JOURNAL OF CHEMICAL EDUCATION

Article

pubs.acs.org/jchemeduc

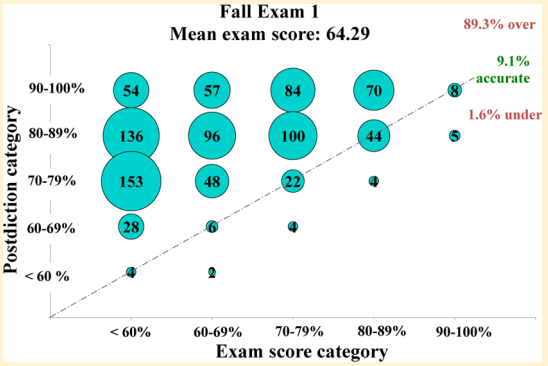
Investigating General Chemistry Students' Metacognitive Monitoring of Their Exam Performance by Measuring Postdiction Accuracies over Time

Morgan J. Hawker, Lisa Dysleski,[†] and Dawn Rickey*[,‡]

Department of Chemistry, Colorado State University, Fort Collins, Colorado 80523, United States

Supporting Information

ABSTRACT: Metacognitive monitoring of one's own understanding plays a key role in learning. An aspect of metacognitive monitoring can be measured by comparing a student's prediction or postdiction of performance (a judgment made before or after completing the relevant task) with the student's actual performance. In this study, we investigated students' postdiction accuracies for a series of exams within a two-semester general chemistry course. The research questions addressed include (1) How accurate are general chemistry students at postdicting their exam scores? Are there gender differences in postdiction accuracy? (2) How do general chemistry students' postdiction accuracies relate to their exam performance? (3) How do general chemistry students' postdiction accuracies and metacognitive monitoring of their exam performance change over time? Results indicate that most general chemistry students are not accurate in their exam score postdictions and that, consistent with research conducted in other domains, higher-performing students make more accurate postdictions than lower-performing students. In addition, although students who were new to a general chemistry course appeared to improve in their metacognitive monitoring on the second course exam compared with the first, monitoring did not significantly improve after that initial adjustment. Given the importance of metacognitive monitoring for student learning of chemistry, these findings suggest that further research and development of interventions to improve the metacognitive monitoring of introductory chemistry students is warranted.



KEYWORDS: *First-Year Undergraduate/General, Chemical Education Research, Testing/Assessment, Student-Centered Learning, Learning Theories*

FEATURE: Chemical Education Research

INTRODUCTION

Metacognitive skills, which involve thinking about one's own thinking, are increasingly being recognized for their important role in learning.[1,2] For example, in one study, metacognitive skillfulness was estimated to account for about 40% of the variance in learning outcomes.[3] Thus, to improve student learning in a specific content area such as chemistry, it is critical to understand both students' metacognitive skillfulness and instructional methods that could be used to improve it in that context.

Monitoring is a component of metacognition that involves assessing one's current knowledge, understanding, and abilities, as well as the task at hand and its difficulty.[1,4] One measure of metacognitive monitoring, often employed in cognitive psychology research, is the accuracy of an individual's predictions or postdictions of performance,[5−10] where a postdiction is a judgment made after completing a task. An individual's accuracy can be expressed as the difference between his or her prediction or postdiction and his or her actual performance on a task. This is also referred to as *calibration*.[11] In this study, we examine the accuracy of general chemistry students' exam performance postdictions.

Measuring metacognitive monitoring with pre- and postdictions of task performance has been approached in a variety of ways. Judgments can be made on performance in absolute terms as described above or relative to others' performance.[12] Generally speaking, people's abilities to pre- or postdict their performance on a task is found to be poor, although pre- and postdiction ability is related to task difficulty.[13]

Undergraduate students' metacognitive skillfulness related to academic tasks has also been investigated. Studies that employ pre- or postdictions of absolute exam performance have been carried out in undergraduate educational psychology,[7,10,14,15] education,[5] cognitive psychology,[8] developmental psychology,[16] introductory physics,[17] and organic chemistry courses;[18] high-school biology[19,20] and psychology classrooms;[21] and upper-elementary classrooms.[22] Most of these studies conclude that

Received: August 27, 2015
Revised: February 12, 2016

DOI: 10.1021/acs.jchemed.5b00705
J. Chem. Educ. XXXX, XXX, XXX−XXX

many students are well-calibrated when it comes to making postdictive judgments of exam performance.[10,14,19,20] A notable exception is the study carried out in organic chemistry, in which the majority of students overestimated their exam performance.[18] All of the studies that explored the relationship between performance and calibration found that higher-performing students were better calibrated on exam postdiction judgments than lower-performing students.[5,10,14,15] Studies that included analyses to determine changes in students' calibration accuracy over time indicate that students only became more accurate in some cases where there was an explicit intervention designed to improve students' metacognitive monitoring throughout the course.[5,7,8,10,15]

Although we did not find previous studies that explored gender differences in accuracy of students' course exam postdictions, some previous work has explored differences in male and female postdiction accuracies on informal knowledge tests. In particular, Beyer[23] and Beyer & Bowden[24] found that, while there were no significant differences between male and female undergraduate students in postdiction accuracy on "neutral" tasks (e.g., tests of common knowledge) or "feminine-gender-typed" tasks (e.g., trivia from movies and TV shows with primarily female audiences), females significantly underestimated their performance and exhibited poorer calibration relative to males on "masculine-gender-typed" tasks (e.g., football, basketball, and baseball trivia). In studies that asked undergraduate students to rate their performance relative to that of other students, Kruger and Dunning did not find any gender differences in postdiction accuracy on tests of humor, logical reasoning, or English grammar.[12]

In the context of general chemistry courses, monitoring research has primarily focused on students' abilities to judge their performance on problems outside of exam situations. For example, in two studies, investigators asked students to report their confidence in their abilities to solve particular chemistry problems, but the students did not actually work the problems presented to them.[25,26] Another study examined the relationship between students' confidence judgments regarding individual stoichiometry questions and their performance on these questions.[27] In one study that did investigate students' judgments of performance on general chemistry exams, students were asked to judge how they performed on exams relative to the average student in the class (without any knowledge of the class average or other students' exam scores).[28] Such relative judgments are fundamentally different tasks compared with the absolute judgments of individual performance examined here.

In this study, we examine general chemistry students' abilities to monitor their examination performance. We investigate students' judgments of their own performance on exams within a two-semester general chemistry course. Our research questions are as follows: (1) How accurate are general chemistry students at postdicting their exam scores? Are there gender differences in postdiction accuracy? (2) How do general chemistry students' postdiction accuracies relate to their exam performance? (3) How do general chemistry students' postdiction accuracies and metacognitive monitoring of their exam performance change over time?

METHODS

We collected data from students enrolled in a two-semester general chemistry sequence (General Chemistry I and General Chemistry II) for science and engineering majors at a large, public university. Although each course included multiple sections and instructors (Table 1), common exams were administered to all students. The instructors for the spring General Chemistry II course were a subset of the instructors for the fall General Chemistry I course. For each semester-long course, five multiple-choice exams were administered at three-week intervals, including one comprehensive final exam. The first four exams were not comprehensive. Example exam questions are shown in Box 1. Students in both fall and spring semesters had the option to use their final exam percentage as a replacement for one previous exam percentage in the calculation of their final course grade.

Table 1. Summary of Course Characteristics

Semester	Sections	Instructors	Students Enrolled[a]	Students in Study
Fall	5	4	1075	925
Spring	3	2	696	491

[a]Total enrollment at the end of each course.

Box 1. Sample questions from first- and second-semester general chemistry exams.

First Semester

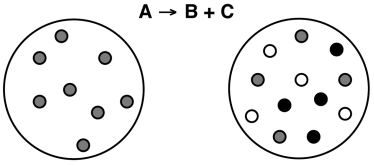
D and E are elements whose identities are unknown. Given that a neutral atom of D has a larger atomic radius than a neutral atom of E, and that the relationships between D and E are consistent with the periodic trends discussed in class, which must be true?

a. Element D is closer to the bottom of the periodic table than element E.
b. An atom of D has more electrons than an atom of E.
c. Element E has a higher first ionization energy than element D.
d. The most stable ion formed from D has a larger atomic radius than the most stable ion formed from E.
e. Both B and D must be true.

Second Semester

Substance A (grey) decomposes into two other substances, B (black) and C (white) according to a zero-order reaction. The molecular scenes below show a portion of the reaction mixture at two times. What is the average rate of disappearance of A during the interval 1 – 2 minutes?

$A \rightarrow B + C$



0 minutes 1 minute

(a) 2 min^{-1} (b) 4 min^{-1} (c) 8 min^{-1} (d) 0.25 min^{-1} (e) 12 min^{-1}

All exams included the postdiction question, *"What percentage score do you expect to earn on this exam? (A) 100%–90%, (B) 89%–80%, (C) 79%–70%, (D) 69%–60%, (E) < 60%."* This question appeared as the last or second-to-last question on each exam,[29] such that each student was asked to make a judgment of his or her own score on the exam immediately after taking the exam. We collected students' answers to all exam questions via optical mark recognition (i.e., Scantron) forms. Each student who answered a postdiction question received credit amounting to 1–2% of the total points for the exam regardless of whether his or her postdiction was accurate. No interventions intended to improve students' metacognitive monitoring of exam performance were implemented in these courses. Students who were missing a

DOI: 10.1021/acs.jchemed.5b00705
J. Chem. Educ. XXXX, XXX, XXX–XXX

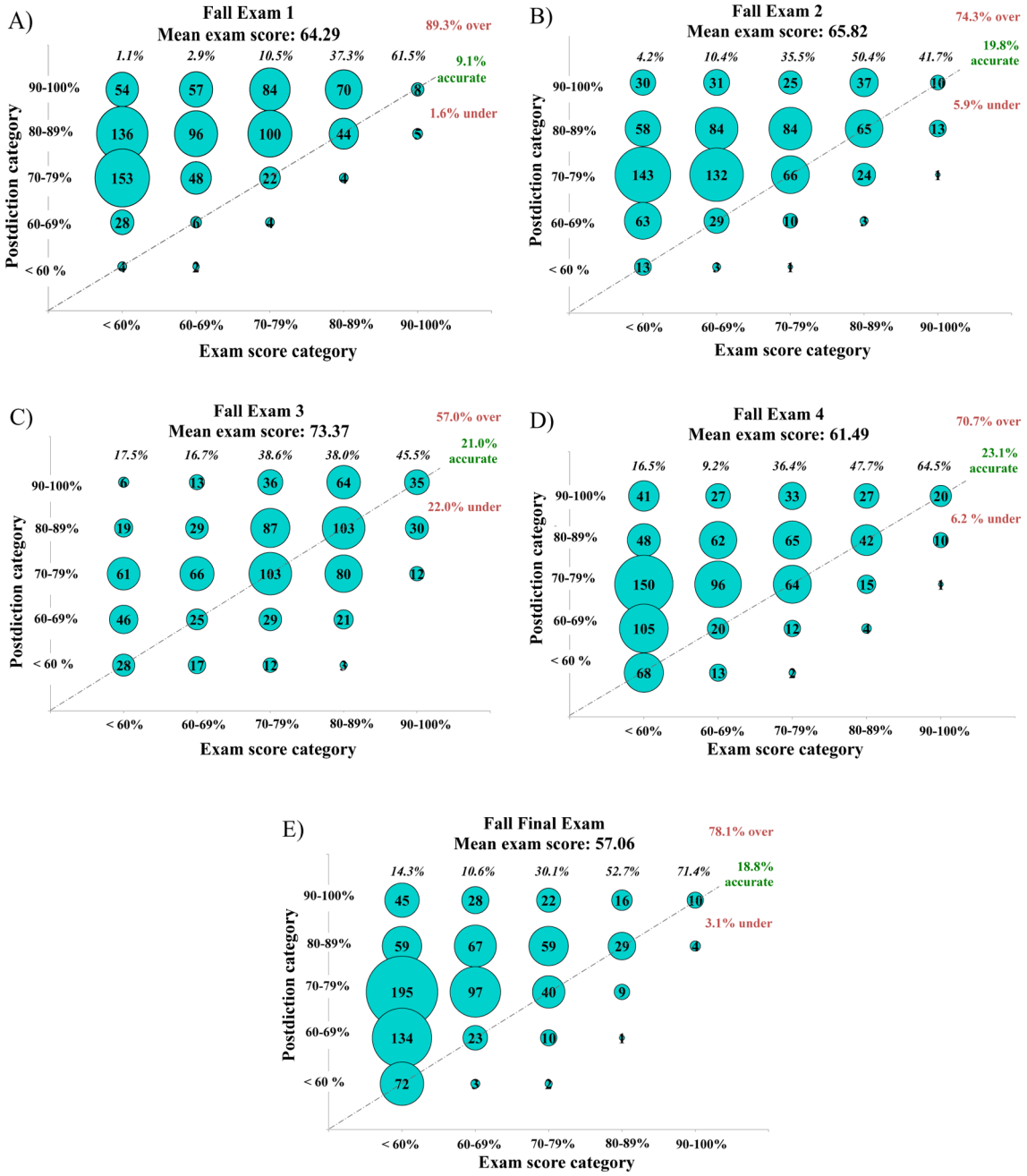


Figure 1. Bubble plots of student exam postdiction score category versus exam score category for each exam in the fall semester (N = 925). Bubble size is proportional to the number of students with a given exam score category and postdiction. In each case, the dashed diagonal line indicates perfect calibration. Overall percentages of overpostdicting, underpostdicting, and perfectly accurate students are included on the right side of each panel. The percentages of perfectly calibrated students in each exam score category are included at the top of the corresponding columns (italicized).

postdiction response or exam score for any exam were excluded from the study, resulting in the final cohorts indicated in Table 1. In addition, we performed separate analyses for the students who took both the General Chemistry I course in the fall and the General Chemistry II course the following spring (Fall and Spring, N = 343), and those who took the General Chemistry II course in the spring, but not the General Chemistry I course in the fall (Spring Only, N = 148). The Spring Only students took the General Chemistry I course in a different semester or at a different institution, or were exempt from the first-semester course because they had Advanced Placement credit or tested out of the course.

To conduct the calibration analyses, we coded each student's actual exam scores and exam postdictions using the categories that were available as answer choices to the postdiction questions. We converted each exam score category and corresponding postdiction to a numerical value on a 4-point scale as follows: 100%−90% = 4, 89%−80% = 3, 79%−70% = 2, 69%−60% = 1, <60% = 0. Finally, we determined the postdiction calibration for each student on each exam by calculating the difference between his or her postdiction category and his or her exam score category (eq 1).

$$\text{postdiction calibration} = \text{postdiction category} - \text{exam score category} \quad (1)$$

DOI: 10.1021/acs.jchemed.5b00705
J. Chem. Educ. XXXX, XXX, XXX−XXX

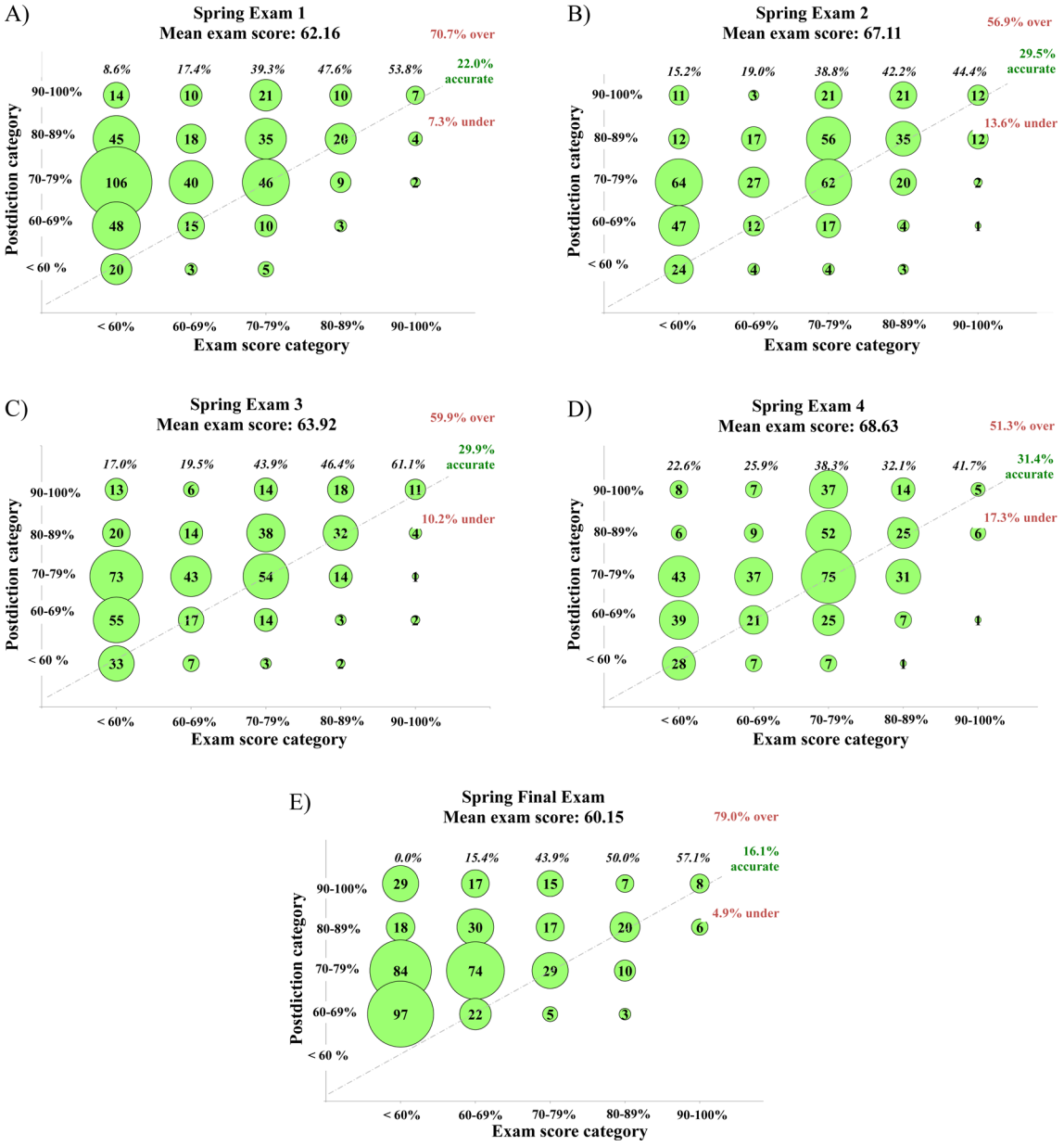


Figure 2. Bubble plots of student exam postdiction category versus exam score category for each exam in the spring semester (N = 491). Bubble size is proportional to the number of students with a given exam score category and postdiction. In each case, the dashed diagonal line indicates perfect calibration. Overall percentages of overpostdicting, underpostdicting, and perfectly accurate students are included on the right side of each panel. The percentages of perfectly calibrated students in each exam score category are included at the top of the corresponding columns (italicized).

A student with positive postdiction calibration postdicted that his or her exam score would be higher than it actually was (*overpostdicted*), whereas a student with negative postdiction calibration postdicted that his or her exam score would be lower than it actually was (*underpostdicted*). A student with postdiction calibration of zero accurately postdicted the score category into which his or her actual exam score would fall (*perfectly calibrated*). We used the absolute value of the postdiction calibration (|calibration|) for each student to determine the average magnitudes of calibration for groups of students. To characterize students' postdiction accuracies in the context of our research questions, we employed the complementary measures of |calibration| and the percentage of students who were perfectly calibrated (% accurate).

To examine relationships between calibration and exam performance, we also established two performance groups of students. One group of consistently "high-performing students," or those who earned exam scores greater than 1/2 standard deviation above the exam mean on every exam, and one group of consistently "low-performing students," or those who earned exam scores lower than 1/2 standard deviation below the exam mean on every exam. These performance groups did not encompass all students in the course, as many students did not consistently score within a single performance group for all exams in a given course.

For both the fall and spring semesters, we calculated descriptive statistics for student exam scores, postdictions, calibration, and the percentage of accurate postdictions for the student groups of interest. We used independent samples t tests to test for significant differences in exam score means of different student groups, except for a case in which the homogeneity of variance assumption was violated for which we

DOI: 10.1021/acs.jchemed.5b00705
J. Chem. Educ. XXXX, XXX, XXX−XXX

Table 2. Performance and Calibration Statistics for Fall and Spring Semesters

Exam	Mean exam score %,[a] *M*	Mean score category[a]	Mean postdiction category[a]	Mean |calibration|[a]	% accurate postdictions
Fall (*N* = 925)					
Exam 1	64.29 (15.25)	1.12 (1.12)	2.95 (0.87)	1.86 (1.03)	9.1
Exam 2	65.82 (13.86)	1.23 (1.13)	2.47 (0.94)	1.37 (1.01)	19.8
Exam 3	73.37 (13.70)	1.95 (1.22)	2.36 (1.10)	0.96 (0.84)	21.0
Exam 4	61.49 (15.54)	1.04 (1.15)	2.23 (1.16)	1.34 (1.05)	23.1
Final	57.06 (15.86)	0.76 (1.00)	2.15 (1.12)	1.45 (1.05)	18.8
Spring (*N* = 491)					
Exam 1	62.16 (13.82)	1.01 (1.14)	2.23 (1.04)	1.40 (1.04)	22.0
Exam 2	67.11 (15.88)	1.51 (1.25)	2.24 (1.10)	1.08 (0.94)	29.5
Exam 3	63.92 (15.45)	1.25 (1.22)	2.10 (1.13)	1.12 (0.99)	29.9
Exam 4	68.63 (12.26)	1.54 (1.10)	2.12 (1.14)	1.01 (0.89)	31.4
Final	60.15 (15.59)	0.92 (1.08)	2.24 (1.01)	1.43 (1.03)	16.1

[a]Values for one standard deviation are included in parentheses.

used nonparametric Mann−Whitney U tests (two-tailed). Additionally, we performed Mann−Whitney U tests to compare the |calibration| and exam score category distributions for different groups. We determined the appropriate effect sizes, with Cohen's *d* effect sizes corresponding to *t* tests and *r* effect sizes corresponding to Mann−Whitney U and Wilcoxon signed rank tests. The *r* effect sizes were calculated according to eq 2, where *Z* is the Mann−Whitney U or Wilcoxon signed rank test Z-score (*Z*) and *N* is the sample size.[30,31]

$$r = \frac{Z}{\sqrt{N}} \tag{2}$$

Effect sizes of 0.1, 0.3, and 0.5 are considered to be small, medium, and large, respectively.[32]

In addition, we utilized Fisher's exact to test for the significance of differences in percent accurate postdictions.[33] We applied the Bonferroni adjustment to correct for multiple comparisons in each semester, resulting in an α level of 0.01.[34] We performed all statistical analyses using SPSS version 20.0 software (SPSS, Inc., Chicago, IL).

RESULTS AND DISCUSSION

Postdiction Accuracy

Figures 1 and 2 present bubble plots of postdiction category versus exam score category for each exam in fall and spring, respectively. The size of each bubble corresponds to the number of students who fell into a specific combination of exam score category and postdiction category. Bubbles that fall on the dashed diagonal lines in the plots represent students who were perfectly calibrated.

In each case, a minority of students fall along the line of perfect calibration, with the percentage of students accurately postdicting their exam score category ranging from 9.1% to 31.4% (also see Table 2). Figure 1A shows that for Fall Exam 1, only 9.1% of students were perfectly calibrated, whereas the vast majority of students (89.3%) overpostdicted. In fact, more students made overpostdictions than under- or accurate postdictions for all exams analyzed.

The mean absolute calibration (|calibration|) for each exam (Table 2) indicates how close students' postdictions were to their actual performance categories on average. Throughout the two-semester general chemistry course, the accuracy of students' postdictions was low, averaging one to two exam score categories away from their actual performance. (Changes in postdiction accuracy over time are discussed later.) We also note that for the comprehensive final exams for both semesters, larger percentages of students overpostdicted, and mean |calibration| was less accurate than in all other cases except for Fall Exam 1. In addition to the more comprehensive nature of the final exams (which increased the difficulty), student postdictions may have been more optimistic than usual since many students hoped to obtain a higher score on their final exam to replace a previous low exam score. Finally, as discussed in more detail later, the lower the exam mean, the less accurate students' mean |calibration| tended to be.

We also compared percent accurate postdictions and mean |calibration| for male and female students on each exam. Although female students had consistently higher rates of percent accurate postdictions compared with male students, none of the differences are statistically significant (Supporting Information Table S1). With respect to mean |calibration|, female students were significantly more accurate on average than male students for three of the five fall-semester exams (exam 2, exam 4, and the final exam), with small effect sizes of about 0.1, while in the spring semester, there were no significant differences in mean |calibration| between male and female students (Supporting Information Table S1). There were no significant performance differences between males and females for any of the exams across the two-semester general chemistry sequence (Supporting Information Table S2).

These findings regarding general chemistry students' postdiction accuracy are consistent with the results of the previously mentioned exam calibration study conducted in organic chemistry courses, where about 60% of students overestimated their exam scores.[18] The exam calibration accuracy results from the studies in chemistry courses differ substantially from results of exam calibration studies conducted in undergraduate psychology courses,[7,8,10,14,16] in which students were generally found to be well-calibrated. Likely explanations include differences in the nature of the exams, the courses, and the student populations.[35] Differences in the nature of tasks are known to influence judgment accuracy,[35,36] and thus, it is likely that the differences between chemistry and psychology courses and exams influence students' abilities to make accurate exam postdictions. For example, questions on the general chemistry exams often required multistep problem solving that was quantitative in nature. This differs from example exam questions reported for the psychology course calibration studies, which consisted of questions that required recall and application of declarative knowledge.[10,14] In addition, some of the student populations studied in the psychology

DOI: 10.1021/acs.jchemed.5b00705
J. Chem. Educ. XXXX, XXX, XXX−XXX

Table 3. Comparison of Exam Scores of Students Who Were Well Calibrated and Those Who Were Not

Exam	Mean exam score[a]		U statistic[b]	Effect size (r)[c]
	Students with calibration = 0, ±1	Students with calibration = ±2, ±3, ±4		
Fall (N = 925)				
Exam 1	75.26 (12.89)	57.76 (12.58)	30226	0.59
Exam 2	71.51 (12.77)	57.51 (10.88)	39075	0.53
Exam 3	75.58 (12.74)	65.93 (12.23)	67522	0.30
Exam 4	65.31 (16.56)	55.70 (11.72)	64515	0.32
Final	60.49 (17.78)	52.91 (11.95)	76490	0.24
Spring (N = 491)				
Exam 1	67.07 (14.29)	56.32 (10.61)	17758	0.41
Exam 2	70.57 (15.29)	58.63 (14.03)	15620	0.36
Exam 3	67.20 (15.40)	56.54 (12.86)	13526	0.35
Exam 4	70.59 (11.56)	62.97 (12.51)	14372	0.26
Final	64.41 (16.78)	55.24 (12.08)	15208	0.27

[a]Values for one standard deviation are included in parentheses. [b]Mann−Whitney U tests compare exam score distributions; population information can be found in Supporting Information Table S3. [c]p-values < 0.0001.

courses were upper-level educational psychology students,[10,14] and the psychology courses in two studies emphasized the importance of making accurate judgments and how those judgments relate to metacognition throughout the courses.[10,15]

Relationships between Student Exam Postdiction Accuracy and Performance

In Figures 1 and 2, the percentages of perfectly calibrated students in each exam score category are included at the top of the corresponding columns. We observe that larger percentages of higher performing students tended to be perfectly accurate in their postdictions compared with lower performing students for every exam across the two semesters. To further explore the relationships between postdiction accuracy and exam performance, we compared the exam scores of students who were relatively well calibrated (calibration = 0, ±1) with those who were not well calibrated (calibration = ±2, ±3, ±4) (Table 3). (Supporting Information Table S3 shows the sample sizes for each group of students.) Mann−Whitney U tests indicate that the exam score distributions in the two groups are statistically different for every exam, with medium-to-large effect sizes for most exams (Table 3),[32] providing further evidence that students who were better calibrated tended to earn higher exam scores than those who were less accurately calibrated. In addition, we compared the distributions of |calibrations| of the consistently high-performing and low-performing student groups for each exam (Supporting Information Tables S4 and S5). The distributions differ significantly, with high performers achieving lower mean |calibration| than low performers for all exams, with medium-to-large effect sizes for most exams (Supporting Information Table S5). In the fall semester, the average mean |calibration| across all exams was 0.68 for high performers and 1.62 for low performers; in the spring semester, it was 0.65 for high performers and 1.45 for low performers (Supporting Information Table S4). These findings regarding postidiction accuracy and performance are consistent with previous exam calibration studies[5,10,14,15] and models of metacognition, which predict that students who are more proficient at monitoring their understanding while studying would make more accurate judgments regarding what they should focus on to enhance their understanding (exercising better metacognitive control), potentially leading to better exam performance. Being better prepared for exams would also allow students to judge their exam performance more accurately.

Changes in Postdiction Accuracy and Metacognitive Monitoring Over Time

We also explored changes in students' exam postdiction accuracies and whether students' metacognitive monitoring changed over time. Previous studies have examined changes in postdiction accuracy across one-semester education and psychology courses.[5,7,8,10,15] Hacker et al. compared the R^2 values for the best fit regression lines for actual score versus postdicted score data for each exam, and asserted that an increase in R^2 values over time indicated improved postdiction accuracy over time.[5] However, we concur with the critique by Nietfeld et al. that "...accounting for increasing amounts of variance does not necessarily mean that the relationship is in the expected direction" (p 23).[14] Thus, it is unclear that the results of the Hacker et al. study regarding changes in postdiction accuracy over time are valid. Other studies compared students' mean |calibration| for exams or quizzes over time for one-semester psychology courses to examine changes in postdiction accuracy.[5,7,8,15]

For our study, in examining trends in mean |calibration| over time (Table 2), we observed that when exam means increase, mean |calibration| improves (values decrease). Since a large proportion of students overpostdicted on each exam, higher exam means may be associated with more accurate postdictions regardless of whether students are more accurately monitoring their performance. While Nietfeld et. al noted similar changes in postdiction accuracy with changing exam difficulty, the range of exam means (76−81%) was narrower compared the range in our study (57−73%), and they did not adjust for changes in exam means over time.[14]

Specifically focusing on the pairs of exams for which mean |calibration| improved from one exam to the next (Table 2), we developed a method to determine whether it was likely that students' improvements in mean |calibration| were due to increases in exam means. In Table 4, we compare each exam with the one immediately following it. For each exam pair, we calculated the effect sizes (eq 2) for the change in |calibration| ($r_{|\text{calibration}|}$) and the change in exam score category (r_{exam}). For the cases in which students' mean |calibration| improved from one exam to the next, we calculated the ratio $r_{|\text{calibration}|}/r_{\text{exam}}$. If the magnitude of the effect size of students' decrease in

DOI: 10.1021/acs.jchemed.5b00705
J. Chem. Educ. XXXX, XXX, XXX−XXX

Table 4. Effect Sizes (r) for Changes in Exam Score Category (r_{exam}) and |Calibration| ($r_{|calibration|}$)

Exam pair	r_{exam}	$r_{\|calibration\|}$	$r_{\|calibration\|}/r_{exam}$
Fall (N = 925)			
Exam 1/Exam 2	0.11	0.38	3.41
Exam 2/Exam 3	0.53	0.34	0.64
Exam 3/Exam 4	0.72	0.32	—[a]
Exam 4/Final	0.36	0.11	—[a]
Spring (N = 491)			
Exam 1/Exam 2	0.35	0.27	0.78
Exam 2/Exam 3	0.25	0.04	—[a]
Exam 3/Exam 4	0.33	0.10	0.31
Exam 4/Final	0.57	0.34	—[a]

[a]Only calculated for pairs of exams where mean |calibration| improved.

|calibration| is larger than the magnitude of the effect size of the increase in mean exam score category, then the ratio $r_{|calibration|}/r_{exam}$ would be greater than 1, indicating that the improvement in students' calibration accuracy may not be fully explained by a higher exam mean.

As seen in Table 4, the only pair of consecutive exams for which students' mean |calibration| improves and $r_{|calibration|}/r_{exam}$ is greater than 1 is for Fall Exams 1 and 2, with $r_{|calibration|}/r_{exam}$ = 3.41. This suggests that students may have improved their metacognitive monitoring of their exam performance on Fall Exam 2 relative to Exam 1. For the other three pairs of exams for which students' mean |calibration| improved over time, the $r_{|calibration|}/r_{exam}$ ratios are less than 1, indicating that the effect sizes for the change in mean exam score categories increased to a greater extent than the corresponding effect sizes of the change in mean |calibration|. This suggests that students' metacognitive monitoring of their performance may not have improved across those exams. For the four other pairs of exams, exam means decreased from one exam to the next and students' mean calibration accuracy decreased, but with a smaller effect size than the mean exam score category decrease in each case.

Thus, the only time for which this analysis indicates that students' metacognitive monitoring of their exam performance may have improved is for Fall Exam 2 relative to Exam 1. Of course, Fall Exam 1 is unique in that it is the first exam of the general chemistry course, and also the first college-level chemistry exam for many of the students. Therefore, Exam 1 provided students with a new, particularly relevant experience to consider in their future judgments of their general chemistry exam performance. In addition, students' mean |calibration| on Exam 1 of 1.86 (an average of almost two exam score categories away from their actual exam score categories) was especially poor. Therefore, students may have adjusted their postdictions between Fall Exams 1 and 2 more than between any other pair of exams in part due to the experience of being miscalibrated by such a large margin for their Exam 1 performance. Previous studies in the psychology laboratory indicate that relevant experience alters the factors people use to make judgments of difficulty.[37,38]

If it was the case in the fall that students' postdiction accuracy improved in part because experience with the first general chemistry exam informed their future judgments of exam performance, then we would expect to see a similar pattern in the spring for the students who did not take the corresponding first-semester general chemistry course at the same institution in the fall (Spring Only, N = 148) to a greater extent than for the students who had the experience of the preceding fall course (Fall and Spring, N = 343). Spring Only students took a first-semester general chemistry course at another institution, took the course in a different semester, or tested out of the course, and therefore, while these students had taken general chemistry exams before, most had no previous experience with the spring instructors' course and assessment styles or with making exam postdictions. As expected, Spring Only students' mean |calibration| improved from Spring Exam 1 to Spring Exam 2 with $r_{|calibration|}/r_{exam}$ = 1.16, while $r_{|calibration|}/r_{exam}$ = 0.69 for the students who had completed the fall course (Supporting Information Tables S7–S10). The Spring Only students' improvement in postdiction accuracy that may be attributed to enhanced metacognitive monitoring is also smaller than that described earlier for all students in the fall cohort ($r_{|calibration|}/r_{exam}$ = 3.41) for which many students were new to college-level chemistry.

We also conducted these analyses for high- and low-performing student groups (Supporting Information Table S6) and for students who took both semesters of general chemistry consecutively (Supporting Information Table S10). In terms of changes in postdiction accuracy and metacognitive monitoring over time, the patterns observed for these subgroups are similar to those of the larger groups. Results indicate that both high- and low-performing students improved in postdiction accuracy and monitoring of their exam performance for Fall Exam 2 relative to Exam 1.

Unlike mean |calibration|, which varied inversely with exam mean, the percentage of students who were perfectly calibrated increased across each semester's exams until the final exam (Table 2). The only statistically significant pairwise increases in percent accurate students from one exam to the next (Fisher's exact $p < 0.01$) were from the first exam to the second exam in each semester.

Adjusting for changes in exam difficulty over time, our findings are generally consistent with previous work indicating that—without an intervention intended to improve student monitoring—students' calibration accuracy did not change much across multiple exams.[5,7,8,10,15] However, it appears that the students in our study, including both high and low performers, may have improved in metacognitive monitoring of their performance on Exam 2 relative to Exam 1 in general chemistry courses that were new to them.

LIMITATIONS

A limitation of this study was that student postdictions were collected using optical mark recognition (i.e., Scantron) forms with only five possible answer choices rather than an open-ended format. This method was chosen to facilitate data collection from larger numbers of students than have been included in previous postdiction calibration studies (more than 1000 students compared with less than 100 students in previous studies), but it also reduced the sensitivity for detecting differences in students' postdictions as well as the precision of the calibration results relative to an open-ended postdiction format. In addition, the postdiction categories we chose were not equal in range. In particular, the <60% postdiction answer choice spanned a 60-point range, whereas the additional four categories spanned only 10-point ranges. Given that the exam means ranged from 57–73%, it may have been preferable to select consistently wider ranges for the postdiction answer choices (e.g., a 20-point range for each). Of course, this would have also had the effect of making it easier

DOI: 10.1021/acs.jchemed.5b00705
J. Chem. Educ. XXXX, XXX, XXX−XXX

for higher-performing students to be better calibrated. Even though our selection of postdiction answer choices in this study theoretically made it easier for lower-performing students to be better calibrated, results illustrated that they were nevertheless significantly less accurate than higher-performing students. Finally, to the extent that postdictions are random guesses not informed by other information, people who perform closer to the middle of a performance scale have a better chance of more accurate |calibration| than those who perform closer to the extremes of the scale.

CONCLUSIONS AND IMPLICATIONS

In this study, we explored aspects of first- and second-semester general chemistry students' metacognitive monitoring of their exam performance by measuring their postdiction accuracies over time. In addition, we determined how postdiction accuracy relates to exam performance. First, we found that a large proportion of students in both semesters of general chemistry were miscalibrated in that they consistently overpostdicted their exam scores. Considering Exams 2−4, during which students' postdiction accuracies were the most accurate and most stable, the average mean |calibrations| were 1.2 in the fall and 1.1 in the spring, which indicates an average miscalibration of more than one exam score category. The extent to which students were miscalibrated is particularly striking because, unlike predictions of performance made without knowledge of the test questions, students made their postdictions immediately after completing each exam while the exam was still in their possession. Our results differ from previous findings in the context of psychology course exam postdiction accuracies, where students were typically found to be well calibrated, but are consistent with a study carried out in organic chemistry courses.[18] We attribute this to course and exam characteristics that are more similar for general and organic chemistry, but differ substantially between chemistry and psychology courses.

Second, we found that general chemistry students who earned higher exam scores also tended to be more accurately calibrated, which is consistent with the findings from previous studies in other courses as well as models of metacognition.[5,10,14,15]

Finally, although we observed improvements in students' postdiction accuracy between some pairs of exams, in all cases the improvements in calibration were seen for cases where the exam mean also increased. Thus, we realized that this could be due to students' tendency to overpostdict as opposed to improvements in their metacognitive monitoring of their exam performance. We developed a method to determine whether improvements in mean |calibration| from one exam to the next were likely due to improvements in students' metacognitive monitoring. Results indicated that, although students who were new to a general chemistry course appeared to improve in their metacognitive monitoring on the second course exam compared with the first, monitoring did not significantly improve after that initial adjustment. Thus, our results are generally consistent with postdiction studies in other domains in which monitoring did not change much without a specific intervention targeted at improving students' monitoring of their exam performance.[5,7,8,10,15] Notably, Bol et al. employed an intervention in which students practiced making pre- and postdictions on quizzes and no significant differences were observed in calibration accuracy on the final exam between intervention and control groups.[5] Interventions that have led to improvements in monitoring exam performance have included offering extra credit for more accurate judgments,[7] and having students complete self-reflection questionnaires that included topics such as how well concepts were understood, identifying strengths/weaknesses in content knowledge, and confidence ratings regarding ability to answer content questions.[19]

Overall, our results show that general chemistry students' perceptions of their own performance do not typically match their actual performance, especially for lower-performing students, and that in the absence of any intervention to improve monitoring, student monitoring of exam performance does not improve much across a year-long general chemistry course. Given the importance of metacognitive monitoring for student learning of chemistry, these findings suggest that further research and development of interventions to improve the metacognitive monitoring of introductory chemistry students is warranted. Increasing chemistry instructors' awareness of both the importance of metacognitive monitoring and the possibility that their students are miscalibrated may encourage them to test their students' calibration accuracy, and design and implement interventions intended to improve student monitoring. Another direction for future research in this area is the exploration of the strategies and reasoning chemistry students use in making judgments of their performance.

IRB/Human Subjects Statement

This study was approved by the IRB at Colorado State University.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available on the ACS Publications website at DOI: 10.1021/acs.jchemed.5b00705.

Data analyzed by gender, student performance groups, students who took both fall and spring courses, and students who only took the spring course (PDF, DOCX)

AUTHOR INFORMATION

Corresponding Author

*E-mail: drickey@nsf.gov.

Present Addresses

†Lisa Dysleski's current affiliation is the College of Natural Sciences, Colorado State University, Fort Collins, CO.

‡Dawn Rickey's current affiliation is the National Science Foundation, Arlington, VA.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation (NSF) under Award Number 0942448. In addition, portions of this manuscript were written while author Dawn Rickey was serving at the NSF, and include NSF support through her Independent Research and Development plan. Findings, conclusions, and recommendations expressed herein are those of the authors, and do not necessarily reflect the views of the NSF. The authors also acknowledge and thank the general chemistry students and instructors who participated in the study, as well as Ellen Fisher, Matthew Rhodes, Melonie Teichert, and anonymous reviewers for insightful comments on drafts of the manuscript.

DOI: 10.1021/acs.jchemed.5b00705
J. Chem. Educ. XXXX, XXX, XXX−XXX

REFERENCES

(1) Flavell, J. H.; Miller, P. H.; Miller, S. A. *Cognitive development*, 4 ed.; Prentice-Hall: Englewood Cliffs, NJ, 2002.

(2) Zohar, A.; Barzilai, S. A review of research on metacognition in science education: current and future directions. *Studies in Science Education* **2013**, *49* (2), 121−169.

(3) Veenman, M. V. Giftedness: Predicting the speed of expertise acquisition by intellectual ability and metacognitive skillfulness of novices. In *Meta-cognition: A Recent Review of Research, Theory, And Perspectives*; Nova Science Publishers, Inc., 2008; pp 207−220.

(4) Nelson, T. O.; Narens, L. Metamemory: A theoretical framework and new findings. *Psychol. Learn. Motiv.* **1990**, *26*, 125−141.

(5) Bol, L.; Hacker, D. J.; O'Shea, P.; Allen, D. The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *J. Exp. Educ.* **2005**, *73* (4), 269−290.

(6) Dunlosky, J.; Metcalfe, J. *Metacognition*; Sage Publications: 2008.

(7) Hacker, D. J.; Bol, L.; Bahbahani, K. Explaining calibration accuracy in classroom contexts: the effects of incentives, reflection, and explanatory style. *Metacognition Learn.* **2008**, *3* (2), 101−121.

(8) Miller, T. M.; Geraci, L. Training metacognition in the classroom: the influence of incentives and feedback on exam predictions. *Metacognition Learn.* **2011**, *6* (3), 303−314.

(9) Pieschl, S. Metacognitive calibration—an extended conceptualization and potential applications. *Metacognition Learn.* **2009**, *4* (1), 3−31.

(10) Hacker, D. J.; Bol, L.; Horgan, D. D.; Rakow, E. A. Test prediction and performance in a classroom context. *J. Educ. Psychol.* **2000**, *92* (1), 160−170.

(11) Lin, L. M.; Zabrucky, K. M. Calibration of comprehension: Research and implications for education and instruction. *Contemp. Educ. Psychol.* **1998**, *23* (4), 345−391.

(12) Kruger, J.; Dunning, D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self−assessments. *J. Pers. Soc. Psychol.* **1999**, *77* (6), 1121−1134.

(13) Burson, K. A.; Larrick, R. P.; Klayman, J. Skilled or unskilled, but still unaware of it: how perceptions of difficulty drive miscalibration in relative comparisons. *J. Pers. Soc. Psychol.* **2006**, *90* (1), 60.

(14) Nietfeld, J. L.; Cao, L.; Osborne, J. W. Metacognitive monitoring accuracy and student performance in the postsecondary classroom. *J. Exp. Educ.* **2005**, 7−28.

(15) Nietfeld, J. L.; Cao, L.; Osborne, J. W. The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self−efficacy. *Metacognition Learn.* **2006**, *1* (2), 159−179.

(16) de Carvalho Filho, M. K. Confidence judgments in real classroom settings: Monitoring performance in different types of tests. *Int. J. Psychol.* **2009**, *44* (2), 93−108.

(17) Galloway, R. K.; Bates, S. P.; Parker, J.; Usoskina, E. In *The Effect of Research-Based Instruction in Introductory Physics on a Common Cognitive Bias*; 2012 Physics Education Research Conference; AIP Publishing: 2013; pp 138−141.

(18) Karatjas, A. G. Comparing College Students' Self−Assessment of Knowledge in Organic Chemistry to Their Actual Performance. *J. Chem. Educ.* **2013**, *90* (8), 1096−1099.

(19) Bol, L.; Hacker, D. J.; Walck, C. C.; Nunnery, J. A. The effects of individual or group guidelines on the calibration accuracy and achievement of high school biology students. *Contemp. Educ. Psychol.* **2012**, *37* (4), 280−287.

(20) Snyder, K. E.; Nietfeld, J. L.; Linnenbrink-Garcia, L. Giftedness and Metacognition A Short−Term Longitudinal Investigation of Metacognitive Monitoring in the Classroom. *Gifted Child Quart.* **2011**, *55* (3), 181−193.

(21) Gillström, Å.; Rönnberg, J. Comprehension calibration and recall prediction accuracy of texts: Reading skill, reading strategies, and effort. *J. Educ. Psychol.* **1995**, *87* (4), 545−558.

(22) Labuhn, A. S.; Zimmerman, B. J.; Hasselhorn, M. Enhancing students' self−regulation and mathematics performance: the influence of feedback and self−evaluative standards. *Metacognition Learn.* **2010**, *5* (2), 173−194.

(23) Beyer, S. Gender differences in the accuracy of self-evaluations of performance. *J. Pers. Soc. Psychol.* **1990**, *59* (5), 960.

(24) Beyer, S.; Bowden, E. M. Gender differences in seff-perceptions: Convergent evidence from three measures of accuracy and bias. *Personality and Social Psychology Bulletin* **1997**, *23* (2), 157−172.

(25) Lindsey, B. A.; Nagel, M. L. Do Students Know What They Know? Exploring the Accuracy of Students' Self-assessments. *Phys. Rev. ST Phys. Educ. Res.* **2015**, *11* (2), 020103.

(26) Bell, P.; Volckmann, D. Knowledge Surveys in General Chemistry: Confidence, Overconfidence, and Performance. *J. Chem. Educ.* **2011**, *88* (11), 1469−1476.

(27) Mathabathe, K. C.; Potgieter, M. Metacognitive monitoring and learning gain in foundation chemistry. *Chem. Educ. Res. Pract.* **2014**, *15* (1), 94−104.

(28) Pazicni, S.; Bauer, C. F. Characterizing illusions of competence in introductory chemistry students. *Chem. Educ. Res. Pract.* **2014**, *15*, 24−34.

(29) On the fall final exam, the last question was: "What letter grade do you expect to earn in [this course]?"

(A) A+, A or A− (B) B+, B or B− (C) C+ or C (D) D (E)F

(30) Fritz, C. O.; Morris, P. E.; Richler, J. J. Effect size estimates: current use, calculations, and interpretation. *Journal of Experimental Psychology: General* **2012**, *141* (1), 2.

(31) Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*; Lawrence Erlbaum Associates, Inc., 1977.

(32) Coolican, H. *Research Methods and Statistics in Psychology*; Psychology Press, 2014.

(33) Upton, G. J. Fisher's exact test. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **1992**, *155*, 395−402.

(34) Bland, J. M.; Altman, D. G. Multiple significance tests: the Bonferroni method. *BMJ.* **1995**, *310* (6973), 170.

(35) Schraw, G.; Roedel, T. D. B. Test difficulty and judgment bias. *Mem. Cognition* **1994**, *22* (1), 63−69.

(36) Pulford, B. D.; Colman, A. M. Overconfidence: Feedback and item difficulty effects. *Pers. Indiv. Differ.* **1997**, *23* (1), 125−133.

(37) Koriat, A. Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *J. Exp. Psych.* **1997**, *126* (4), 349.

(38) Kelley, C. M.; Jacoby, L. L. Adult egocentrism: Subjective experience versus analytic bases for judgment. *Journal of Memory and language* **1996**, *35* (2), 157−175.

DOI: 10.1021/acs.jchemed.5b00705
J. Chem. Educ. XXXX, XXX, XXX−XXX