2013

Identifying Differential Performance in General Chemistry: Differential Item Functioning Analysis of ACS General Chemistry Trial Tests

Lisa Kendhammer
University of Wisconsin - Milwaukee

Thomas Holme
Iowa State University, taholme@iastate.edu

Kristen Murphy
University of Wisconsin - Milwaukee

Identifying Differential Performance in General Chemistry: Differential Item Functioning Analysis of ACS General Chemistry Trial Tests

Abstract
The development of high-quality assessments can be an intensive process. At the American Chemical Society Examinations Institute (ACS-EI), this follows a general process of test design and content mapping, item construction, trial testing, item analysis, and final test setting. The item analysis portion of this procedure is an important step in using field-testing results to select the best items. This selection is based on validity analysis by experts, both field-test users and test writers, and by students in field testing. The traditional item analysis of the ACS-EI now includes a differential item functioning analysis with subgroups by gender when a sufficient data set is available. The results of this analysis from six trial tests were further evaluated by both content and format. Trends of specific content areas and by format of items reveal that classes of items favored one subgroup over another in at least four instances.

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

JOURNAL OF CHEMICAL EDUCATION

Article

pubs.acs.org/jchemeduc

Identifying Differential Performance in General Chemistry: Differential Item Functioning Analysis of ACS General Chemistry Trial Tests

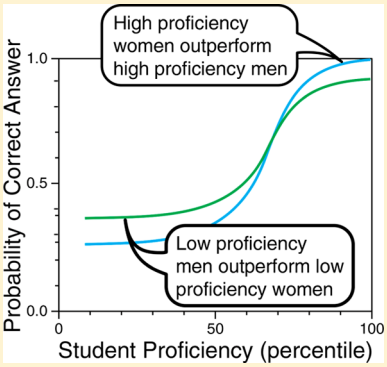
Lisa Kendhammer,[†] Thomas Holme,[‡] and Kristen Murphy*[,†]

[†]Department of Chemistry and Biochemistry, University of Wisconsin−Milwaukee, Milwaukee, Wisconsin 53201 United States

[‡]Department of Chemistry, Iowa State University, Ames, Iowa 50010 United States

Supporting Information

ABSTRACT: The development of high-quality assessments can be an intensive process. At the American Chemical Society Examinations Institute (ACS-EI), this follows a general process of test design and content mapping, item construction, trial testing, item analysis, and final test setting. The item analysis portion of this procedure is an important step in using field-testing results to select the best items. This selection is based on validity analysis by experts, both field-test users and test writers, and by students in field testing. The traditional item analysis of the ACS-EI now includes a differential item functioning analysis with subgroups by gender when a sufficient data set is available. The results of this analysis from six trial tests were further evaluated by both content and format. Trends of specific content areas and by format of items reveal that classes of items favored one subgroup over another in at least four instances.



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

FEATURE: Chemical Education Research

INTRODUCTION

Instruction and assessment are primary classroom activities of any educator. The assessment component, or testing whether the learning has occurred and to what degree, can take many different forms. Following the testing event, data are compiled and judgments are made about the data. These lead to decisions about the students, including what grade a student will get on an exam or in a course. Ideally, test items are checked for validity, in order for these decisions to be as free of bias or favor as possible. Part of these checks can include examining for differential performance on test items by different subgroups. Differential item functioning (DIF) occurs when two groups matched on a relevant measure of proficiency should perform the same on a test item yet show a statistical difference in performance.[1]

The American Chemical Society Examinations Institute (ACS-EI) of the ACS Division of Chemical Education provides norm-referenced, standardized exams for all levels of chemistry from second-year-level high school to graduate entrance exams. The process by which the exams are developed mitigates some concerns about item and test validity.[2] The trial testing phase of all ACS-EI exams provides statistical analysis for the selection of items for the active, norm-referenced exam. As part of this process, the ACS-EI conducts DIF analysis and provides the results of this analysis to the exam development committee when selecting the final set of test items for the active exam. Items that show significant statistical measures of DIF may be excluded from the final version. By virtue of the analysis, several possible test items have been putatively identified as showing DIF. These items, from exam development over the last four years of general chemistry exams, are described and analyzed here.

LITERATURE REVIEW

Differences between gender subgroups have been studied extensively in the past 35 years, including a 1974 study that examined gender differences in 1600 studies in eight areas, including achievement. One conclusion (among others) reached was that although female students have greater verbal ability, male students have better math and visual−spatial ability.[3] More recently, Educational Testing Services (ETS) has conducted a large-scale gender study examining multiple exams over multiple grade levels (from fourth grade to graduate entrance) from low-stakes to high-stakes testing and found that the previously identified gaps between genders in overall performance on mathematics and science (which favored male students) has decreased substantially, although the gender gap on overall performance on verbal assessments still favors female students.[4] This study also found that the previously asserted advantage for male students on multiple-choice (MC) items is nonexistent, although constructed-response (CR) items were found to sometimes favor female students. Finally, another

Published: May 7, 2013

1997 study found that male students tend to perform higher on tasks involving visual−spatial reasoning, although the link between this ability and the performance on mathematics or physical science assessments was not conclusive.[5] While these studies investigated the possibility of bias within an overall testing environment, the role of item content and format remains important to understand further. Indeed, with enhanced data about how individual items exhibit bias, the ability to incorporate this knowledge may be more likely to be included in the writing process of a standardized assessment.

Differential item functioning occurs when two groups matched on a relevant measure of proficiency should perform the same on a test item yet show a statistical difference in performance.[1] Subgroups can be matched either internally, where the overall performance on the assessment serves as the measure of ability, or externally, using a separate and relevant measure of ability. DIF can be identified on both MC and CR assessments and is identified through a number of methods, including item response theory (IRT),[6] simultaneous item bias statistic (SIBTEST),[7] Mantel−Haenszel statistic,[8] and logistic regression.[9] Subgroups can be differentiated by gender, ethnicity, socioeconomic status, language ability, or others characteristics. Using these methods, DIF can be identified as uniform or nonuniform. Uniform DIF applies when one subgroup consistently outperforms the other at all ability levels. Nonuniform DIF pertains when one subgroup starts out having higher performance than the other for a lower ability group, but when moving to a higher ability group the performance of the subgroups switch (e.g., lower-performing male students may outperform lower-performing female students, but higher-performing female students may outperform higher-performing male students). Examining an item characteristic curve (ICC) will normally reveal instances of nonuniform DIF.[10]

Many DIF studies are conducted via standardized exams combining performances from many institutions (e.g., statewide standardized testing of students in K−12), often requiring that abilities be matched internally, which consists of the score they received on that exam. The students must be matched on their ability levels, otherwise bias may be created by matching a high-performing student against a low-performing student. Because possible DIF exists, the matching criterion is affected by the presence of DIF.[11] In order to remedy this, a two-stage iterative process has been proposed[8] in which a DIF analysis is conducted, DIF items are identified (via a threshold for that particular method), and these items removed (the first stage). The matching criterion (the test score) is now refined and the DIF analysis is conducted again with any remaining DIF items identified.

This process was used in two studies of large-scale assessments administered to elementary, middle, and high school students using a process called the "weighted two-stage conditional *p*-value comparison procedure".[12,13] In the first study, the two-stage process versus the single-stage process for DIF analysis was examined in three subject areas (language arts, mathematics, and science) and found that there was a change between using a two-stage versus a one-stage process (in either an addition or reduction in the number of items exhibiting potential DIF) in 15 out of the 18 tests examined (three subject areas, three grade levels, and two test forms). In most cases, the two-stage process resulted in more items found as exhibiting possible DIF with the greatest difference (between a one- and two-stage process) in the mathematics and science tests given on the high school level.[12]

In the second study, the two-stage process was used to identify possible DIF in science only and categorize these items (based on content, visual−spatial or reference component, and item type). Of all the items analyzed, both MC and open-ended response (OR), 60 items were found to exhibit minor to high DIF, with 52 (87%) favoring male students, and all of the moderate to high DIF items (29 of the 60) favoring male students. Categorizing these flagged items by content, 17 were earth and space science, 8 were inquiry, 7 were life science, 13 were physical science, and 15 were technology. Of these items, only 1 in technology, 2 in both life science and earth and space science, and 3 in inquiry favored female students (with none in physical science). Categorizing these items by visual−spatial or reference components revealed that of the full set of items, 82 items contained a visual−spatial or reference component (with 71 items containing one component and 11 items contained two components), including graphs, maps, tables, and so on. Further, 30% of the items containing one or two of these components were flagged. Of the flagged items, the majority had three specific components: maps, diagrams, or MC with picture responses. All of these items, with one exception (a diagram item) favored male students. It is important to note that many of these items were in content areas in which DIF was previously found to favor male students, and content areas in which female students were favored used fewer visual−spatial or reference components (life science). Finally, item type was categorized as MC or OR, and of the 37 OR items, only 3 items were flagged and all favored female students. Accordingly, the remaining flagged items were MC and overwhelmingly favored male students.[13]

Research into identifying and categorizing DIF on mathematics assessments has been conducted through a number of studies. Many assessments measure higher-order cognitive processes requiring examinees to use multiple skills (e.g., reading ability and mathematics ability[14]). Research has shown that DIF can occur when test items measure more than one skill and differences exist between the types of skills tested.[15] Sources of DIF were examined in a 2006 study of 12th grade student performance on the British Columbia Principles of Mathematics Exam.[16] When DIF was identified as favoring male students, it was in the content area of problem solving (story problems or noncontext specific) and when multiple skills were required. When DIF was identified as favoring female students, it was bundled as computational items when an equation was not provided. Finally, a study of a mathematics test (Midwestern Mathematics Placement Exam) administered to first-year college students was examined for DIF with regards to study item order and item content.[17] No significant DIF was found between genders for changes in item order, however, significant DIF that favored male students was found for word problems.

Fewer studies examining DIF on science assessments have been conducted. A 1998 study[18] examined the 1994 State Performance Assessment for fifth and eighth grade students in science for DIF and compared DIF between the two grades using a combination of examining item characteristic curves[10] and IRT. DIF was found for 18 items described as science knowledge items, for 5 science and reading combination items, and for 8 science and mathematics combination items. In all these cases of DIF, the female students were favored, which is consistent with studies mentioned above given the format of the items (CR) and writing components that were necessary to perform well on the items. In another study,[19,20] CR items on a

dx.doi.org/10.1021/ed4000298 | J. Chem. Educ. 2013, 90, 846−853

Table 1. Trial Tests Analyzed for Differential Item Functioning by Sex in General Chemistry

Respondents	General Chemistry (First Term) 2010, N		General Chemistry (First Term) 2012, N		General Chemistry (Full Year) 2011, N		General Chemistry (Full Year) 2013, N	
	Form A	Form B	Form A	Form B	Form A	Form B	Form A	Form B
Male students	735	758	374	346	214	143	158	359
Female students	504	518	286	330	265	212	140	306
Number of items analyzed	70	70	70	70	70	0[a]	0[a]	70
Number of institutions participating in trial testing	7	7	6[b]	3	5			11

[a]Analysis was not done because there was not a necessary minimum of students in either or both subgroups. [b]One institution provided 11 performances that were all of the same subgroup.

science assessment administered as part of the National Education Longitudinal Study of 1988 given to 12th graders was studied for DIF using a logistic regression procedure called logistic discrimination function analysis (LDFA)[21] appropriate for polytomously scored items. Of the four items studied, one item in particular exhibited a large DIF favoring male students, large enough for a significant difference in overall test performance between male and female students. In addition, 25 MC items also administered to 12th graders as part of the same assessment study were examined using factor analysis and found to load to three dimensions: spatial−mechanical reasoning (visualization and prediction), quantitative science (application of specific factual knowledge), and basic knowledge and reasoning (verbal reasoning ability), which were based on item content and student responses during interviews.[22,23] Using these dimensions, the item found with the largest DIF, which favored male students regardless of the matching criterion used, was based on content (physical science) coupled with spatial−mechanical reasoning.

More recently, an analysis was conducted on two tests of formal reasoning ability: Test of Logical Thinking (TOLT)[24] and Group Assessment of Logical Thinking (GALT).[25,26] The TOLT has 8 item pairs (combining for a single correct or incorrect score) and 2 additional open-response items (scored dichotomously).[24] The GALT has 10 item pairs and two additional items testing concrete reasoning.[25] Both uniform[8] and nonuniform DIF[9] analyses were conducted on items from both tests. Additionally, both tests were examined with two different populations: students in general chemistry I and students in preparatory chemistry. The results of the DIF analysis on the TOLT revealed 1 item pair (of the 8 total) with uniform DIF for either population. Depending on the population studied, the results for the GALT were between 1 and 2 items with nonuniform DIF and 2 and 3 items with uniform DIF.

METHOD

Use of the word "gender" in this paper is not meant to imply that the study was conducted on the socially constructed associations of gender but rather on the biological differences of sex (in which students would be classified as male or female).[27,28] However, because the literature on DIF uses the term gender almost universally to discuss subgroups differentiated by sex, the use of the word "gender" in this paper reflects usage consistent with the literature. Within the tables, the word "sex" is used because these were the data collected.

DIF was investigated on ACS-EI trial exams for college general chemistry, examining both forms of the trial tests (form A and form B). Trial tests as produced by ACS-EI consist of two exams with almost 100% unique items. Occasionally, an exam committee will opt to include the same item on both trial tests. Of the tests examined in this study, all items on all trial tests were unique. Therefore of any two trial exams, each with 70 items, 140 unique items would be tested. Trial exam performance data were collected by requesting instructors nationally to use the trial tests for a single semester in place of a regular final exam (typically an active ACS-EI exam). Trial testing involves sampling of students from multiple schools and, in some cases, multiple instructors at an individual school. Instructors self-select for participation in the process and are often associated with the test development committee or have previous experience with ACS exam development. Nonetheless, there is little reason to expect that the national samples analyzed in this study are based on students who are exposed to unusual content or pedagogical circumstances. An instructor who agrees to trial test an exam will receive exam booklet copies, optical answer sheets ("bubble sheets"), instructions for administering the exam, and a report to be filled out by the instructor detailing how the exam was administered. All materials must be returned to the ACS-EI for use in assembling the trial test statistics. Instructors who return student performance data are also requested to submit gender data for their students in order for the DIF analyses to be conducted. If incomplete data are returned by an instructor, an attempt is made to assign gender based on a student's name using a check for the probability of gender.[29] Gender is assigned if the probability is more than 50% greater for the name to be a particular sex. Thus, for example, a name such as Taylor, which is only 1.18 times more likely to be a male, is not assigned and that student's performance is not included in the DIF analysis.

Of the trial tests given in general chemistry over the past four years, there have been four trial exams with the necessary minimum of student performances on at least one trial exam to conduct the analysis.[30,31] The types (full-year or first-term exams), number of students, number of test items, number of institutions providing data for the analysis, and the forms studied are given in Table 1. All of the institutions provided performance data for students in both subgroups with the exception of one institution for GC12F (general chemistry 2012 first-term exam), form A, for which 11 performances were from one subgroup.

Analyses for uniform DIF were conducted using a Mantel−Haenszel (MH) statistic[8] in conjunction with a simple difference in difficulties (the number of students who answered correctly out of the number of students who answered the question)[32] by ability within subgroup and examining the ICC for each item.[10] The Mantel−Haenszel statistic was selected for the analysis because of the availability of the technique (accessed through SPSS) and the common use of the method in the literature. The results of the MH method were confirmed by the uniform DIF results from the logistic regression analysis.

dx.doi.org/10.1021/ed4000298 | J. Chem. Educ. 2013, 90, 846−853

A one-stage analysis was conducted for determining the items for examination; however, a two-stage analysis is also conducted routinely (to examine for changes to the number of DIF items identified). Regularly, the number of DIF items remains constant, with the significance values changing only slightly. Analyses for nonuniform DIF were conducted using logistic regression, with the threshold of significance established as 0.01 to lower the probability of a false positive.[9] All analyses were conducted using SPSS.

From the perspective of exam development, portions of the resulting data are provided to the development committees. In particular, the committees were provided with the items that exhibited possible DIF, along with instructions on how to consider acting on this information. Committees were instructed in some cases that items were unable to be used on the active exam because of the probability that they would contribute to an instrument with a built-in bias toward a subgroup of test-takers. Committees were also instructed that this was a single analysis with no additional information implying that the statistical probability that the item may exhibit DIF does not guarantee that it does exhibit DIF. Finally, committees were provided only with item numbers and not with any direction (or favor) of possible DIF.

The DIF-flagged items were further examined for both content and construct (or format of the items). The content was broken down into general categories (of specific content areas covered in a typical general chemistry course) and analyzed separately by three general chemistry instructors (two of the authors and one additional rater). The initial assignments were analyzed for agreement and where there was not full agreement, an additional discussion occurred until full agreement had been reached. Additionally, the items were examined for inclusion of a visual−spatial or reference component, reasoning, computation, or specific chemical knowledge. Again, the initial assignments were analyzed for agreement and occasionally, where there was not full agreement, an additional discussion occurred until full agreement had been reached.

RESULTS

General Chemistry First-Term Trial Tests

From the four general chemistry first-term trial tests (GCF; GC10F and GC12F), 280 items were examined for both uniform and nonuniform DIF. A total of 61 items (22%) had a significant value via either the MH or logistic regression analysis that suggests the possibility of either uniform or nonuniform DIF. The number of items that were found to statistically exhibit uniform or nonuniform DIF by test year and form are given in Table 2, with the direction of the possible DIF indicated.

In addition to the DIF statistics, the items were further examined for content and format. The results of this analysis are shown in Figure 1 (content area) and Figure 2 (format of item), with the number of items within the specific areas broken down by the direction of favor of possible DIF. Tables of items by content or format and direction of favor of possible DIF broken down by individual test can be found in the Supporting Information. The results do not indicate any clear trends with regards to content areas (Figure 1). There is a weak trend with regards to the format of the items, particularly considering items on the GC12F exams, where inclusion of a visual−spatial or reference component, or testing specific chemical knowledge, exhibited possible DIF favoring female students (Figure 2). The trend also shows more items containing a computational or reasoning component favoring male students.

Table 2. Uniform and Nonuniform DIF Items on First-Term General Chemistry Trial Exams, Separated by Subgroup

Item Uniformity or Nonuniformity	Respondents	Number of Items on These Trial Exams Exhibiting DIF[a]			
		GC10F[b] Form A	GC10F[b] Form B	GC12F[c] Form A	GC12F[c] Form B
Exhibiting possible uniform DIF[a]	Females	7	9	4	9
	Males	9	12	3	6
Exhibiting possible nonuniform DIF[a]	—	1	0	1	0

[a]Differential item functioning (DIF) by sex. [b]GC10F: General chemistry, 2010, first-term. [c]GC12F: General chemistry, 2012, first-term.

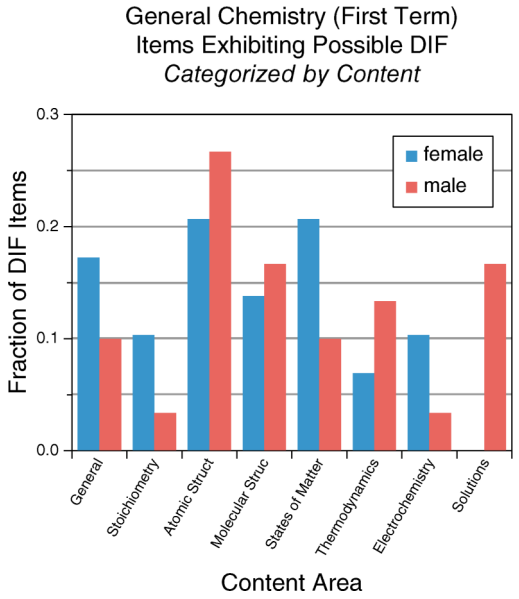


Figure 1. Possible DIF items on general chemistry first-term trial tests categorized by content and favor of possible DIF. Items from all four trial exams are combined. The fraction represents the number of items in that category out of a total of 59 items.

General Chemistry Full-Year Trial Tests

From the two general chemistry full-year trial tests (GC; GC11 and GC13), 140 items were examined for both uniform and nonuniform DIF. A total of 27 items (19%) had a significant value via either the MH or logistic regression analysis, suggesting the possibility of either uniform or nonuniform DIF, as given in Table 3, which also reports the direction of the possible DIF.

Once again, the items were further examined by content and format. The results of this analysis are shown in Figure 3 (content area) and Figure 4 (format of item) with the number of items within the specific areas broken down by the direction of favor of possible DIF. Given the low number of items within each group, there is no observable trend by individual trial exam or in an aggregate of both trial exams.

dx.doi.org/10.1021/ed4000298 | J. Chem. Educ. 2013, 90, 846−853

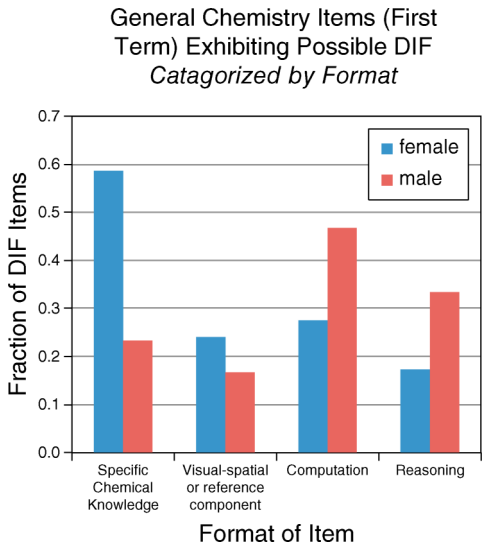


Figure 2. Possible DIF items on general chemistry first-term trial tests categorized by format and favor of possible DIF. Items from all four trial exams are combined. Items could be placed in more than one category (so the totals are greater than 100%). The fraction represents the number of items in that category out of a total of 59 items.

Table 3. Uniform and Nonuniform DIF Items on Full-Year General Chemistry Trial Exams, Separated by Subgroup

Item DIF Uniformity or Nonuniformity[a]	Respondents	GC11[b] Form A	GC13[c] Form B
Exhibiting possible uniform DIF	Females	6	4
	Males	7	4
Exhibiting possible nonuniform DIF	—	3	3

[a]Differential item functioning (DIF) by sex. [b]GC11: General chemistry, 2011. [c]GC13: General chemistry, 2013.

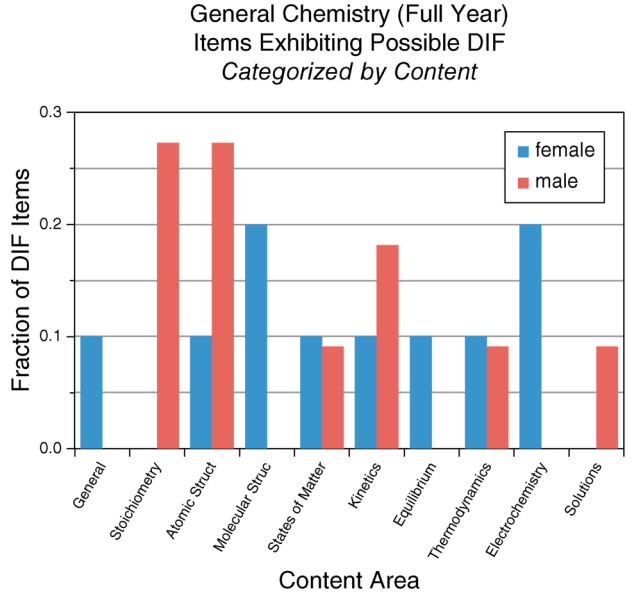


Figure 3. Possible DIF items on general chemistry full-year trial tests categorized by content and favor of possible DIF. Items from both trial exams are combined. The fraction represents the number of items in that category out of a total of 21 items.

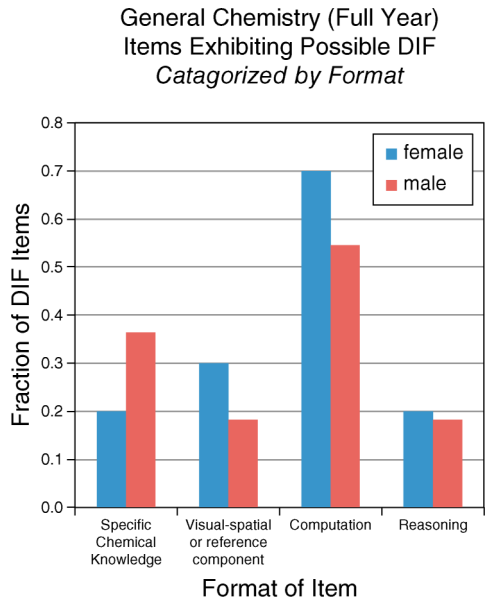


Figure 4. Possible DIF items on general chemistry full-year trial tests categorized by format and favor of possible DIF. Items from both trial exams are combined. Items could be placed in more than one category (so the totals are greater than 100%). The fraction represents the number of items in that category out of a total of 21 items.

Combined Categorization

When combining the results by both content area and format of items, the lack of a trend (favoring one gender over the other) remains when categorizing the items by content (Figure 5). However, when examining the items by format, the previous trend from the first-term trial exams remains (Figure 6). Format categories of visual–spatial or reference component and specific chemical knowledge tend to be included in items that favored female students; inclusion of a computational or reasoning component tends to favor male students. Because the

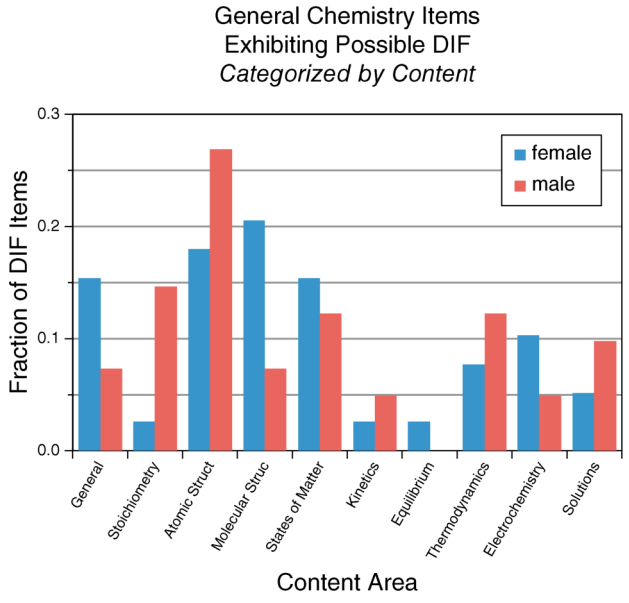


Figure 5. Possible DIF items on all general chemistry trial tests, categorized by content and favor of possible DIF. The fraction represents the number of items in that category out of a total of 80 items.

dx.doi.org/10.1021/ed4000298 | J. Chem. Educ. 2013, 90, 846–853

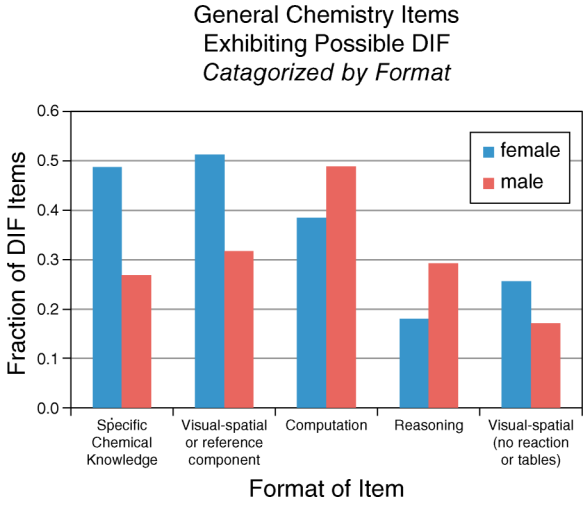


Figure 6. Possible DIF items on all general chemistry trial tests categorized by format and favor of possible DIF. Items could be placed in more than one category (so the totals are greater than 100%). The fraction represents the number of items in that category out of a total of 80 items.

inclusion of a visual–spatial or reference component favoring female students is unexpected considering the literature, the type of visual–spatial or reference component was examined more closely. Any chemical reaction, equation, table, figure, or graph was classified as a visual–spatial or reference component. Because even a small table (with a single molar mass of a substance) was included in this broad definition, the items that contained balanced equations or tables of data were separated from items containing a graph or figure. This is also shown in Figure 6, where the percentage of items still favored female students, although the gap is reduced.

The lack of a trend within specific content areas was expected given the broad categories within these content areas. However, when examining the items on all trial tests more closely, some interesting groupings emerged.

Because ACS Exams are secure exams, we cannot show the items or give any more specific information beyond the specific content area that was tested. However, in order to offer two examples of the items described here as well as the differences in format, two examples of trial test items that are not included in the released exams are provided in Figures 7 and 8.

A sample of chlorine gas at 298 K and 1.00 atm has a density of 2.90 g·L^{-1}. If the pressure of the gas is tripled at constant temperature, the density will

(A) stay the same.
(B) increase by a factor of three.
(C) decrease by a factor of three.
(D) increase by a factor of six.

Figure 7. Gas law item that favored male students and incorporated the format of reasoning.

The first grouping was within the content area of "general". This included laboratory measurement, density, classification of matter, and nomenclature items. Examining the nomenclature items, all five items that exhibit DIF favored female students from five different trial tests given to multiple institutions (Table 4), with a strong trend of the specific type of

What mass of SF_6(g) will exert a pressure of 2.00 atm in a 3.00 L container at 25.0 °C?

Molar mass / g·mol^{-1}	
SF_6	146.0

(A) 35.8 g
(B) 427 g
(C) 49.9 g
(D) 595 g

Figure 8. Gas law item that favored female students and incorporated the format of computation and a reference component.

nomenclature item. The format of all of these items was the same.

The second content area, stoichiometry, includes reactions, stoichiometry, and formula calculations. Within this content area, there were four items that exhibit DIF that all tested formula calculations (all the same type of calculation) and all four items favored male students. The format of these items was all the same as well (all computation items).

Within the content area of atomic structure and periodicity, there were five items that exhibited DIF; all five tested the same aspect of isotopes and favored male students. The format of these items varied, with one containing a visual–spatial or reference component and all items involving a computation or reasoning component. Still within the content area of atomic structure and periodicity, there were also three items that tested electron configurations or electronic structure: all of these items favored female students.

The content area of states of matter included the topics of gases, liquids, and solids. Of the items that specifically tested students on gases, 8 items exhibited DIF and not all of these favored either all female or all male students. However, when examining the format of the items, the 2 items that favored male students involved a reasoning component or solving a gas law problem without a calculation (one example is shown in Figure 7). The remaining items favored female students, and of the gas law problems, these involved a calculation (one example is shown in Figure 8). Items testing states of matter (changes of states) and classification of matter that exhibited DIF favored male students.

Of the items within thermodynamics that focused on thermochemistry, 7 items exhibited DIF. Of these, those involving a calculation related to enthalpy or calorimetry favored male students, while calculations based on Hess's law favored female students. One additional definition item favored female students.

The final content area that had many items that exhibited DIF was molecules, compounds, and molecular structure. Within this content area, the items overwhelming favored female students, with observable trends based on specific content areas. This includes formal charges and molecular orbital theory.

■ CONCLUSION

Differential item functioning analysis of exam items has been routinely carried out on many tests, including ACS Exams Institute trial exams that have enough student performance data available for the analysis. This analysis, when presented to committees, can invoke strong responses. These responses can range from dismissal of the results ("this cannot be") to asking why the DIF was identified and conjecture about the reason behind the results. However, these results must be considered in the context of how trial tests are given, specifically as an isolated testing event of a high-stakes test, with the item

dx.doi.org/10.1021/ed4000298 | J. Chem. Educ. 2013, 90, 846–853

Table 4. Some DIF Items from All Trial Tests by Content Area, Specific Content Area, Format of Item, Direction of DIF, and Number of Items

Content Area[a] of Trial Tests	Specific Content Area	Direction of DIF[b]	Format(s)[c]	Number of Items
General	Nomenclature	F	SCK	5
Stoichiometry	Formula calculations	M	C	4
Atomic structure	Isotopes	M	R, VS, and/or C	5
Atomic structure	Electron configuration	F	SCK and/or R	2
Atomic structure	Electron structure	F	SCK and VS	1
States of matter	Gas laws	M	R	2
States of matter	Gas laws	F	C and/or VS	3
States of matter	Gas stoichiometry	F	C and/or VS	1
States of matter	Ideal behavior of gases	F	SCK	1
States of matter	Kinetic molecular theory	F	R and VS	1
States of matter	States of matter	M	SCK or R	2
General	Classification of matter	M	SCK and VS	1
Thermodynamics	Calorimetry	M	C	1
Thermodynamics	Enthalpy	M	VS and C or SCK and R	3
Thermodynamics	Definition	F	SCK and VS	1
Thermodynamics	Hess's law	F	C and VS	2
Molecular structure	Formal charge	F	VS and SCK or VS and C	2
Molecular structure	Lattice energy	F	R	1
Molecular structure	Lewis dot structure	M	C	1
Molecular structure	Bonding	F	R	1
Molecular structure	Hybridization	F	SCK and VS	1
Molecular structure	Molecular orbital theory	F	SCK or VS and C	2
Molecular structure	Polarity	M	R and VS	1
Molecular structure	Shape	F	R	1
Molecular structure	Shape	M	R	1

[a]Trial tests were GC10FA, GC10FB, GC12FA, GC12FB, GC11A, and GC13B. [b]Differential item functioning (DIF) by sex: male (M) or female (F). [c]One or more items contain a single or multiple format using the coding of SCK, specific chemical knowledge; R, reasoning; VS, visual–spatial or reference component; C, computation.

responses submitted to the ACS-EI for analysis. This method does not allow access to multiple testing of these items (given that it is a single testing event of a trial test) nor do participants in trial testing provide any other relevant measures of proficiency for each student (a necessary component of a DIF analysis) to conduct a more thorough examination. Therefore, the results shown here can only represent a beginning of an investigation of DIF on general chemistry exams rather than a statement of broad classes or types of items to avoid including on an exam to minimize DIF. On occasion, an item shows the highest level of statistical significance for DIF, and exam committees are told they cannot use those items in the released version of the exam because they are less likely to be the result of small-sample-size-based fluctuations.

However, when a grouping of items that test the same specific content area using the same format of the item favor only one subgroup, these items are worthy of discussion. For example, the grouping of nomenclature items that favored female students includes transition metal ions in four out of the five items. The grouping of atomic structure items that favored male students includes a conceptual question about isotopes. The grouping of formula calculation items that favored male students includes a conceptual question about numbers of atoms. Finally, the grouping within gases splits between favoring female or male students; however, the items that incorporate a reasoning component to solve a gas law question favored male students, while similar computational items favored female students.

In addition to the constraints of how trial tests are administered, there is no way to extricate the format of the item from the content of the item. One may suppose that asking a question about gas laws without numbers may favor male students, while one with numbers may favor female students, but without further testing, this conclusion is premature. As with many testing analyses, the results shown here present more of an opportunity to investigate whether a specific content area or the inclusion of a specific format in a test item promotes the probability of DIF; these studies are already underway and will be reported separately. Nonetheless, the results of DIF analyses, when conducted as part of an ACS exam development process, will continue to be included in the suite of results given to testing-writing committees. The most likely action taken will be to remove possible DIF items from production exams to increase the probability of producing the highest-quality exams. In some cases, a balance can be struck by including an item that possibly favors female students with an item that possibly favors male students. It is perhaps likely that items flagged as potential DIF items will include random fluctuations associated with modest sample sizes of student performance data for trial tests, yet given the ability to choose items that do not show possible DIF, it is prudent to make such a choice to avoid possible item bias. Ultimately, a test most likely cannot be designed that entirely avoids the existence of possible DIF via random fluctuation or otherwise. Therefore, knowing more about methods to identify DIF and consider trends related to DIF items contributes to constructing tests that minimize DIF and improve the quality of the measurement and associated judgment from the test.

dx.doi.org/10.1021/ed4000298 | J. Chem. Educ. 2013, 90, 846–853

■ ASSOCIATED CONTENT

Supporting Information

Tables of a breakdown by number of items per content area or format of item per trial test. This material is available via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: kmurphy@uwm.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to thank April Zenisky for sharing her knowledge and expertise on DIF and Anja Blecking for serving as an expert rater.

■ REFERENCES

(1) Osterlind, S. J.; Everson, H. T. *Differential Item Functioning*, 2nd ed.; Quantitative Applications in the Social Sciences, 161; Sage Publications: Thousand Oaks, CA, 2009.

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

(3) Maccoby, E. E.; Jacklin, C. N. *The Psychology of Sex Differences*; Stanford University: Stanford, CA, 1974.

(4) Cole, N. S. *The ETS Gender Study: How Females and Males Perform in Educational Settings*; Educational Testing Service: Princeton, NJ, 1997.

(5) Halpern, D. F. *Am. Psychol.* **1997**, *52*, 1091−1102.

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

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

(8) Holland, P. W.; Thayer, D. T. Differential Item Functioning and the Mantel−Haenszel Procedure. In *Test Validity*, Wainer, H., Braun, H. I., Eds.; Lawrence Erlbaum: Hillsdale, NJ, 1988; pp 129−145.

(9) Swaminathan, H.; Rogers, H. J. *J. Educ. Meas.* **1990**, *27*, 361−370.

(10) Hambleton, R. K.; Swaminathan, H. *Item Response Theory: Principles and Applications*; Kluwer Nijhoff Publishing: Boston, MA, 1985.

(11) Clauser, B.; Mazor, K.; Hambleton, R. K. *Appl. Meas. Educ.* **1993**, *6*, 269−279.

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

(13) Zenisky, A. L.; Hambleton, R. K.; Robin, F. *Educ. Assess.* **2003**, *9*, 61−78.

(14) Walker, C. M.; Zhang, B.; Surber, J. *Appl. Meas. Educ.* **2008**, *21*, 162−181.

(15) Walker, C. M.; Beretvas, S. N. *J. Educ. Meas.* **2001**, *38*, 147−163.

(16) Mendes-Barnett, S.; Ercikan, K. *Appl. Meas. Educ.* **2006**, *19*, 289−304.

(17) Ryan, K. E.; Chiu, S. *Appl. Meas. Educ.* **2001**, *14*, 73−90.

(18) Holweger, N.; Taylor, G. Differential Item Functioning by Gender on a Large-Scale Science Performance Assessment: A Comparison across Grade Levels. In *Education Resources Information Center*; 1998; pp 1−31 (ED423282), http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED423282 (accessed April 2013).

(19) Hamilton, L. S.; Snow, R. E. Exploring Differential Item Functioning on Science Achievement Tests. *Education Resources Information Center*; 1998; pp 1−44 (ED427077), http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED427077 (accessed April 2013).

(20) Hamilton, L. S. *Appl. Meas. Educ.* **1999**, *12*, 211−235.

(21) Miller, T. R.; Spray, J. A. *J. Educ. Meas.* **1993**, *30*, 107−122.

(22) Hamilton, L. S.; Nussbaum, E. M.; Snow, R. E. *Appl. Meas. Educ.* **1997**, *10*, 181−200.

(23) Nussbaum, E. M.; Hamilton, L. S.; Snow, R. E. *Am. Educ. Res. J.* **1997**, *34*, 151−173.

(24) Tobin, K. G.; Capie, W. *Educ. Psychol. Meas.* **1981**, *41*, 413−423.

(25) Roadrangka, V.; Yeany, R. H.; Padilla, M. J. *The Construction and Validation of Group Assessment of Logical Thinking (GALT)*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Dallas, TX, April 1983.

(26) Jiang, B.; Xu, X.; Garcia, A.; Lewis, J. E. *J. Chem. Educ.* **2011**, *87*, 1430−1437.

(27) Crawford, M.; Marecek. *J. Psychol. Women Q.* **1989**, *13*, 147−166.

(28) Weisstein, N. *Kinder, Küche, Kirche, as Scientific Law: Psychology Constructs the Female*; New England Free Press: Boston, MA, 1968.

(29) Baby Name Guesser Web site. http://www.gpeters.com/names/baby-names.php? (accessed April 2013).

(30) Fidalgo, A. A.; Ferreres, D.; Muniz, J. *Educ. Psych. Measure.* **2004**, *64*, 925−936.

(31) Woolson, R. F.; Bean, J. A.; Rojas, P. B. *Biometrics* **1986**, *42*, 927−932.

(32) Crocker, L.; Algira, J. *Introduction to Classical and Modern Test Theory*; Holt, Reinhart and Wilson: New York, 1986.