The specific nature of the different position can also play a significant role in the observed difference.[15] In particular, when items were pretested early on an exam, and then moved to late in the exam, they were more difficult. When items appear late on the pretest and are moved to the early portion of the released exam, they were less difficult. Furthermore, item-order effects may affect low-proficiency students more than high-proficiency students.[16] Survey research has also focused considerable effort on response-order (answer-order) effects. Cognitive theory has been advanced that suggests these effects are more pronounced when respondents have "low cognitive sophistication".[17] Finally, it is important to note that for norm-referenced exams, cancelation of errors owing to item-order effects has been identified as rather common.[8,10]

Because the answer order on different versions of ACS exams in many cases is different, it is important to consider this aspect of item performance as well. Tellinghuisen and Sulikowski[18] have proposed that a particular statistical anomaly on the 2002 ACS First Term General Chemistry exam was due primarily to answer order effects. Their analysis led them to conclude that students were more successful at answering questions correctly when the correct answer for that question appeared earlier in the list of possible choices. Literature on answer order effects is less unanimous in terms of its conclusions. Again, most of the interest in this effect has been generated from survey research rather than studies on achievement tests.

One answer order effect that may be important at an exam level exists when one version of the exam contains a significantly larger number of correct answers in the first position, while the other version contains a large number of correct answers in the last position.[19] This observation may correspond to intuition about test taking. Students who see the correct answer first may take less time looking at the other possible answers on those items. As a consequence, if a significant number of "first-answer" items are on one version of the exam, students who take that version essentially have more time for other items.

In addition to statistical studies that identify anomalies related to either item-order or answer-order effects, it is important to consider reasons why they might arise. One aspect of cognition that may be helpful in understanding item-order and answer-order effects is "cognitive load".[20] Cognitive load theory is tied to observations that when an individual is working on cognitive tasks, a limit is reached in the capacity of the working memory that is used. In the case of a timed test, the variation in cognitive load as students move from one item to the next can ultimately help explain the role of cognitive fatigue in test taking.[21]

Another helpful view of learning is the concept of dual-processing theories of cognitive processing. A recent summary of these accounts[22] refers to them as system 1 processing (heuristic reasoning) and system 2 processing (analytical reasoning). In this sense, heuristic processes are fast processes that people may use without expending much cognitive effort—a frugal choice in test taking or similar tasks. Alternatively, analytical processes commonly engage more of the working memory—a more time-consuming method for achieving the cognitive task (answering the test item).

DATA COLLECTION

The ACS EI produces several variations of exams for the college general chemistry course, the most common being the full year, first term, and second term exams. The full year exam is released every other year during odd numbered years; thus, for example, the 2001 general chemistry exam will be referred to as GC01. The other two exams divide the content according to semester of instruction. They are typically released every three or four years and the shorthand notation adds an *F* for first term and *S* for second term (e.g., GC02F refers to the 2002 First Term General Chemistry exam.) For each of these exams, two versions are produced in which item positions and answer positions are scrambled. For older exams described here, the forms were distinguished by being printed on blue or gray paper, and for more recent exams, the paper colors were yellow or gray. This distinction will be noted parenthetically after the exam code. For example the 2002 First Term exam on gray paper is referred to as GC02F(G).

Users of ACS exams voluntarily return the results of student performances from national samples of typically 20 schools for the purpose of norm calculations. The EI has an online system that provides instant comparisons between user score inputs and national samples.[23] Once an instructor inputs such scores, they are contacted and encouraged to send in data for item statistics generation. Roughly half of instructors are able to provide these data. Thus, for the purpose of the data reported here, there is no a priori design for obtaining data; rather, the analysis represents an empirical study of student groups whose main commonality is that they took a general chemistry course that used an ACS Exam as a final. Table 1 reports the number of student performances for each sample.

Table 1. Sample Sizes, *N*, for Student Performances on Exams Analyzed

Exam	GC97	GC99	GC01	GC02F	GC03	GC05
Blue/Yellow	531	441	296	1106	783	503
Gray	514	209	465	1178	553	369

Within the constraint of this empirical sample, perhaps the simplest method to examine performance on an item level is to look at the difficulty index (DI) or the fraction of students that answered the item correctly. However, examining differences in DI can be misleading if the two groups under examination are not equivalent in proficiency. For example, if one subgroup has a higher proficiency, then it would be expected that they would perform better on an item and the corresponding DI would be higher. This would not be due to a differential performance of the item. If the two subgroups were matched on proficiency and a differential performance on the item remains, this would generally be an undesirable result for a test.

Differential item functioning (DIF) is an item-level characteristic in which an item may be found to be statistically easier for members of one subgroup than another.[24] DIF analyses typically involve matching examinees from different subgroups on proficiency, carrying out item analysis for each group, and evaluating the results for statistical significance. Where DIF is present, the item "favors" one group over another. Statistical techniques for detecting DIF include item response theory (IRT),[25] simultaneous item bias statistic (SIBTEST),[26] Mantel−Haenszel (M−H) statistic,[27] conditional *p*-value differences,[28] and logistic regression.[29] These techniques can be carried out on both multiple-choice and constructed-response items.

Subgroups are commonly based on demographics or those designated as germane to research or psychometric evaluation (such as gender, race or ethnicity, socioeconomic status, language ability). For the analyses presented here, the M−H statistic was used, matching examinees based on their overall performance on the exam with subgroups based *on the two different versions* of the test. This is not a common subgroup definition, yet it does serve the purpose of identifying the statistical significance of items that perform differentially on different forms of the exam. Both the M−H χ^2 value and significance (asymptotically distributed as a one degree of freedom χ^2 distribution) are calculated using SPSS.

■ RESULTS

One potential source of errors in a multiple-choice exam arises from answer order effects. A simple way to assess the importance of answer order on student performance is to identify where the correct answer is placed in an item, and to determine overall differences in difficulty based on that placement. Details of such an analysis can be obtained by identifying the distribution of earlier answers for each version of six general chemistry exams and checking student performance. This analysis is presented in Table 2. In addition to identifying the exams in column 1, this table shows the number of items with an earlier answer for each form of the exam in column 2. Note that all tests in the sample are 70-item tests, and therefore, the difference between the sum of items in column 2 and 70 is the number of items for which the correct answer appears in the same location on both forms. Column 3 provides the percentage of instances in which the item with the higher performance is also the item with the earlier correct answer, regardless of which form has that item. Columns 4−7 provide the difference in difficulty, both the maximum and average. For example, on the 2002 First Term General Chemistry exam (fourth row of data), 66 items had the answer earlier on either the blue or gray form of the exam. Students performed better 56% of the time when the correct answer appeared earlier. The maximum difference in difficulty reflects the two items at each extreme—the largest difficulty difference in favor of an earlier answer (9.3%) and the largest in favor of a later answer (6.2%).

Table 2. Observations Related to Answer-Order Effects on ACS General Chemistry Exams

			Maximum Difficulty Difference, %		Average Difficulty Difference, %	
Exam Code	Earlier Answer	Earlier Answer: Better Performance, %	Earlier	Later	Earlier	Later
GC97	29B; 26G	43.6	12.0	14.1	4.3	3.6
GC99	27B; 31G	51.7	16.0	10.3	4.4	4.3
GC01	38B; 31G	52.1	18.4	16.2	7.7	6.3
GC02F	33B; 33G	56.1	9.3	6.2	2.8	2.0
GC03	36B; 33G	50.7	9.7	7.7	3.1	2.9
GC05	30Y; 31G	54.1	19.7	21.0	5.6	6.0

Several key observations can be made from these data. First, the general trend is that students taking an exam with the correct answer appearing earlier are more successful than those with later correct answers roughly 50% of the time. In other words, students seem to perform equally well, regardless of the position of the answer choices. Second, looking at the extremes via the maximum difficulty difference identifies differences on individual items, and in some cases these differences can be rather large. However, when the differences are included for all items, these effects appear to offset leading to little difference in overall performance between the two versions of the exam. Finally, it should be noted that the earlier report on answer order effect from the GC02F exam[18] happens to be for the exam that has the highest performance difference of all tests studied.

On the basis of this final observation, it is possible to consider some details related to answer order for items on the GC02F exam. In addition to identifying the statistical difference of student performance on some items, Tellinghuisen and Sulikowski also speculate on possible reasons for the observed differences.[18] Specifically, they indicate that students may use noncontent strategies for answering questions, including the primacy effect (choosing an earlier answer) for accomplishing the cognitive task of answering the test item. While it is challenging to disprove the possibility that a primacy effect is operative, it is noteworthy that primacy is not normally associated with quantitative test items, while it is more commonly invoked in survey research. The answer-order result may just as likely be explained in terms of test economy—students who see the correct answer first may not look at the other choices (the distractors). At the same time, if the most common incorrect answer were placed earlier in the set of choices, students may second-guess their own content understanding and choose the earlier, incorrect answer.

Further item analysis of this exam highlights 11 items that exhibit significant differences between the two versions (Table

Table 3. GC02F Item Statistics Related to Answer-Order Effects

Item Number (Blue Version)	Difficulty Difference, %	M−H Statistics χ^2 Values	M−H Statistics *P* Values	Earlier Answer	Favors
10	7.1	14.870	<0.001	Gray	Gray
11	6.2	13.821	<0.001	Blue	Gray
20[a]	9.3	28.071	<0.001	Gray	Gray
27	3.6	4.834	0.028	Gray	Gray
37	5.3	10.479	0.001	Gray	Gray
38[b]	4.7	5.140	0.023	Blue	Blue
42	5.5	4.152	0.042	Blue	Blue
47	7.5	21.338	<0.001	Blue	Blue
52	3.5	5.818	0.016	Blue	Gray
57	5.6	8.564	0.003	Blue	Blue
59	4.1	4.052	0.044	Gray	Blue

[a] Item that requires explicit calculation to obtain answer. [b] Item that includes numeric reasoning to obtain answer.

Table 4. Characteristics and Difficulty of Prior Items Leading Up to 20B and 27G

Preceding Items	Item Concept	Difficulty Index[a]
17B	Identifying a compound with the same empirical and molecular formula	0.419
18B	Identifying an exothermic reaction	0.493
19B	Balancing equations	0.734
20B	Calculate moles of an atom given grams of the compound	0.450
24G	Calculate the percentage by mass of an atom given molar mass of a compound	0.742
25G	Balancing equations	0.765
26G	Calculate a dilution given the molarity of a solution	0.640
27G	Calculate moles of an atom given grams of the compound	0.543

[a] Lower difficulty index values indicate a more difficult item, one with lower student performance.

3). Using M−H statistics,[27] the difference in student performance is verified to be statistically significant. Of these 11 items, only one requires the students to choose an answer that results from a numerical calculation, and one other item includes quantitative content but not explicit calculations. Overall, the percentage of numerical items on GC02F was 28.6%, so the frequency of numerical items showing potential answer-order effects of 18.2% is lower than for the test as a whole. Indeed, when considering calculation-based answers, the overall exam has 20% of these items and answer order appears to arise on only 1 of the 14 such items. Thus, a disproportionately large number of the items that show differential performance are conceptual in nature. Moreover, student performance is higher when the correct answer is earlier on 8 of the 11 items. These data suggest that answer order may indeed play a role on item performance, particularly for conceptual items.

An important caveat related to student performance on these items is that in addition to the variability in answer order for the two versions of the exam is the variability in the item order itself. Even if answer order contributes to performance differences in many cases, answer order by itself is not capable of explaining all of the observed differences in item performance for different exam versions. In many cases in which significant differences in item performance exist, it seems reasonable to conclude that a combination of item-order effects and cognitive processing may play an important role. Consider the GC02F item pair with the largest difference in performance (20B, 27G). If one considers the content and performance of the preceding three items, a trend emerges that may also explain the differential performance (Table 4).

The prior questions leading up to item 20B are largely qualitative, requiring students to recall terminology (exothermic reactions, empirical and molecular formulas), yet 20B is quantitative. By comparison, the prior questions leading up to item 27G are mostly quantitative. Thus, while the data for these exams arise from test administrations "in the wild" rather than in controlled laboratory settings, the possibility of priming effects for student performances exists, akin to those seen in memory research.[30] Another trend exemplified in these data involves the pattern of the difficulty index for previous items. For students taking the blue exam, item difficulty index on the previous three items is, on average, lower than for the three preceding items on the gray exam. Not only was the content type different, the typical student success probability was also lower on the blue exam and students scored significantly lower on the target item. Thus, in addition to strictly cognitive models (priming quantitative skills on the gray exam), students who struggle with several items prior to the target item may also have differentially lower performance. *This effect (lower performance for a target item arising after lower performance on several preceding items) is the most common pattern observed for all instances of differential item performance on ACS General Chemistry exams.* Possible cognitive origins of this effect—for example, self-efficacy versus fatigue—are currently being investigated further.

■ DISCUSSION

Insofar as examinations are instruments that measure student proficiency in a particular content domain, the presence of measurement error is unavoidable. Nonetheless, the ability to determine factors that may exacerbate the severity of measurement error is a critical research step in the development of improved tests. The existence of multiple forms of ACS Exams provides one means to identify some such factors, as reported here. Moreover, because exams are used to evaluate student work for grades, additional factors, such as security, must be considered. The prime motivation for having multiple forms of exams is to reduce opportunities for cheating, which is also a factor in measurement error. In a real sense, therefore, a need exists for a thorough analysis of many variables related to the fidelity of measurement of student learning.

At least three factors have been identified from the analysis presented here that may be worth attention for both national exam programs and instructor-generated exams that use multiple forms.[1] Answer-order effects may be important on chemistry tests, and are potentially more important for conceptual questions. A number of possible reasons why answer order can influence item performance are known, including the fact that students may not read all possible answers once they find one they believe to be correct. If the correct answer is earlier on the list, students may be less likely to be distracted by incorrect answers that follow it.[2] Item-order effects can arise from priming—when students on one form of the exam are carrying out cognitive tasks that are similar to those needed in a specific item, just prior to answering that item. Such priming may allow students who are less fluent in the content to answer correctly more often than more proficient students, essentially because they have used hints available to them.[3] The role of cognitive complexity, as manifested in item difficulty, is another important factor in student performance. Specifically, if students are required to do several challenging test items in a row, their performance on a subsequent item is often lower than similarly proficient students who do not have a set of challenging items prior to the question on their form.

The Exams Institute has already devised an example of an adjustment that is sensitive to this final factor. A recently released examination, the Diagnostic of Undergraduate Chemistry Knowledge,[31] was constructed from 15 scenarios, each with four test items. Because all of the scenarios were chosen based on data from trial testing, the ordering of the scenarios was intentionally adjusted to ramp from the scenario with the least difficult items on average, to the scenario with the most difficult items. While the preliminary assignment of item difficulty from trial testing is not an assurance of ultimate performance of the item on the released exam,[11−14] building exams with this type of structure may help reduce student errors related to cognitive load. Item statistics from the released version of the exam do show that the goal of ramping scenarios by difficulty based on the trial test data has been largely successful.

It is important to acknowledge at this point that a post-hoc analysis of exams that have been used as large-scale assessments has inherent limitations. For example, in the case of ACS Exams, the different forms of general chemistry exams invariably scramble both item order and answer order. Without specific control for one of these variables, it is inherently impossible to identify which factor (or indeed any other currently unforeseen factor) is responsible for differential performance on a test item. Moreover, ACS Exams are used in high-stakes testing, so intentional experiments that might disadvantage any group of students are inherently unethical.

Thus, the current findings mostly provide preliminary evidence about factors that influence student performance on test items. These factors can inform hypotheses that might be tested in lower-stakes environments, such as practice exams. Indeed, this methodology for research is readily incorporated into the recently developed practice exam system from the Institute.[32] As the practice exam is ported into an electronic delivery system and obtains larger numbers of student performances, the ability to control for specific aspects of item design (for example, either item order or answer order) will be available to advance the data collection for questions such as those raised by the observations presented here.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: taholme@iastate.edu.

■ ACKNOWLEDGMENTS

The authors gratefully acknowledge conversations with A. Zenisky on several aspects of this work. Portions of the work described here were funded through the National Science Foundation via grants DUE 0817409 and DUE 0943783.

■ REFERENCES

(1) Holme, T. A. *J. Chem. Educ.* **2003**, *80*, 594−598.

(2) Haladyna, T. M.; Downing, S. M.; Rodriguez, M. C. *Appl. Meas. Educ.* **2002**, *15*, 309−333.

(3) Frey, B. B.; Petersen, S.; Edwards, L. M.; Teramoto Pedrotti, J.; Peyton, V. *Teach. Teach. Educ.* **2005**, *21*, 357−364.

(4) Bar-Hillel, M; Attali, Y. *Am. Stat.* **2002**, *56*, 299−303.

(5) Bar-Hillel, M.; Budescu, D.; Attali, Y. *Mind & Society* **2005**, *4*, 3−12.

(6) Christenfeld, N. *Psychol. Sci.* **1995**, *6*, 50−55.

(7) Attali, Y.; Bar-Hillel, M. *J. Educ. Meas.* **2003**, *40*, 109−128.

(8) McLeod, I.; Zhang, Y.; Yu, H. *J. Stat. Educ* **2003**, *11* (1), 1−6.

(9) Rodriguez, M. C. *Educ. Meas.: Issues and Practice* **2005**, *24*, 3−13.

(10) Meyers, J. L.; Miller, G. E.; Way, W. D. *App. Meas. Educ.* **2009**, *22*, 38−60.

(11) Whitely, S. E.; Dawis, R. V. *Educ. Psych. Meas.* **1976**, *36*, 329−337.

(12) Eignor, D. R.; Stocking, M. L. *An Investigation of Possible Causes for the Inadequacy of IRT Pre-Equating*, Educational Testing Services Report ETS-RR-86-14; Educational Testing Services: Princeton, NJ, 1986. ERIC number ED275695; available online at http://eric.ed.gov/ERICWebPortal/search/simpleSearch.jsp;jsessionid=mdy9Vz6OphWCkAHZVKZokA__.ericsrv004?newSearch=true&eric_sortField=&searchtype=keyword&pageSize=10&ERICExtSearch_SearchValue_0=ED275695&eric_displayStartCount=1&_pageLabel=ERICSearchResult&ERICExtSearch_SearchType_0=no (accessed Jan 2012).

(13) Doerner, W. M.; Calhoun, J. P. The Impact of the Order of Test Questions in Introductory Economics. Available at SSRN: http://ssrn.com/abstract=1321906 (accessed Jan 2012).

(14) Sue, D. L. *J. Econ. Educators* **2009**, *9*, 32−41.

(15) Meyers, J. L.; Miller, G. E.; Way, W. D. *App. Meas. Educ.* **2009**, *22*, 38−60.

(16) Huntley, R. M.; Welch, C. *The Effect of Answer Location on Item Difficulty and Discrimination in Language-Usage Tests* (ACT Research Report); American College Testing: Iowa City, IA, 1988.

(17) Krosnick, J. A.; Alwin, D. F. *Public Opin. Q.* **1987**, *51*, 201−219.

(18) Tellinghuisen, J.; Sulikowski, M. M. *J. Chem. Educ.* **2008**, *85*, 572−575.

(19) Bresnock, A. E.; Graves, P. E.; White, N. *J. Econ. Educ.* **1989**, 239−245.

(20) Paas, F. G. W. C.; van Merrienboër, J. J. G. *J. Educ. Psych.* **1994**, *86*, 122−133.

(21) Paas, F.; Tuovinen, J. E.; Tabbers, H.; van Gerven, P. W. M. *Educ. Psychol.* **2003**, *38*, 63−71.

(22) Evans, J. St. B. T. *Ann. Rev. Psych.* **2008**, *59*, 255−278.

(23) American Chemical Society, Division of Chemical Education Examinations Institute, 2010. http://chemexams.chem.iastate.edu/stats/score_reporting/index.cfm (accessed Jan 2012).

(24) Osterlind, S. J.; Everson, H. T. *Differential Item Functioning*, 2nd ed. (161 in the Series of Quantitative Applications in the Social Sciences); Sage Publications: Newbury Park, CA, 2009.

(25) Hambleton, R. K.; Swaminathan, H.; Rogers, H. J. *Fundamentals of Item Response Theory*; Sage: Newbury Park, CA, 1991.

(26) Shealy, R.; Stout, W. *Psychometrika* **1993**, *58*, 159−194.

(27) Holland, P. W.; Thayer, D. T. In *Test Validity*; Wainer, H., Braun, H. I., Eds.; Lawrence Erlbaum: Hillsdale, NJ, 1988; pp 27, 129−145.

(28) Zenisky, A. L.; Hambleton, R. K.; Robin, F. *Educ. Psych. Meas.* **2003**, *63*, 51−64.

(29) Swaminathan, H.; Rogers, H. J. *J. Educ. Meas.* **1990**, *37*, 361−370.

(30) Johns, E. E.; Mewhort, D. J. K. *J. Exp. Psychol. Learn. Mem. Cogn.* **2009**, *35*, 1162−1174.

(31) Holme, T. A.; Murphy, K. L. *Diagnostic of Undergraduate Chemistry Knowledge Exam*; ACS Examinations Institute: Ames, IA, 2009.

(32) Knaus, K.; Murphy, K.; Holme, T. *J. Chem. Educ.* **2009**, *86*, 827−832.