JOURNAL OF CHEMICAL EDUCATION

Article

pubs.acs.org/jchemeduc

Comparing Student Performance Using Computer and Paper-Based Tests: Results from Two Studies in General Chemistry

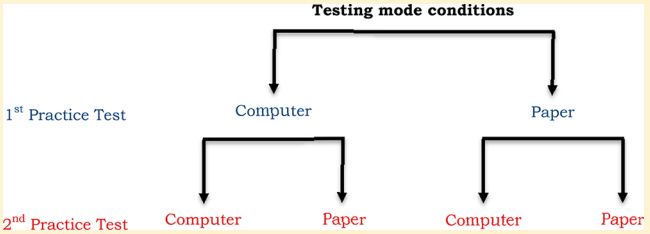
Anna A. Prisacari,[†] Thomas A. Holme,*[,†,‡] and Jared Danielson[†]

[†]Human Computer Interaction, Iowa State University, Ames, Iowa 50011, United States

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

Supporting Information

ABSTRACT: Taking a test online rather than on paper is becoming increasingly common. However, there has been little research directly addressing the testing mode (taking a test on paper or online) in chemistry courses, particularly when students take multiple practice tests before an exam. Two studies were conducted to investigate student performance on two proctored general chemistry practice tests as a function of the testing mode. Data were collected in 2013 (Study 1) and again in 2015 (Study 2). The participants were 422 undergraduate students (Study 1 N = 207 and Study 2 N = 215) from a first-semester general chemistry course at a midwestern university. In each study students took two practice tests. Each test included 17 algorithmic, 5 conceptual, and 2 definition questions and was administered on computer or paper. The mode combination of Test 1−Test 2 identified the four conditions: Computer−Computer, Computer−Paper, Paper−Computer, and Paper−Paper. The results show minor differences between online and paper modes. In particular, no significant difference was found between Computer−Paper and Paper−Paper conditions. This pattern suggests that online testing is a promising alternative to the traditional paper-and-pencil mode most often used in chemistry.



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

FEATURE: Chemical Education Research

INTRODUCTION

Recently it has become more common to assess students' learning in online settings. For the past several years, the American Chemical Society Examinations Institute (ACS-EI) has been working to design online modes for its tests. In addition to norm-referenced exams, ACS-EI offers several student practice tests online such as full-year general chemistry, full-year organic chemistry, analytical chemistry, and first-term general chemistry. With such an increase in computer-based assessment this study was undertaken to examine the equivalency of paper-based and computer-based general chemistry practice tests by measuring and comparing performance of students who take the same practice test in one of two modes. The results of this study may provide specific insights into how to deliver general chemistry practice tests so student learning from them is maximized. Instructors who are considering or already using an online platform to test their students may find the comparative data obtained here of interest.

Online Mode in Chemistry

As technology becomes more affordable and powerful, computer-based activities in chemistry classroom settings provide powerful means for fostering student learning[1,2] and enhancing the quality of assessment.[3] While teaching chemistry has traditionally been done in face-to-face classroom settings using hands-on lab experience and paper-and-pencil methods for assessing student learning, recent research shows alternative and reasonably effective strategies to teach chemistry and deepen student understanding by using technology. For example, computer-based animations and simulations can help students learn abstract concepts.[4,5]

The opportunities to assess student learning have also changed with the development of new tools that offer unique advantages over traditional, paper-based tests.[6] Notably, the growing body of research regarding testing mode, which typically involves comparing student test performance on a paper-based test to test performance on the same test delivered online, has often shown mixed findings. In addition to the mixed results, these studies have three important limitations. First, the effect of the testing mode has not been explored with multiple tests. Is there a testing mode order effect? Even though students routinely take multiple tests to illustrate their learning in a class, little empirical data has been gathered regarding whether taking one test on paper and the next test online is as effective as taking both tests only via paper or online. Investigating the testing mode order would help to construct

Received: April 19, 2017
Revised: September 9, 2017
Published: October 19, 2017

DOI: 10.1021/acs.jchemed.7b00274
J. Chem. Educ. 2017, 94, 1822−1830

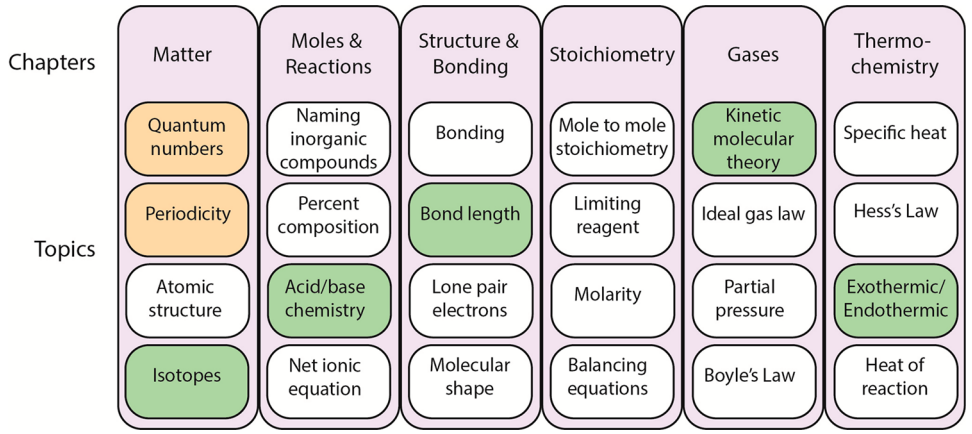


Figure 1. Listing of 17 algorithmic (white), 5 conceptual (green), and 2 definition (yellow) topics by 6 chapters. Each chapter contains 4 topics, and each topic includes 1 question pair that is composed of 2 questions similar in content and level of difficulty.

a more complete picture about technology-infused student learning. Such information is essential to instructors who teach blended classes or have opportunities to assess their students using either paper-based or computer-based testing modes. In addition, chemistry instructors would benefit from knowing in which mode to deliver practice tests and exams in order to maximize student performance.

Second, testing mode has not been explored in detail in the context of general chemistry. Efforts to evaluate the comparability of a paper-based and computer-based modes have been demonstrated with material from disciplines like anatomy,[7,8] biology,[9,10] English,[11,12] mathematics,[10,13] and reading,[14] but these studies produced mixed results. While some studies report no significant differences between computer-based and paper-based tests,[7,8,15−18] other studies report that two modes do not produce equivalent test performance[19] and differences generally favor the paper mode.[11,13,20] Because many undergraduate students take general chemistry courses to fulfill their degree requirements, it would be worthwhile to examine whether the effect of the testing mode is present in a general chemistry setting. Lastly, Cumming[21] suggests that "a single study is rarely, if ever, definitive; additional related evidence is required" to increase precision and robustness of the original work. To our knowledge, a direct replication of testing mode study has never been tested. Therefore, two years after the initial study, we replicated it with new students to ensure reliability of the results.

PURPOSE OF THE STUDY

The goals of this study were to examine whether (a) the testing mode and the testing mode order are related to student performance on assessments in a general chemistry practice exam setting and (b) the results of the initial study can be replicated with new students. To investigate the testing mode and its order, general chemistry students took two proctored practice tests in one of four conditions that were defined by the testing mode of the initial practice test and the second practice test. The conditions were the following: Computer−Computer, Computer−Paper, Paper−Computer, and Paper−Paper. Course-relevant material included items classified as algorithmic, conceptual, and definition questions, question types that are common in general chemistry.[22−27] Different question types allowed us to examine the testing mode not only at the broad level, but also at the level of individual items, based on their type. Using the settings that are common to undergraduate general chemistry tests, all sessions were timed, proctored, and conducted in classrooms. To examine whether the testing mode results can be replicated, the experiment was repeated after a two year delay with new students using the same design, material, and classroom setting. Thus, the data were collected twice: in 2013 (i.e., Study 1) and 2015 (i.e., Study 2) fall semesters. The research questions of Study 1 and Study 2 were the following:

(1) Is there a difference in test performance for students across the four testing mode conditions?
(2) Is there a difference in test performance for algorithmic and conceptual questions based on the mode of tests?

STUDY 1

Methods

Participants. A total of 221 students were recruited from a first-semester general chemistry at a large midwestern university. However, data from three students were excluded for the following reasons: One student left in the middle of the session, and two students experienced some issues with their laptops that prevented their data from being fully recorded, resulting in 218 students as the original sample size. Results from 11 more students were removed because they had answered some component set of the test items, such as algorithmic or conceptual questions, correctly in their first practice. Such a score on the first test precludes a student demonstrating any gain from repeated practice for that type of question. Thus, the final sample size contained 207 students (male = 37.2%, female = 61.4%, no response = 1.4%; mean age = 19). The course from which students were recruited lasted 15 weeks and included four sessions a week: three 50 min lectures and one 50 min recitation that was led by a senior undergraduate or graduate student. Notably, in addition to receiving the opportunity to practice before final course examination, all students were given complementary access to the online ACS general chemistry practice test for participating in the project. The study was approved by the Institutional Review Board (IRB), and the consent form was obtained from all participants at the beginning of the session (see Procedure section).

DOI: 10.1021/acs.jchemed.7b00274
J. Chem. Educ. 2017, 94, 1822−1830

Materials and Design. Because the final exam of this course was either the 2012 (in Study 1) or 2015 (in Study 2) ACS first-semester general chemistry exam, the practice test items were constructed using items that had items constructs and content coverage commensurate with an ACS Exam. *The items were not taken from any released ACS Exam.* To compose the test items, we used six areas that are typically taught at the first-semester general chemistry university level. For convenience, these areas will be referred to as "chapters" in the subsequent descriptions of the project. As depicted in Figure 1, each chapter contained four topics. For example, the four topics of Chapter 1 on matter were quantum numbers, periodicity, atomic structure, and isotopes. For each topic, two closely related multiple-choice questions were composed, resulting in four question pairs for each chapter or 24 question pairs in total (see Figure 1 where each cell represents a question pair).

Because students took two practice tests, one question of each pair was displayed on Test 1 and another question on Test 2. Thus, each practice test contained 24 questions, and each test displayed different questions on the same set of topics. An example of a complete 24-item test is included with the Supporting Information. We illustrate an example of a "related pair" of multiple-choice items in Boxes 1 and 2, where the pair of questions was composed to test knowledge of lone pairs versus bonding pairs of electrons (i.e., Chapter 3, Topic 3):

Box 1. Question A

When the correct Lewis structure is drawn for acetylide ion, C_2^{2-}, what is the total number of electron lone pairs present?

A. 0
B. 1
C. 2
D. 3

Box 2. Question B

When the correct Lewis structure is drawn for acetylide ion, C_2^{2-}, what is the total number of bonding electron pairs present?

A. 0
B. 1
C. 2
D. 3

Because students took two tests, the material on the tests was counterbalanced, a technique that is often used with repeated measures design.[28] First, we counterbalanced the question format so all questions within each topic appeared equally often in multiple-choice (MC) and open-ended format (OE). To accomplish this, all OE items (algorithmic, conceptual, and definition) were composed by removing the answer options from MC items. As a result, half of the questions were displayed to students in MC format and half in OE format. Second, for each topic, we counterbalanced the question format order (i.e., seeing Question A and Question B in only MC format versus seeing Question A in MC and Question B in OE format), which resulted in four Test 1−Test 2 combinations (MC−MC, MC−OE, OE−MC, and OE−OE). These format combinations were then rotated through the four topics of each chapter, creating four test versions. For example, version 1 test represented Topic 1 of all chapters in Test 1−Test 2 combination 1 (Question A MC, Question B MC), Topic 2 in combination 2 (Question A MC, Question B OE), Topic 3 in combination 3 (Question A OE, Question B MC), and Topic 4 in combination 4 (Question A OE, Question B OE). Lastly, in order to randomize possible item order effects, we manipulated the order of Question A and Question B by reversing the test order of each test version. For example, version 5 test became the reversed version 1 test (i.e., Topic 1: Question B MC, Question A MC; Topic 2: Question B OE, Question A MC; Topic 3: Question B MC, Question A OE; and Topic 4: Question B OE, Question A OE). This step doubles the number of test versions, resulting in eight possible tests that were administered to participants. Figure 2 illustrates all steps of counterbalancing and eight possible combinations for question pairs which were used to generate eight equally balanced test versions. Each student took the tests in only one version.

Differences based on item construct (OE versus MC) were observed and were consistent with prior work that generally finds that open-ended items are more difficult. A detailed description of how the material was counterbalanced and discussion of the item construct findings are presented in another publication.[29] Because this component of the study is largely a replication of earlier work, a graphical summary of the item construct results is included only in Supporting Information.

Lastly, all questions were classified by five experts into three categories: algorithmic, conceptual, and definition. Figure 1 shows the distribution of the six chapters and their topics,

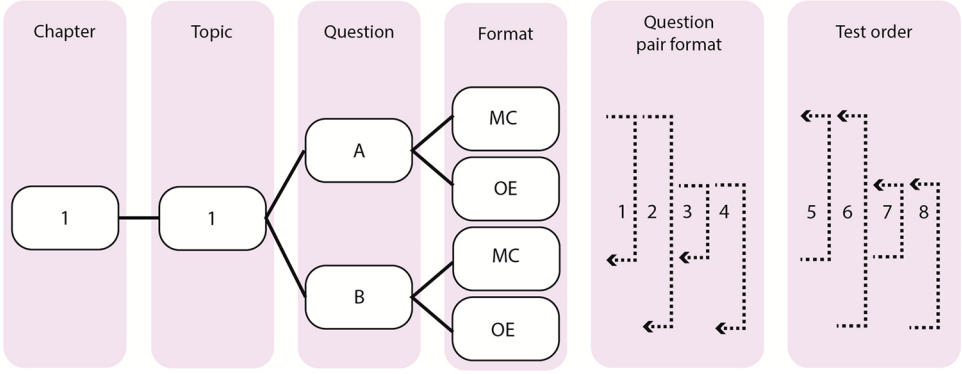


Figure 2. Example of counterbalancing process for one of the topics of Chapter 1 and eight combinations for its question pair. Combinations 5−8 represent the same question pair and the question format, only in the reverse order of combinations 1−4.

DOI: 10.1021/acs.jchemed.7b00274
J. Chem. Educ. 2017, 94, 1822−1830

indicating 17 algorithmic, 5 conceptual, and 2 definition question pairs. The material was presented on either paper or computer, and students were given a particular form depending on the counterbalanced condition to which they were assigned. All students were assigned to one of four conditions (see Procedure section for more information) that were identified by the testing mode of Test 1 and Test 2: Computer−Computer (N = 22%), Computer−Paper (N = 21%), Paper−Computer (N = 33%), and Paper−Paper (N = 24%) (see Table 1). While some students experienced only one testing mode (i.e., Computer−Computer and Paper−Paper), other students took two tests in different modes (i.e., Computer−Paper and Paper−Computer). This design allowed for examining not only the differences in test performance by the testing mode, but the effect of testing mode order as well.

Table 1. Original and Final Sample Size Distribution among Four Conditions of Study 1

Condition	Sample Size, N (%)	
	Original	Final
Computer−Computer	48 (22.0)	47 (22.7)
Computer−Paper	45 (20.6)	41 (19.8)
Paper−Computer	73 (33.5)	68 (32.9)
Paper−Paper	52 (23.9)	51 (24.6)
Total	218 (100)	207 (100)

Procedure. Students in the general chemistry class were invited to participate in this study, after the class had covered most of the content. The value of participation was described only in terms of preparation for the final exam. No course points were offered for participation. Students who were interested in participating received a link that contained a scheduling interface that listed multiple options of when they could take the practice tests and served as a registration process. The days and times for each condition were determined prior to student recruitment but not disclosed to students. Thus, students signed up for conditions on the basis of their time preference rather than the testing mode preference. After registration was closed, all students received an e-mail that notified them of the location of the study, included a copy of the consent form, and listed items that they needed to bring with them to the session. All students were asked to bring a basic calculator, scratch paper, and a pencil. In addition to these items, students in the conditions that included at least one online practice test were instructed to bring a laptop. To replicate the setting where testing commonly occurs, all sessions were conducted in classrooms and were proctored by the main researcher.

All practice tests were designed to mimic timed tests that the students would take as their final exam in the course. Before the session started, the researcher distributed a copy of the periodic table and a general chemistry data sheet that contained some general chemistry formulas, because these materials were permitted during the class's final examination. After signing the consent form, students received instructions according to their condition. For example, for the paper-based test, students were informed when to open the test and at which page to stop whereas for the online-based test, students received instructions regarding how to log on. All online tests were delivered via Qualtrics software and were protected by a password. The password was given to students before Test 1 and could be used only during the session time, which did not permit students to share their password with other students nor retake the test after the session. The timeline of the student experience of each session is shown in Figure 3.

After receiving the printed package or activating Test 1, students were given 35 min to complete Test 1. The proctor announced when there were 10 min remaining. After Test 1, students received feedback in the form of correct answers, and feedback was given in the same testing mode as Test 1. If a student took Test 1 online, then the correct answers along with the student's submitted answers appeared on the screen. If the student took Test 1 on paper, then he or she needed to flip to the feedback page to see the list of the correct answers and flip the pages of the test back and forth to compare the answers. Students were allowed to review their test performance for several minutes and were instructed to begin the next test only after the proctor confirmed that all students in the session had finished their review. After Test 1, all students repeated half of Test 1 (i.e., 12 questions; time = 10 min) and again were provided with feedback. This step in the design was included to assess memory factors that were reported on elsewhere.[29] Next, all students participated in a 20 min distractor task that involved answering five, non-chemistry-related trivia questions. This task allowed students to take a short break and also served as a check on the impact of short-term recall of answers to Test 1. Following the distractor task, students took Test 2 and, after completing it, were given the feedback as in Test 1. Finally, students completed a short demographics survey after which the researcher collected students' scratch paper and paper-based test or verified the online submission, gave access to the ACS online practice test, and thanked and dismissed the students. All sessions took place 1−15 days before the course's comprehensive exam, were conducted in groups of 3−33 students, and lasted approximately 1.5−2 h.

Scoring. Using a rubric, each question was scored as 1 (i.e., correct) or 0 (i.e., incorrect) by the principal investigator and several chemistry graduate students and postdoctoral associates. Next, using the proportion of correctly answered questions of Test 1 and Test 2, individual normalized gains[30] were calculated for overall performance on the practice exams. In addition, individual normalized gains were determined for the subsets of items characterized as algorithmic and conceptual. Definition questions were excluded from the subset calculation because with only two such questions many students answered both correctly on the first practice test and could not demonstrate gains in this category as a result.

Results

To test if the assumptions of ANOVA have been met, several tests were performed. Levene's test for homogeneity of variance between conditions showed a nonsignificant result (i.e., p-value = 0.95), and gains appeared to show a reasonably normal distribution in the histogram and Normal Q−Q Plot. Because

Figure 3. Timeline of the student experience during the study.

DOI: 10.1021/acs.jchemed.7b00274
J. Chem. Educ. 2017, 94, 1822−1830

this project was conceived to look at student performance gains due to practice tests, a few students who scored 100% either on Test 1 or a particular question type set from Test 1 (for question type analysis) were removed from the analysis where they could not show any gain. Thus, the final sample size contains students who have the possibility of gains in all three categories: overall, algorithmic, and conceptual. Table 1 shows the original and final sample sizes by four conditions.

The α-level for all analyses was set at 0.05, and all results were depicted graphically with confidence intervals instead of standard errors, as suggested by Cumming.[21] First, we examined the effect of testing mode on the overall gains. To answer the first research question, a one-way analysis of variance (ANOVA) was performed with conditions as the independent variable and overall normalized gains as the dependent variable. The results, $F(3, 203) = 7.47$, $p = 0.0001$, $\eta^2 = 0.1$, indicated a significant difference among the conditions and a medium effect.[31] Posthoc comparisons using the Tukey HSD test indicated that student gain in the Paper−Paper condition ($M = 0.32$, SD = 0.30) was significantly higher than the gain in the Computer−Computer condition ($M = 0.12$, SD = 0.29), the gain in the Computer−Paper condition ($M = 0.27$, SD = 0.33) was significantly higher than the gain in the Paper−Computer condition ($M = 0.08$, SD = 0.30), and the gain in the Paper−Paper condition ($M = 0.32$, SD = 0.30) was significantly higher than the gain in Paper−Computer ($M = 0.08$, SD = 0.30). The results are displayed in Figure 4.

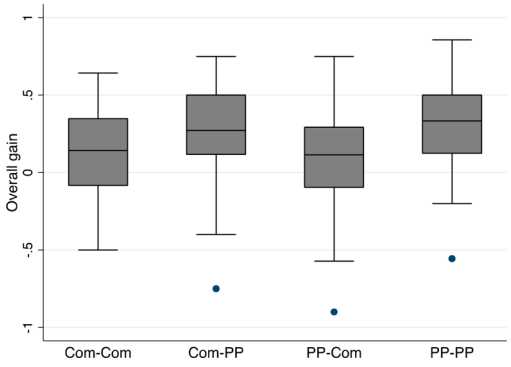


Figure 4. Boxplots for overall gains by conditions for Fall 2013.

To answer the second research question, two ANOVA analyses were performed by grouping items as algorithmic and conceptual and considering gain score differences in the four previously defined conditions for each group. For the algorithmic gain, results showed significance, $F(3, 203) = 3.72$, $p = 0.01$, and a small effect size ($\eta^2 = 0.05$). The posthoc comparisons showed that Paper−Paper gain ($M = 0.31$, SD = 0.42) was significantly higher than Paper−Computer gain ($M = 0.06$, SD = 0.42). The mean algorithmic gains are presented in Figure 5.

Similar to algorithmic gain, analysis of conceptual gain revealed a significant effect of the testing mode with a small effect, $F(3, 203) = 2.85$, $p = 0.04$, $\eta^2 = 0.04$, but the posthoc Tukey test showed no significant differences among conditions ($p > 0.05$) (see Figure 6). While the F test is significant, the Tukey test may not always show significance, because it is a conservative test that attempts to control the overall α-level.[32] These results in our initial study suggest that although the effect of testing mode was found to be significant at the overall level with gains for both algorithmic and conceptual items, the

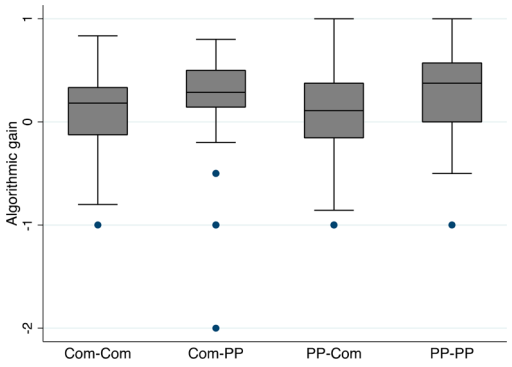


Figure 5. Boxplots for algorithmic item gains by conditions for Fall 2013.

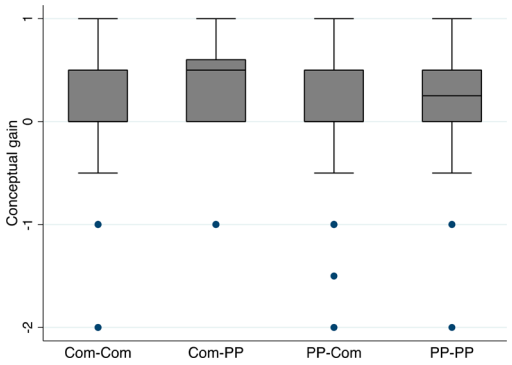


Figure 6. Boxplots for conceptual item gains by conditions for Fall 2013.

difference was small, effectively arguing that the testing mode did not produce meaningful changes in gains. These 2013 results were intriguing, so a replication study with a new cohort of general chemistry students was carried out in fall 2015.

STUDY 2

Methods

Prior to student recruitment, a total sample size was calculated to determine how many students were needed to detect any differences in gains among testing mode conditions. First, we estimated the means of overall gain for different conditions (Computer−Computer = 0.12; Computer−Paper = 0.27; Paper−Computer = 0.08; and Paper−Paper = 0.32) and error variance (i.e., 0.09) of Study 1. Using these estimates we computed the required sample assuming 5% significance level and 80% power in STATA 13.1.[32] The required minimum sample size was 108. Since dropout was anticipated (i.e., not all participating students may complete the study successfully), the required sample size was inflated from 108 to 122 or 31 participants per condition using the following formula: $n_d = n/(1 - R_d)^2$, where n_d is the inflated estimated sample size, n is the original estimated sample size, and R_d is the rate of dropout[33] that was calculated using Study 1 data (out of 221 students who participated, only 207 were included in the analysis = 6% dropout rate).

A total of 250 new students were recruited from students taking the same general chemistry course 4 semesters later than those who participated in the initial study. However, 7 students were not included in the analyses (3 students were under the age of 18, 1 student forgot to charge their laptop and thus had to switch the condition in the middle of the session, 1 student

DOI: 10.1021/acs.jchemed.7b00274
J. Chem. Educ. 2017, 94, 1822−1830

did not follow the instructions correctly, and 2 students left in the middle of the session). Thus, the original sample size was 243. Next, 28 students were removed as they answered either all test questions correctly, all algorithmic questions, or all conceptual questions correctly on Test 1. Therefore, the final sample size was 215 students (male = 35%, female = 59%, no response = 6%; mean age = 18.6). None of these students participated in Study 1, and the only benefit they received was the additional opportunity to practice before the final exam. In other words, students in the replication study did not receive an access code to the online ACS practice exam in addition to the preparations arising from the participation in the project. Student sample size by condition is shown in Table 2. All

Table 2. Original and Final Sample Size Distribution among Four Conditions of Study 2

Condition	Sample Size, *N* (%)	
	Original	Final
Computer–Computer	57 (23.5)	52 (24.2)
Computer–Paper	57 (23.5)	55 (25.6)
Paper–Computer	56 (23.0)	48 (22.3)
Paper–Paper	73 (30.0)	60 (27.9)
Total	243 (100)	215 (100)

sessions took place 3–7 days before the course's comprehensive exam and were conducted in groups of 58–73 students. The material, design, procedure, and scoring were the same as in Study 1.

Results

Before analyses were conducted, ANOVA assumptions have been verified. Levene's test for homogeneity of variance between conditions showed a nonsignificant result (i.e., *p*-value = 0.26) and the histogram and Normal *Q*–*Q* Plot graphs of gains appeared to show a normal distribution. By contrast to Study 1, the results of overall normalized gains showed no support for the testing mode ($F(3, 211) = 2.05$, $p = 0.1077$) as illustrated in Figure 7.

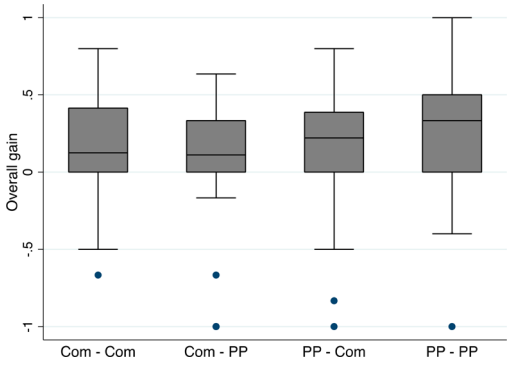


Figure 7. Boxplots for overall gains by conditions for Fall 2015.

Next, analyses were performed to identify gains for only the algorithmic and conceptual item sets. For algorithmic gain, the results revealed no significant difference among conditions $F(3, 211) = 1.23$, $p = 0.3001$ (Figure 8), while significance was detected with gain of conceptual questions $F(3, 211) = 3.67$, $p = 0.01$, $\eta^2 = 0.05$ (Figure 9). The posthoc comparisons showed that Paper–Paper gain ($M = 0.48$, SD = 0.60) was significantly higher than Computer–Paper gain ($M = 0.07$, SD = 0.76).

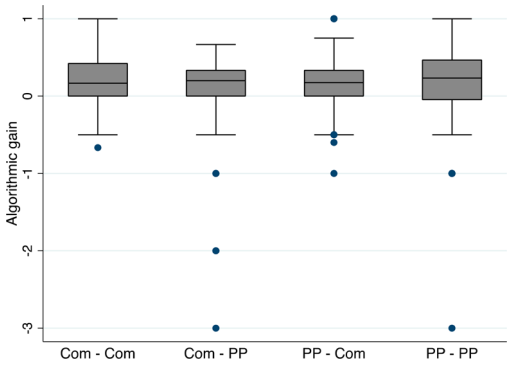


Figure 8. Boxplots for algorithmic item gains by conditions for Fall 2015.

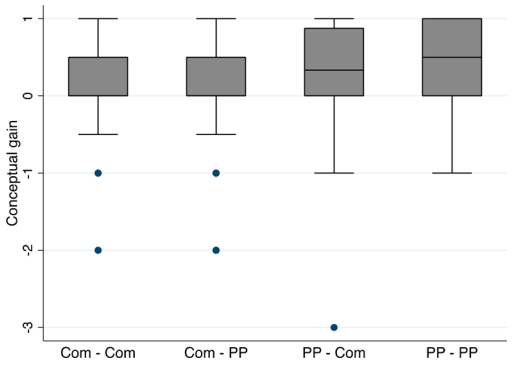


Figure 9. Boxplots for conceptual gains by conditions for Fall 2015.

DISCUSSION

Students face learning in computerized environments in many contexts including chemistry. The first chemistry massive open online course, or MOOC, was offered in 2012,[34] and since then, many chemistry education researchers have explored the landscape of the online environment and its effect on student learning.[35−37] Even though technology offers unique opportunities to support assessment activities in science,[38] little is known regarding whether students taking online assessments in general chemistry can produce similar test performance as students who take the same test in traditional, paper-and-pencil mode.

An interest in the relative impact of computer-based assessment derives from well-known learning effects, such as the testing effect.[39] In this case, the investigations were carried out in a low-stakes environment using a practice test methodology. Because the students self-selected and were practicing for an upcoming class exam, however, they were likely to be sufficiently motivated to perform well on the practice tests. Not surprisingly, performance gains were always present when comparing Test 1 to Test 2. The results of Study 1 in 2013 showed the presence of testing mode effects, indicating the benefit of taking the second test on paper regardless of the testing mode of Test 1. The highest gains were observed in Computer–Paper and Paper–Paper conditions, and no significant difference was found between these two conditions. This study suggested that although a combination of online and paper testing can be attractive to both instructors and students, the testing mode order, or which test mode comes first, might be worth consideration in designing study activities. The results showed a significant difference between Paper–Computer and Computer–Paper conditions, suggesting

DOI: 10.1021/acs.jchemed.7b00274
J. Chem. Educ. 2017, 94, 1822−1830

that student test performance would be enhanced if students take the first practice test online and the next one on paper. Further analyses based on categorizing the question type revealed that the testing mode differences were potentially attributable to performance gains on conceptual questions. Finally, it would be ideal to include student performance on the final exam in this study, but all students in the class had a variety of preparation tools available in addition to these practice tests, so within the design parameters of the study, it is not possible to make meaningful comparisons and attribute them to a testing effect for participants.

Study 1 was replicated with 215 new general chemistry students two years later (e.g., Study 2). Importantly while some differences were detected among student performances in Study 1 with a medium effect, no testing mode effect was found in Study 2. The fact that the second study, in 2015, showed no significant effects associated with testing mode, while the earlier study detected the testing mode effect is worth further investigation. Ultimately, there are a number of factors that are likely to influence student performance differences such as students' prior knowledge. It is possible that students in one study were better prepared for the tests than students in the other study. To test this hypothesis, we compared the mean performance on the initial test and found that students in Study 2 (M = 0.67) significantly outperformed students on Test 1 in Study 1 (M = 0.59). Thus, this group appears to have been better prepared. Because we used a convenience sample, it is difficult to attribute this difference to any other external factor such as instructors or content coverage of the two courses. Nonetheless, prior research suggests that the effect of testing may be moderated by students' prior knowledge.[39−41] Future research could investigate further whether testing mode differences are more predominant with students who are less prepared versus students who are better prepared for general chemistry tests.

■ LIMITATIONS

While this study is informative regarding the effect of testing mode on student performance in general chemistry, it has some limitations. First, the participants were self-selected students who might be different from those students who chose to not participate. The weight of the final exam on student grades in the two years was similar, and both courses were using an ACS exam, but there still could have been motivational differences for students between the two courses. No information was obtained about students who chose not to participate, so little additional information can be gleaned about the relative importance of the self-selection process for recruiting participants. Second, data for some students were excluded from the analyses as normalized gains could not be calculated due to their perfect scores on the initial test. This was particularly problematic when calculating performance gains for items categorized as definition items. Because there were only two definition questions, a high percentage of students (34% in Study 1 and 65% in Study 2) answered both questions correctly, so gain scores could not be defined for those students. If student performance on definition style items is a concern for chemistry instruction, a higher number of items would be needed to study this type of question in subsequent studies. Third, the material contained an unequal number of questions in each question type category. Out of 24 questions, there were 17 algorithmic, 5 conceptual, and 2 definition questions. While this distribution of questions was a good representation of the type of questions general chemistry students typically see on their final exam, it complicates investigations as to whether the testing mode equally affects all chemistry question types or it is predominant for one type of item more than others.

Finally, the time period between two practice tests was short. Students took both tests on the same day with a 20 min time interval. While a short delay between activities has been previously used in experimental studies of memory[42,43] and it simplifies the study procedure (i.e., helps to keep the sample size high by avoiding reduction in sample size as the result of fewer students returning to the next session), a short delay between two tests does not reflect the authentic classroom setting well as students do not typically take back-to-back tests that cover the same content. Thus, having students take the tests on different days is a better strategy.[44] Inserting an additional time gap between learning activities (in this case a practice exam) represents a more common teaching strategy and, as shown by previous research, enhances long-term learning more than back-to-back or massed learning.[45−49] Known as the spacing effect, it represents the desirable difficulty, or better condition of learning, for learners.[50] The more difficult the retrieval task, the more it benefits learning.[51,52] If the material is easily available and still accessible in working memory, as is the case in massed learning, the individual only needs to review this information. However, when some time interval is inserted between the initial learning and testing, some deactivation or forgetting occurs. Accordingly, the person must regenerate the full retrieval process, and thus, recall becomes more challenging.[53] Because the spacing process requires time to process new material, during which the student rehearses and connects new material to knowledge already stored in his or her long-term memory, spaced practice works better than multiple learning sessions that take place one after another. The experiments by Karpicke and Roediger[48] showed that when participants recalled the material 2 days after the study session, delaying the initial test increased the difficulty of retrieval and, consequently, boosted student performance on the final test. To study ways to maximize student learning, future studies on the testing mode should extend the time interval between the tests. That is, learning activities should be separated by a period of at least 12 h that include a night sleep.[44] Such studies are more challenging to recruit student participants, but would likely show effects that are unachievable with the shorter-term experiments described here. The literature on distributed practice clearly suggests that separating learning activities by 1 day rather than conducting all learning on the same day aids students to remember content for an extended period of time.[44,46,54]

■ IMPLICATIONS

The results of our studies have interesting educational implications for chemistry instructors. Frequent and distributed tests represent an effective instructional method,[55] and the work presented here suggests that testing methods can be used with either online or paper testing modes. Although university instructors tend to adopt technology at a slow rate,[56] this study indicates that online practice tests, which may be easier to administer and grade, can work as an instructional method. Giving a practice test online and delivering the final exam on paper represents a good alternative to assessing students' knowledge with only the traditional paper-and-pencil mode. Using the mix of online and paper modes would allow

DOI: 10.1021/acs.jchemed.7b00274
J. Chem. Educ. 2017, 94, 1822−1830

instructors and students to derive benefits of both modes without compromising students' test performance. For instance, the online mode would permit instructors to effortlessly document and store student test scores for future content evaluation and comparison purposes, whereas it would provide instant feedback to students,[7] a feature that chemistry students have mentioned as their top reason to prefer online to paper mode.[29] Giving the last test on paper would permit chemistry students to take the test in what is commonly their preferred testing mode and work things out on paper, another common feature that was previously reported by chemistry students.[29]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available on the ACS Publications website at DOI: 10.1021/acs.jchemed.7b00274.

Graphical depictions of the impact of item type (open-ended versus multiple-choice) between tests, and example test with 24 items (PDF, DOCX)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: taholme@iastate.edu.

ORCID

Thomas A. Holme: 0000-0003-0590-5848

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Funding for this project was provided by the ACS Exams Institute as part of its research into the development of online testing methods. We would like to thank chemistry instructors who allowed recruiting their general chemistry students and students who were part of this study. Also, we would like to thank John Baluyut, Jennifer Brigham, Chamila DeSilva, Kimberly Linenberger, Cynthia Luxford, Jeffrey Raker, and Jessica Reed for their help with material writing and test grading.

■ REFERENCES

(1) Ardac, D.; Akaygun, S. Effectiveness Of Multimedia-based Instruction That Emphasizes Molecular Representations On Students' Understanding Of Chemical Change. *J. Res. Sci. Teach.* **2004**, *41* (4), 317−337.

(2) Frailich, M.; Kesner, M.; Hofstein, A. Enhancing Students' Understanding Of The Concept Of Chemical Bonding By Using Activities Provided On An Interactive Website. *J. Res. Sci. Teach.* **2009**, *46* (3), 289−310.

(3) Brandriet, A.; Holme, T. Development Of The Exams Data Analysis Spreadsheet As A Tool To Help Instructors Conduct Customizable Analyses Of Student ACS Exam Data. *J. Chem. Educ.* **2015**, *92* (12), 2054−2061.

(4) Suits, J. P. Design of Dynamic Visualizations To Enhance Conceptual Understanding in Chemistry Courses. In *Chemistry Education: Best Practices, Opportunities and Trends*; Garcia-Martinez, J., Serrano-Torregrosa, E., Eds.; Wiley-VCH: Weinheim, Germany, 2015; pp 595−619.

(5) Tang, H.; Abraham, M. R. Effect Of Computer Simulations At The Particulate And Macroscopic Levels On Students' Understanding Of The Particulate Nature Of Matter. *J. Chem. Educ.* **2016**, *93* (1), 31−38.

(6) Quellmalz, E. S.; Pellegrino, J. W. Technology And Testing. *Science* **2009**, *75* (5910), 75−79.

(7) Meyer, A. J.; Innes, S. I.; Stomski, N. J.; Armson, A. J. Student Performance On Practical Gross Anatomy Examinations Is Not Affected By Assessment Modality. *Anat. Sci. Educ.* **2016**, *9* (2), 111−120.

(8) Inuwa, I. M.; Taranikanti, V.; Al-Rawahy, M.; Habbal, O. Anatomy Practical Examinations: How Does Student Performance On Computerized Evaluation Compare With The Traditional Format? *Anat. Sci. Educ.* **2012**, *5* (1), 27−32.

(9) Chua, Y. P.; Don, Z. M. Effects Of Computer-based Educational Achievement Test On Test Performance And Test Takers' Motivation. *Comput. Human Behav.* **2013**, *29* (5), 1889−1895.

(10) Kim, D.; Huynh, H. Comparability Of Computer And Paper-and-pencil Versions Of Algebra And Biology Assessments. *J. Technol. Learn. Assess.* **2007**, *6* (4), 1−31.

(11) Emerson, L.; MacKay, B. A Comparison Between Paper-based And Online Learning In Higher Education. *Br. J. Educ. Technol.* **2011**, *42* (5), 727−735.

(12) Kim, D.; Huynh, H. Computer-based And Paper-and-pencil Administration Mode Effects On A Statewide End-of-course English Test. *Educ. Psychol. Meas.* **2008**, *68* (4), 554−570.

(13) Bennett, R. E.; Braswell, J.; Oranje, A.; Sandene, B.; Kaplan, B.; Yan, F. Does It Matter if I Take My Mathematics Test on Computer? A Second Empirical Study of Mode Effects in NAEP. *J. Technol. Learn. Assess.* **2008**, *6* (9). https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1639/1472 (accessed Aug 2017).

(14) Kim, H. J.; Kim, J. Reading From An LCD Monitor Versus Paper: Teenagers' Reading Performance. *Int. J. Res. Stud. Educ. Technol.* **2013**, *2* (1), 15−24.

(15) Horkay, N.; Bennett, R. E.; Allen, N.; Kaplan, B.; Yan, F. Does It Matter If I Take My Writing Test On Computer? An Empirical Study Of Mode Effects In NAEP. *J. Technol. Learn. Assess.* **2006**, *5* (2), 1−49.

(16) Hochlehnert, A.; Brass, K.; Moeltner, A.; Juenger, J. Does Medical Students' Preference Of Test Format (Computer-based Vs. Paper-based) Have An Influence On Performance? *BMC Med. Educ.* **2011**, *11* (1), 89.

(17) Tsai, T.; Shin, C. D. A Score Comparability Study For The NBDHE: Paper-pencil Versus Computer Versions. *Eval. Health Prof.* **2013**, *36* (2), 228−239.

(18) Alexander, M. W.; Bartlett, J. E.; Truell, A. D.; Ouwenga, K. Testing In A Computer Technology Course: An Investigation Of Equivalency In Performance Between Online And Paper And Pencil Methods. *J. Career Technol. Educ.* **2000**, *18* (1), 69−80.

(19) Clariana, R.; Wallace, P. Paper-based Versus Computer-based Assessment: Key Factors Associated With The Test Mode Effect. *Br. J. Educ. Technol.* **2002**, *33* (5), 593−602.

(20) Keng, L.; McClarty, K. L.; Davis, L. L. Item-Level Comparative Analysis Of Online And Paper Administrations Of The Texas Assessment Of Knowledge And Skills. *Appl. Meas. Educ.* **2008**, *21* (3), 207−226.

(21) Cumming, G. The New Statistics: Why And How. *Psychol. Sci.* **2014**, *25* (1), 7−29.

(22) Nakhleh, M. B.; Mitchell, R. Concept Learning Versus Problem Solving: There Is A Difference. *J. Chem. Educ.* **1993**, *70* (3), 190−192.

(23) Nakhleh, M. B. Are Our Students Conceptual Thinkers Or Algorithmic Problem Solvers? Identifying Conceptual Students In General Chemistry. *J. Chem. Educ.* **1993**, *70* (1), 52−55.

(24) Nurrenbern, S.; Pickering, M. Concept Learning Versus Problem Solving: Is There A Difference? *J. Chem. Educ.* **1987**, *64*, 508−510.

(25) Pickering, M. Further Studies On Concept Learning Versus Problem Solving. *J. Chem. Educ.* **1990**, *67*, 254−255.

(26) Sawrey, B. A. Concept Learning Versus Problem Solving: Revisited. *J. Chem. Educ.* **1990**, *67* (3), 253.

(27) Smith, K. C.; Nakhleh, M. B.; Bretz, S. L. An Expanded Framework For Analyzing General Chemistry Exams. *Chem. Educ. Res. Pract.* **2010**, *11* (3), 147−153.

DOI: 10.1021/acs.jchemed.7b00274
J. Chem. Educ. 2017, 94, 1822−1830

(28) Butler, A. C.; Roediger, H. L. Feedback Enhances The Positive Effects And Reduces The Negative Effects Of Multiple-choice Testing. *Mem. Cognit.* **2008**, *36* (3), 604−616.
(29) Prisacari, A. A. *The Testing Effect in General Chemistry: Effects of Repeated Testing on Student Performance across Different Test Modes*. M.S. Thesis, Iowa State University, Ames, IA, 2015.
(30) Hake, R. Interactive-engagement Vs. Traditional Methods: A Six-thousand-student Survey Or Mechanics Test Data For Introductory Physics Courses. *Am. J. Phys.* **1998**, *66* (1), 64−74.
(31) Cohen, J. The Analysis Of Variance. In *Statistical Power Analysis for the Behavioral Sciences*; Lawrence Erlbaum Associates, Inc.: Hillsdale, NJ, 1988; pp 273−406.
(32) Simon, S. When the F Test Is Significant, But Tukey Is Not. http://www.pmean.com/05/TukeyTest.html (accessed Aug 2017).
(33) *STATA Power And Sample-Size Reference Manual*, Release 13; STATA Press Publication: College Station, TX, 2013.
(34) Leontyev, A.; Baranov, D. Massive Open Online Courses In Chemistry: A Comparative Overview Of Platforms And Features. *J. Chem. Educ.* **2013**, *90* (11), 1533−1539.
(35) Amaral, K. E.; Shank, J. D.; Shibley, I. A.; Shibley, L. R. Web-enhanced General Chemistry Increases Student Completion Rates, Success, And Satisfaction. *J. Chem. Educ.* **2013**, *90* (3), 296−302.
(36) Winschel, G. A.; Everett, R. K.; Coppola, B. P.; Shultz, G. V.; Lonn, S. Using Jigsaw-style Spectroscopy Problem-solving To Elucidate Molecular Structure Through Online Cooperative Learning. *J. Chem. Educ.* **2015**, *92* (7), 1188−1193.
(37) Eichler, J. F.; Peeples, J. Online Homework Put To The Test: A Report On The Impact Of Two Online Learning Systems On Student Performance In General Chemistry. *J. Chem. Educ.* **2013**, *90* (9), 1137−1143.
(38) Kuo, C.-Y.; Wu, H.-K. Toward An Integrated Model For Designing Assessment Systems: An Analysis Of The Current Status Of Computer-based Assessments In Science. *Comput. Educ.* **2013**, *68*, 388−403.
(39) Pyburn, D. T.; Pazicni, S.; Benassi, V. A.; Tappin, E. M. The Testing Effect: An Intervention On Behalf Of Low-skilled Comprehenders In General Chemistry. *J. Chem. Educ.* **2014**, *91* (12), 2045−2057.
(40) Leppink, J.; Broers, N. J.; Imbos, T.; van der Vleuten, C. P. M.; Berger, M. P. F. Self-explanation In The Domain Of Statistics: An Expertise Reversal Effect. *High. Educ.* **2012**, *63* (6), 771−785.
(41) Lee, H.; Plass, J. L.; Homer, B. D. Optimizing Cognitive Load For Learning From Computer-based Science Simulations. *J. Educ. Psychol.* **2006**, *98* (4), 902−913.
(42) Carpenter, S. K.; DeLosh, E. L. Application Of The Testing And Spacing Effects To Name Learning. *Appl. Cogn. Psychol.* **2005**, *19* (5), 619−636.
(43) Peterson, D. J.; Mulligan, N. W. The Negative Testing Effect And Multifactor Account. *J. Exp. Psychol. Learn. Mem. Cogn.* **2013**, *39* (4), 1287−1293.
(44) Mazza, S.; Gerbier, E.; Gustin, M.-P.; Kasikci, Z.; Koenig, O.; Toppino, T. C.; Magnin, M. Relearn Faster And Retain Longer: Along With Practice, Sleep Makes Perfect. *Psychol. Sci.* **2016**, *27*, 1321−1330.
(45) Kornell, N.; Castel, A. D.; Eich, T. S.; Bjork, R. A. Spacing As The Friend Of Both Memory And Induction In Young And Older Adults. *Psychol. Aging* **2010**, *25* (2), 498−503.
(46) Cepeda, N. J.; Pashler, H.; Vul, E.; Wixted, J. T.; Rohrer, D. Distributed Practice In Verbal Recall Tasks: A Review And Quantitative Synthesis. *Psychol. Bull.* **2006**, *132* (3), 354−380.
(47) Kornell, N.; Bjork, R. Learning Concepts And Categories: Is Spacing The "enemy Of Induction"? *Psychol. Sci.* **2008**, *19* (6), 585−592.
(48) Karpicke, J. D.; Roediger, H. L. Expanding Retrieval Practice Promotes Short-term Retention, But Equally Spaced Retrieval Enhances Long-term Retention. *J. Exp. Psychol. Learn. Mem. Cogn.* **2007**, *33* (4), 704−719.
(49) Sobel, H. S.; Cepeda, N. J.; Kapler, I. V. Spacing Effects In Real World Classroom Vocabulary Learning. *Appl. Cogn. Psychol.* **2011**, *25* (5), 763−767.
(50) Bjork, E. L.; Bjork, R. A. Making Things Hard on Yourself, but in a Good Way: Creating Desirable Difficulties To Enhance Learning. In *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*; Gernsbacher, M., Pew, R., Hough, L., Pomerantz, J., Eds.; Worth: New York, 2011; pp 56−64.
(51) Bjork, R. A.; Allen, T. W. The Spacing Effect: Consolidation Or Differential Encoding? *J. Verbal Learning Verbal Behav.* **1970**, *9* (5), 567−572.
(52) Cuddy, L. J.; Jacoby, L. L. When Forgetting Helps Memory: An Analysis Of Repetition Effects. *J. Verbal Learning Verbal Behav.* **1982**, *21* (4), 451−467.
(53) Dellarosa, D.; Bourne, L. E. Surface Form And The Spacing Effect. *Mem. Cognit.* **1985**, *13*, 529−537.
(54) Kornell, N. Optimising Learning Using Flashcards: Spacing Is More Effective Than Cramming. *Appl. Cogn. Psychol.* **2009**, *23*, 1297−1317.
(55) Glass, A. L.; Sinha, N. Multiple-choice Questioning Is An Efficient Instructional Methodology That May Be Widely Implemented In Academic Courses To Improve Exam Performance. *Curr. Dir. Psychol. Sci.* **2013**, *22* (6), 471−477.
(56) Hora, M. T.; Holden, J. Exploring The Role Of Instructional Technology In Course Planning And Classroom Teaching: Implications For Pedagogical Reform. *J. Comput. High. Educ.* **2013**, *25* (2), 68−92.

DOI: 10.1021/acs.jchemed.7b00274
J. Chem. Educ. 2017, 94, 1822−1830