CBE—Life Sciences Education
Vol. 11, 294–306, Fall 2012

Article

Multiple-Choice Exams: An Obstacle for Higher-Level Thinking in Introductory Science Classes

Kathrin F. Stanger-Hall

Department of Plant Biology, University of Georgia, Athens, GA 30602

Submitted November 16, 2011; Revised May 22, 2012; Accepted May 22, 2012
Monitoring Editor: Eric H. Chudler

Learning science requires higher-level (critical) thinking skills that need to be practiced in science classes. This study tested the effect of exam format on critical-thinking skills. Multiple-choice (MC) testing is common in introductory science courses, and students in these classes tend to associate memorization with MC questions and may not see the need to modify their study strategies for critical thinking, because the MC exam format has not changed. To test the effect of exam format, I used two sections of an introductory biology class. One section was assessed with exams in the traditional MC format, the other section was assessed with both MC and constructed-response (CR) questions. The mixed exam format was correlated with significantly more cognitively active study behaviors and a significantly better performance on the cumulative final exam (after accounting for grade point average and gender). There was also less gender-bias in the CR answers. This suggests that the MC-only exam format indeed hinders critical thinking in introductory science classes. Introducing CR questions encouraged students to learn more and to be better critical thinkers and reduced gender bias. However, student resistance increased as students adjusted their perceptions of their own critical-thinking abilities.

INTRODUCTION

Higher-level processing of learned information, or critical thinking, is generally viewed as an essential part of college training (American Association for the Advancement of Science [AAAS], 1990, 1993, 2010; Boyer Commission on Educating Undergraduates in the Research University, 1998; National Research Council [NRC], 2003). Given that in 2009, only 21% of twelfth graders in the United States performed at or above the proficiency level in science (National Center for Education Statistics [NCES], 2009), the development of higher-level scientific-thinking skills in college poses a considerable challenge for both students and instructors. Repeated calls to reinvent science teaching and learning by the AAAS, the NRC, and the Boyer Commission on Educating Undergraduates in the Research University (AAAS, 1990, 1993, 2010; Boyer Commission, 1998; NRC, 2003) have been answered with many teaching innovations to promote student engagement and/or active learning strategies in science classes. These innovations include problem-based learning (Allen *et al.*, 1996; Eberlein *et al.*, 2008); process-oriented, guided-inquiry learning (Eberlein *et al.*, 2008; Moog and Spencer, 2008); collaborative learning (Crouch and Mazur, 2001; Smith *et al.*, 2009); peer-led team learning (Gosser and Roth, 1998; Stanger-Hal1 *et al.*, 2010); a new emphasis on "scientific teaching" methodologies (Handelsman *et al.*, 2004, 2007; Pfund *et al.*, 2009); and the use of technological innovations, such as personal-response systems for student engagement and immediate in-class feedback (Caldwell, 2007; Smith *et al.*, 2009), among others (Ebert-May and Brewer, 1997). Despite the increasing adoption of these innovative instruction methods over simple lecturing in college (66.3% of assistant professors and 49.6% of associate and full professors used student-centered and inquiry-based instruction methods in 2008; DeAngelo *et al.*, 2009), and 99.6% of all university professors indicating that helping students develop critical-thinking skills is very important (DeAngelo *et al.*, 2009), the outcomes for critical thinking have been disappointing so far. A recent study on student learning in U.S.

DOI: 10.1187/cbe.11-11-0100
Address correspondence to: Kathrin F. Stanger-Hall (ksh@uga.edu).

colleges and universities documented that 46% of college students did not gain critical-thinking skills during their first 2 yr of college, and 36% had not gained critical-thinking skills after 4 yr (Arum and Roksa, 2011). These data highlight the difficulties for both teaching and learning of critical-thinking skills in college, despite universal agreement on the importance of these skills. This raises the question whether this shortcoming is due to a lack of a critical-thinking challenge by instructors (Paul *et al.*, 1997; Haas and Keeley 1998; Crowe *et al.*, 2008; Zheng *et al.*, 2008; Momsen *et al.*, 2010; Arum and Roksa, 2011), student resistance to such a challenge (Keeley *et al.*, 1995; Weimer, 2002; Arum and Roksa, 2011), or a combination of both. This study focuses on student resistance and possible influences on resistance, specifically the exam format used to assess student learning.

Exam Format

There has been considerable discussion on the advantages and disadvantages of different exam formats (Biggs, 1973; Simkin and Kuechler, 2005), both from a pedagogical (focused on learning outcomes) and from a practical (time and cost) perspective. In general, from an instructor standpoint, multiple-choice (MC) questions can be advantageous with respect to ease of scoring, perceived objectivity in grading, fast return of scores in large classes, and the capacity to ask more questions (Simkin and Kuechler, 2005). Limitations of MC exams include the work-intensive construction of high-quality question banks (Simkin and Kuechler, 2005), the difficulty of assessing critical-thinking skills (Martinez, 1999), and possibly false indication of student knowledge and understanding (Dufresne *et al.*, 2002). While it is possible to write good critical-thinking (e.g., application, analysis, and evaluation) MC questions, they are usually difficult and time-intensive to create (Simkin and Kuechler, 2005), and synthesis (creative) skills cannot be assessed. In contrast, constructed-response (CR), such as fill-in-the-blank, short answer (SA), or essay questions that require students to create their own answer, can assess a wider range of thinking skills (Martinez, 1999), including critical thinking. In addition, CR questions give students the opportunity to express what they know (on all thinking levels). However, CR questions tend to be criticized for more subjective grading, intergrader variability, and time requirement for grading (Simkin and Kuechler, 2005). Students view MC exams as easier than essay exams and feel MC exams are easier to prepare for (require less time and effort), find the availability of options comforting, and like to be able to guess the right answer (Zeidner, 1987). Students also tend to expect knowledge questions from an MC exam and use surface (lower-level) learning when preparing for them (Scouller, 1998; Martinez, 1999). In contrast, students view essay exams as somewhat more appropriate for assessing the depth of their knowledge (Zeidner, 1987; Simkin and Kuechler, 2005). They may find grading more subjective, but some students believe that subjective grading could work to their advantage (Zeidner, 1987). Students also tend to expect more higher-level questions from CR exams and employ more deep-learning strategies in preparation for them (Scouller, 1998).

Finally, there is some inconclusive evidence for gender bias in assessment format. Some studies have found an advantage for males with MC questions, whereas others have found an advantage for males with CR questions (Bolger and Kellaghan, 1990; DeMars, 1998; Simkin and Kuechler, 2005).

MC in Introductory Science Classes

In a 2008 national survey, 33.1% of college instructors reported the use of MC exams (DeAngelo *et al.*, 2009), but MC testing is especially common in introductory science classes at research universities, because of logistics and institutional policies (e.g., class size and grading support). Students in these classes tend to have been successful using memorization to prepare for and perform well on MC exams that emphasize lower-level thinking skills (e.g., Scholastic Assessment Test, advanced placement, other introductory science classes: Zheng *et al.*, 2008). As a consequence, they have learned to associate MC questions with memorization (and other lower-level learning strategies: Scouller, 1998; Watters and Watters, 2007). This association likely undermines the credibility of the instructor and his or her attempts to convince introductory science students of the value of higher-level (critical) thinking, when students are tested with MC-only exams, even if these exams include higher-level MC questions. To test the hypothesis that the MC-only exam format hinders the development of higher-level thinking skills in introductory science students, I changed the exam format in a large introductory biology class with traditionally MC-only exams to include MC as well as several CR questions. I predicted that changing exam formats would result in 1) a change in how students studied and 2) a change in learning outcomes as assessed on the final exam, including 3) improved performance in critical (higher-level) thinking questions.

Study Design

The change in exam format was implemented in a large introductory biology class for biology majors (250–330 students per section). This section size is typical for general biology classes at research-intensive universities (Momsen *et al.*, 2010). The class in this study was the second of two introductory biology classes in the major sequence, and it focused on organismal diversity, phylogenetics, the evolution of structures and functions in plants and animals, and ecology. Critical thinking was emphasized in this class, and Bloom's taxonomy of thinking skills (Bloom, 1956; Anderson and Krathwohl, 2001) was taught to students as a communication and study tool during the first week of class. Critical thinking was defined in the framework of Bloom's taxonomy (as Bloom levels 3–6: application, analysis, evaluation, and synthesis), because this hierarchical model of thinking skills creates specific expectations for practice and assessment at each thinking level (see also Simkin and Kuechler, 2005; Crowe *et al.*, 2008). Furthermore, by using the term *higher-level thinking*, rather than the more abstract term *critical thinking*, it was the instructor's intent to remind students that these are progressive thinking skills that build upon others and can be practiced.

After introducing Bloom's taxonomy, the instructor explained to students that 25–30% of the questions on each exam would be asked at Bloom levels 3–5. As a result, students who desired to earn a grade of "C" or higher had to master these higher-level thinking skills (due to the MC exam format, level 6 [synthesis] was not assessed in the

lecture portion of this class). After teaching Bloom's taxonomy, the instructor demonstrated examples of higher-level thinking skills with sample clicker questions, and through interactive questioning during class. Throughout the semester students could practice their thinking skills by answering clicker questions, and subsequently discussing which evidence could be used to support or to eliminate different answer options (rather than being simply given the "correct" answer; see also Tanner and Allen, 2005). The students were also given question skeletons to design their own study questions at all Bloom levels, and instructed to practice the higher-level thinking skills during studying. However, despite these instructions (given to students every semester), students in this class usually struggle with the higher-level MC questions on the exams, and tend to perceive them as "tricky" (on the part of the instructor), rather than challenging (i.e., requiring higher-level thinking skills). After trying different approaches to convince the introductory biology students in this class of the value of practicing both lower- and higher-level thinking skills during studying (i.e., learning on all learning levels) while being assessed with a MC-only exam format (Stanger-Hall *et al.*, 2011), I decided to test whether this exam format posed an obstacle for motivating students to practice higher-level thinking.

MATERIALS AND METHODS

During Spring 2009, two sections of this introductory biology class were offered, both taught by the same instructor (K.S.-H.). Instruction, assignments, and study tips were identical for both classes. The larger section ($N = 282$ consenting students) was assessed using the traditional MC-only exam format, and the smaller section ($N = 192$ consenting students) was assessed using a combination of MC, SA, and other CR questions (denoted by MC+SA hereafter). Students were not aware of the assessment format when signing up for one of the two class sections. Students in both sections answered four online surveys during the semester and took four exams and a comprehensive final exam. I used data from the four surveys (study behavior) and the cumulative final exam (learning outcomes) to test the predictions for this study.

Exam Format

The exams of the MC section consisted of 50 MC questions, the exams of the MC+SA section consisted of 30 MC and 3–4 SA questions. SA questions could usually be answered in three or four sentences or by labeling diagrams. For each exam, 25–30% of the questions were higher-level thinking questions (application or analysis), and for each exam, this proportion was the same for the two sections. Whenever possible, the MC questions in the MC-only section and the SA questions in the MC+SA section were designed to assess the equivalent content and thinking skills (see Supplemental Material 1 for examples). The students in this study were not specifically trained how to answer SA questions, and the instructor provided the same study recommendations and sample exam questions to both class sections. The sample exam questions (from previous semesters) were in MC format, but the instructions asked students to answer the questions as essay questions first (without looking at the answer options) and to use logical step-by-step reasoning to arrive at the answer. Only after reasoning out their answers were students to look at the answer options and chose the matching answer. In addition, students in both sections were assigned a two-page reading on how to answer an essay question. The handout used an example from class and gave the students different examples of student answers (different quality) that students were asked to compare and score (see Supplemental Material 2).

Scoring Rubrics for Exams. For each SA question the instructor worked with the grader (graduate teaching assistant) to create grading rubrics for content and reasoning skills (see Supplemental Material 3 for an example). For fill-in or labeling questions, only content was scored. The original rubrics were tested and modified with ~40 exam answers, and the resulting rubrics were used to grade all exams. During grading, the rubrics were further adjusted as necessary. The grader graded all exams to ensure grader consistency. The results of the MC questions were released within 1 d of the exam; the results of the SA questions were posted within 1 wk, along with the answer keys. Students did not receive individual feedback on their exam answers.

Regrade Requests. Students in both sections could submit a written regrade request for any exam question (explaining why they should receive credit and using scientific facts and reasoning) within 1 wk of the answer key being posted. These regrade requests were considered by the instructor and returned the following week with written feedback on the validity of the request.

Student Surveys

Students in both sections filled out four online surveys during the semester. They received 3 points of class credit (of 1100) for each survey. All surveys were given during non–exam weeks. The first survey (week 2) was given before exam 1, the second survey (week 6) was given before exam 2, the third survey (week 12) was given before exam 4, and the final survey was given during the last week (week 15) of class.

During the first survey, students were asked which exam format they would prefer if given the choice (MC only, MC+SA, SA only), how many hours (per 7-d week) they usually studied for a science class during exam weeks and during non–exam weeks, and to report their current (start-of-the-semester) grade point average (GPA). In each of the four surveys, they were asked how much they had actually studied for the biology class during the previous 7-d week (a non–exam week), and which study behaviors they had used. The list of study behaviors contained nine cognitively passive (surface-learning) study behaviors and 13 cognitively active (deep-learning) study behaviors (Table 1), which had been developed from open-response student surveys in previous semesters and included study behaviors that had been recommended by the instructor at the beginning of the semester. Students could check as many study behaviors as they had used.

Although the term *active learning* is used in the literature (e.g., Prince, 2004) for both physically active (e.g., rewriting notes, making index cards) and cognitively active (e.g., making new connections, asking and answering new questions) learning behaviors, critical thinking cannot be achieved without the latter (Stanger-Hall *et al.*, in preparation). For this

Table 1. List of cognitively passive and active learning behaviors that students reported in their study surveys

Cognitively passive learning behaviors	Cognitively active learning behaviors
I previewed the reading before class.	I asked myself: "How does it work?" and "Why does it work this way?"
I came to class.	I drew my own flowcharts or diagrams.
I read the assigned text.	I broke down complex processes step-by-step.
I reviewed my class notes.	I wrote my own study questions.
I rewrote my notes.	I reorganized the class information.
I made index cards.	I compared and contrasted.
I highlighted the text.	I fit all the facts into a bigger picture.
I looked up information.	I tried to figure out the answer before looking it up.
I asked a classmate or tutor to explain the material to me.	I closed my notes and tested how much I remembered.
	I asked myself: "How are individual steps connected?" and "Why are they connected?"
	I drew and labeled diagrams from memory and figured out missing pieces.
	I asked myself: "How does this impact my life?" and "What does it tell me about my body?"
	I used Bloom's taxonomy to write my own study questions

reason, the present analysis differentiated between cognitively passive (can be physically active) and cognitively active behaviors.

In a series of questions, the students were also asked to rate the strength of their own ability to perform certain assessment tasks (e.g., to remember facts and explanations, to apply what they have learned to different situations, to analyze problems, to evaluate different solutions to a problem) on a 5-point Likert scale (from 1: "I struggle with it" to 5: "I am excellent at it"). In addition, they were asked to rate the statement "I see the value of learning on all learning levels" on a Likert scale from 1 (completely disagree) to 5 (completely agree).

Final Exams

Student performance on the cumulative final exam was used to assess whether exam format affected student learning (final exams are not returned in this class, and none of the final exam questions had appeared on previous exams). For the MC section, the final exam consisted of 125 MC questions (90 of these questions were also asked in the MC+SA section) and three CR questions: a 24-item, fill-in question in table format; a 12-item, fill-in flowchart; and an extra credit SA question in short essay format. For the MC+SA section, the final exam consisted of 90 MC questions; a 24-item, fill-in question in table format; a 12-item, fill-in flowchart; and five SA questions (four of these were in short essay format). Students had 3 h to complete the final exam, and in both sections only very few (<10) students remained until the end of the allotted time.

For the final exam comparison between the two sections, I used the 90 identical MC questions (29 higher-level and 61 lower-level thinking questions: categorized based on class content and activities, assignments, and assigned reading); the cumulative 24-item, fill-in table (could be answered with lower-level skills, e.g., remembering from classes throughout the semester, but higher-level thinking would have helped retrieval and checking for errors); the 12-item, fill-in flowchart (remembering from class 2 wk before the final exam); and one common higher-level short essay question (extra credit for the MC class and part of the exam for the MC+SA class).

GPA

Due to class logistics, it was not possible to assign students randomly to the two exam formats, therefore previous student achievement (GPA) was used to account for potential student differences between treatment groups and was used as a covariate.

Gender Differences

To address potential gender differences in student performance on different exam questions, I coded the gender of the participating students based on their first names as male or female. Ambiguous names (e.g., Ashley, Kerry, Tyler, and names from other cultures) were not coded, and these students were not included in this analysis. The final gender sample size was $N = 323$ ($N = 195$ in MC class and $N = 128$ in MC+SA class).

Student Evaluations

Summary statistics for the anonymous end-of-semester class evaluations are reported. Evaluations were submitted by 207 students from the MC class and 130 students from the MC+SA class.

Statistical Analysis

I used SPSS version 19.0 (2011) for all quantitative statistical analyses. For each class, I tested all variables for normality (goodness of fit: Shapiro-Wilk test; SPSS 19.0). Only the total MC (90 questions) and the higher-level MC (29 questions) scores of the final exam were normally distributed. As a result, I report the results of nonparametric tests for all analyses. For the comparison of GPA, final exam scores, and study data between the two classes, I used the nonparametric Mann-Whitney *U*-test for independent samples. This is a test for both location and shape to test for differences between distributions of ranked variables. To compare data from the study surveys throughout the semester (repeated samples), I used the related-samples Wilcoxon signed-rank test. To correct for multiple comparisons (inflated type I error), I applied a false discovery rate correction (Benjamini and Hochberg, 1995) and report the adjusted p values. To compare student preferences for exam format and their attitudes regarding the value of learning on all learning levels between the two classes (i.e., exam formats), I applied a Pearson χ^2 test, using the data from the MC class to calculate the expected values for the MC+SA class. I used a Spearman correlation to assess a possible relationship between the change in value ratings (value

rating after four exams minus value rating before the first exam) and the change in higher-level ability ratings (average in higher-level ability ratings after four exams minus average in higher-level ability ratings before the first exam). For a comprehensive analysis of the influences of previous student achievement (GPA), exam format, and gender on final exam performance, I conducted a two-way analysis of covariance (ANCOVA) with GPA as a covariate. For all variables we report means and standard error of the mean (mean ± SEM). All reported statistical results are based on two-tailed tests and significance levels of $p < 0.05$.

Sample Sizes for Analysis. In the MC class, 282 students consented to participate in this study, 231 students reported their start-of-semester GPA, and 195 students were identified (based on their first names) as male ($N = 78$) or female ($N = 117$). A total of 242 students finished the class (took the final exam). Of these, 172 students took all four study surveys (for longitudinal comparison) and reported on study times and study activities. Thus, the reported results of the MC class are based on sample sizes of 242 (final exam performance), 231 (GPA), 195 (gender), or 172 (study time and activities). When asked their preferred exam format (MC only, MC+SA, or SA only) at the beginning of the semester, 60% of the students in this class preferred MC only, 36% preferred MC+SA, and 4% preferred SA only.

In the MC+SA class, 192 students consented to participate in this study, 155 students reported their start-of-semester GPA, and 128 students were identified (based on their first names) as male ($N = 63$) or female ($N = 65$). A total of 164 students finished the class (took the final exam). Of these, 121 students took all four study surveys and reported on study times and study activities. Thus, the reported results are based on sample sizes of 164 (final exam performance), 155 (GPA), 128 (gender), or 121 (study time and activities). When asked their preferred exam format (MC only, MC+SA, or SA only) at the beginning of the semester, 58% of the students in this class preferred MC only, 36% preferred MC+SA, and 6% preferred SA only. This was not significantly different from the MC class ($\chi^2 = 2.117$, $p = 0.347$).

ANCOVA for GPA, Gender, and Exam Format. To control for the influence of previous student achievement and gender on final exam performance I included GPA as a covariate in a two-way (gender and exam format) ANCOVA.

RESULTS

Exam Format and Studying

Study Time for Science Classes. The students in the two sections did not differ in their reported study times for a science class in general (exam weeks: Mann-Whitney $U = 9850.5$, $p = 0.430$; non–exam weeks: Mann-Whitney $U = 9350.5$, $p = 0.213$), and there was no difference between male ($N = 125$) and female ($N = 172$) students (exam weeks: Mann-Whitney $U = 11{,}528.5$, $p = 0.280$; non–exam weeks: Mann-Whitney $U = 11{,}961$, $p = 0.092$). All students combined ($N = 323$) reported an average study time of 3.16 ± 0.098 h/wk during non–exam weeks (range: 0 min to more than 9 h), and an average of 8.39 ± 0.211 h/wk for exam weeks (range: 2 h to more than 20 h) for a science class.

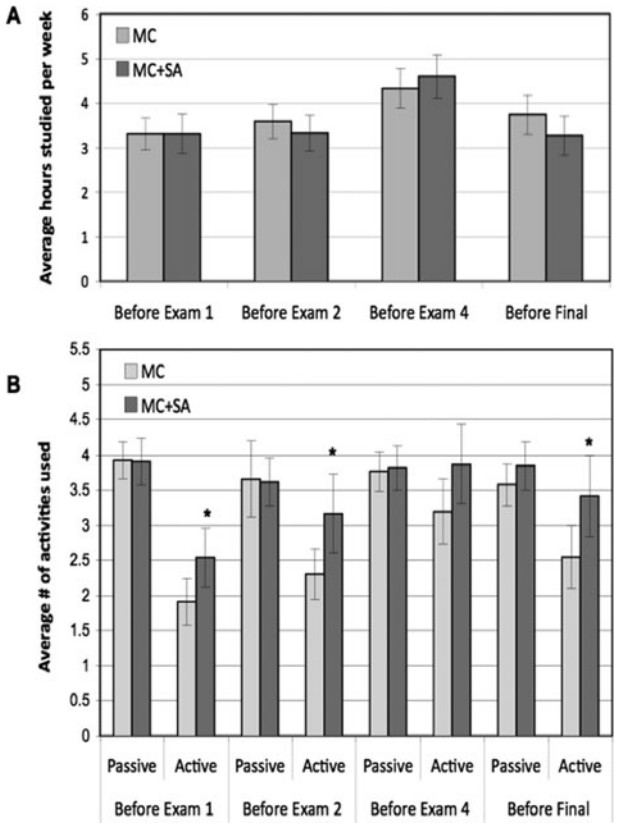

Figure 1. Study times and behaviors for introductory biology. (A) Reported study times (mean h/wk ± 2 SEM) during non–exam weeks for the introductory biology class in this study. (B) Reported number of study behaviors (mean ± 2 SEM) during non–exam weeks: cognitively passive (surface) and cognitively active (deep) learning behaviors are shown separately.

Study Time for Biology Class. In line with their other science classes, students in the MC section reported an average study time of (mean ± SEM) 3.31 ± 0.18 h/wk for the biology class in this study, and students in the MC+SA section reported on average of 3.31 ± 0.21 h/wk (Figure 1A) for the second week of class. After the first exam, students in both sections increased their study time (Table 2), but neither of these changes was significant (Table 3). Between the second and the fourth exam, students increased their weekly study time significantly to an average of 4.33 ± 0.23 h in the MC section and an average of 4.60 ± 0.24 h in the MC+SA section. Before the final exam, both sections significantly decreased their weekly study time (Table 3). There was no significant difference in study time between the two sections at any point in the semester (Figure 1A and Table 3).

Study Behavior. At the beginning of the semester, students in the MC section reported an average of 3.93 (± 0.84) cognitively passive learning behaviors (of nine options), while the students in the MC+SA section reported an average of 3.91 (±0.17) cognitively passive learning behaviors during studying for this class (non–exam week; Table 2); this was not significantly different (Mann-Whitney $U = 10{,}249$, $p = 0.824$, Table 2). This trend continued for the remainder of the semester (Figure 1B): students in the two sections did not change their cognitively passive learning behaviors

Table 2. Reported study time and study activities through the semester[a]

Change over time			Time 1	Time 2	Time 3	Time 4	Compare 1–2		Compare 2–3		Compare 3–4	
Section	*N*		Mean ± SEM	Mean ± SEM	Mean ± SEM	Mean ± SEM	*Z*	*p*	*Z*	*p*	*Z*	*p*
MC	172	Time	3.31 ± 0.18	3.59 ± 0.19	4.33 ± 0.23	3.74 ± 0.22	4,374	>0.05	5,388.5	<0.05*	2,045.5	<0.05*
		Passive	3.93 ± 0.13	3.66 ± 0.27	3.77 ± 0.14	3.58 ± 0.15	4,861	>0.05	3,699	>0.05	3,778.5	>0.05
		Active	1.91 ± 0.17	2.31 ± 0.18	3.20 ± 0.23	2.55 ± 0.22	3,915.5	<0.05*	5,966	<0.05*	2,519.5	<0.05*
MC+SA	121	Time	3.31 ± 0.22	3.33 ± 0.20	4.60 ± 0.24	3.27 ± 0.22	1,840	>0.05	3,403	<0.05*	1,444	<0.05*
		Passive	3.91 ± 0.17	3.62 ± 0.17	3.82 ± 0.16	3.85 ± 0.17	1,600	>0.05	2,644	>0.05	2,049.5	>0.05
		Active	2.54 ± 0.21	3.17 ± 0.28	3.87 ± 0.28	3.42 ± 0.29	3,159	< 0.05*	3,043.5	<0.05*	1,707	<0.05*

*Statistically significant at $p < 0.05$.

[a]Study time is reported in hours per week (mean ± SEM), study activities (passive and active) are reported as number of activities used in that week (mean ± SEM). Time 1: before exam 1; time 2: before exam 2; time 3: before exam 4; time 4: before final exam. Sample sizes (MC: 172; MC + SA: 121) are based on students that answered all four surveys. Pairwise comparisons: related samples Wilcoxon signed-rank test (*Z* values). *p* values are reported after Benjamini-Hochberg correction for multiple comparisons.

Table 3. Section differences in study time and study activities through the semester[a]

Study behavior comparison			Before exam 1		Before exam 2		Before exam 4		Before final exam	
	MC (*N*)	MC+SA (*N*)	Mann-Whitney *U*	*p*	Mann-Whitney *U*	*p*	Mann-Whitney *U*	*p*	Mann-Whitney *U*	*p*
Time	172	121	10,381	0.973	9,949.5	0.518	11,121	0.314	10,639.5	0.677
Passive	172	121	10,249	0.824	10,101	0.665	10,708.5	0.668	11,264.5	0.224
Active	172	121	11,264.5	<0.05*	11,883	<0.05*	11,752	0.057	12,001	<0.05*

*Significant at the $p < 0.05$ level.

[a]An independent sample Mann-Whitney *U*-test was used. *p* values are reported after Benjamini-Hochberg correction for multiple comparisons.

significantly (Table 2), and they did not differ significantly from one another in the number of cognitively passive learning behaviors they reported at any point in the semester (Table 3). In contrast, the reported cognitively active learning behaviors (of 13 options) increased significantly from the beginning of the semester (MC: 1.91 ± 0.172; MC+SA: 2.54 ± 0.212) to the second exam (MC: 2.31 ± 0.177; MC+SA: 3.17 ± 0.276) and from the second to the fourth exam (MC: 3.20 ± 0.226; MC+SA: 3.87 ± 0.283) in both sections (Figure 1B). Between the fourth and the final exam, there was a significant drop in cognitively active learning behaviors in both sections (Table 3). These changes in cognitively active learning behaviors remained significant (at $p < 0.05$) or marginally significant (at $p = 0.05$) after correcting for multiple comparisons ($N = 3$). Even though both sections changed their cognitively active learning behavior throughout the semester, students in the MC+SA section reported significantly more cognitively active learning behaviors than the students in the MC section in three of the four surveys (Figure 1B). All differences between the two sections remained significant (at the $p < 0.05$ level) after correcting for multiple ($N = 4$) comparisons.

Exam Format and Final Exam Performance

Total Final Exam Scores. Within each section (MC and MC+SA), students did not differ significantly in how they performed on the final exam regardless of the preference for exam format (MC only, MC+SA, SA only) they had stated at the beginning of the semester (independent samples Kruskal-Wallis test [$df = 2$] for MC section: 2.265, $p = 0.322$; for MC+SA section: 1206, $p = 0.547$). But the students in the MC+SA section scored significantly higher (67.34%) on the final exam than the students in the MC-only section (63.82%, Mann-Whitney $U = 23{,}622$, $p = 0.001$).

CR Question Scores. The MC and MC+SA sections had three CR questions in common on the final exam. The students in the MC+SA section scored significantly higher (67.27 ± 1.00) on these three CR questions (SA, fill-in table, and fill-in flowchart) than the students in the MC section (61.97 ± 0.85, Mann-Whitney $U = 24{,}540.5$, $p < 0.001$; Figure 2A and Tables 4 and 5).

MC Question Scores. Students in the MC and MC+SA sections answered 90 identical MC questions on their final exams. The students in the MC+SA section scored significantly higher (67.35%) on these 90 MC questions than the students in the MC section (64.23%, Mann-Whitney $U = 23{,}095.5$, $p = 0.005$; Figure 2A). This difference was mostly due to a significantly better performance on the higher-level MC questions: students in the MC+SA section scored significantly higher (64.4%) on the higher-level questions than the students in the MC section (59.54%; Mann-Whitney $U = 24{,}035$, $p < 0.001$). The difference between the two sections on lower level MC questions (68.76% vs. 66.46%) was marginally significant (Mann-Whitney $U = 22{,}114$, $p = 0.05$). All differences remained significant after adjustment for multiple comparisons ($N = 3$; Table 5).

Previous Student Achievement (GPA). At the beginning of the semester students in the MC section reported an average GPA of 3.3 ± 0.025, and students in the MC+SA section reported an average GPA of 3.2 ± 0.035 (Table 4). This was not significantly different (Mann-Whitney $U = 15{,}869$, $p = 0.053$, Table 5).

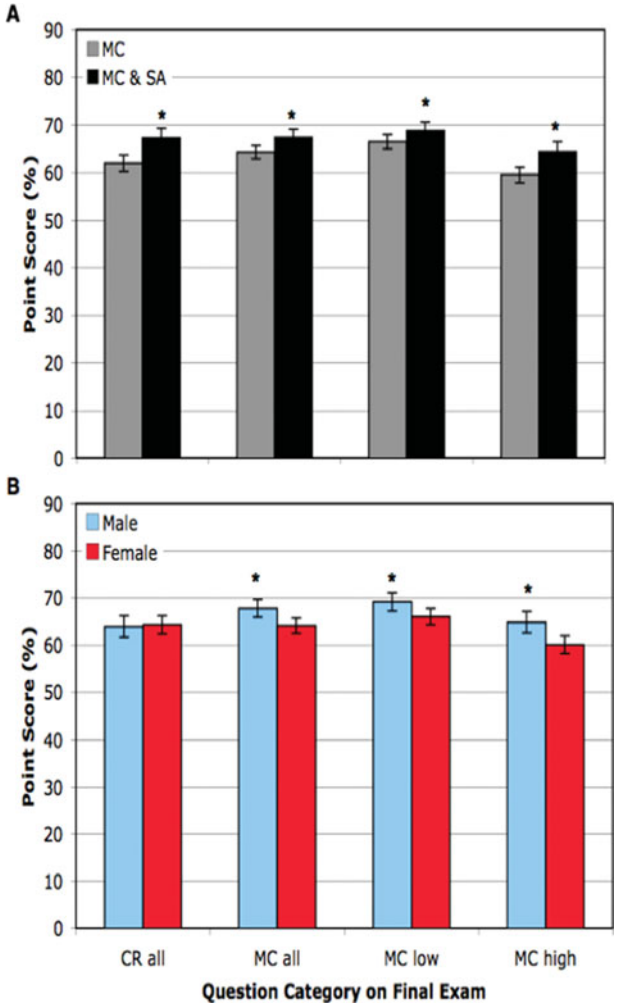

Figure 2. Student performance on the final exam (mean ± 2 SEM) by (A) exam format and (B) gender. CR all (3 CR questions), MC all (all 90 MC questions), MC low (61 lower-level MC questions), MC high (29 higher-level MC questions).

Gender. Overall, male students scored significantly higher (67.10%) than female students (64.18%, Mann-Whitney $U = 1104.5$, $p = 0.032$) on the final exam (Table 4). Male students did not significantly differ from female students in the CR questions on the final exam (63.95% vs. 64.34%, Mann-Whitney $U = 12{,}942.5$, $p = 0.893$), but they performed significantly better on both lower-level (69.19% vs. 66.05%, Mann-Whitney $U = 10{,}989$, $p = 0.027$) and higher-level (64.88% vs. 60.14%, Mann-Whitney $U = 10{,}206.5$, $p = 0.002$) MC questions (significant after adjustment for multiple comparisons: $N = 3$; Figure 2B). With reference to the overall student scores (mean ± SD: lower level = 67.39 ± 11.58; higher level = 61.5 ± 13.2), male students performed 0.273 SD units better on the lower level, and 0.359 SD units better on the higher-level MC questions than female students. These gender differences were at least partially influenced by differences in exam format. When considering the two class sections separately (Table 6), the gender differences in average scores were less pronounced (and not significant: Table 7), but in each section,

Table 4. Summary statistics of GPA and final exam performance

	GPA[a]	Final exam question category				
Exam format and gender	Mean ± SEM	FE total (%)	CR total (%)	MC total (%)	MC lower (%)	MC higher (%)
MC (242)	3.30 ± 0.025	63.82 ± 0.68	61.97 ± 0.85	64.23 ± 0.71	66.46 ± 0.74	59.54 ± 0.82
MC+SA (164)	3.20 ± 0.035	67.34 ± 0.13	67.27 ± 1.00	67.35 ± 0.87	68.76 ± 0.89	64.40 ± 1.05
Male (141)	3.25 ± 0.039	67.10 ± 0.39	63.95 ± 1.13	67.80 ± 0.93	69.19 ± 0.95	64.88 ± 1.14
Female (182)	3.29 ± 0.029	64.18 ± 0.78	64.34 ± 0.97	64.14 ± 0.82	66.05 ± 0.86	60.14 ± 0.93

[a]GPA (mean ± SEM) is based on smaller sample sizes (MC: 231; MC + SA: 155; male: 129; female: 176) of students who reported their start-of-semester GPA. Final exam performance (mean ± SEM) is reported as percent of total for the different performance categories: FE total (total final exam score), CR total (3 CR questions), MC total (all 90 MC questions), MC lower (61 lower-level MC questions), MC higher (29 higher-level MC questions).

Table 5. Exam format and gender differences in GPA and final exam performance[a]

			Final exam question category									
	GPA[b]		FE total		CR total		MC total		MC lower		MC higher	
Exam format and gender	Mann-Whitney *U*	P	Mann-Whitney *U*	*p*	Mann-Whitney *U*	*p*	Mann-Whitney *U*	*p*	Mann-Whitney *U*	*p*	Mann-Whitney *U*	*p*
Exam format	15,829	0.053	23,622	0.001*	24,541	<0.001*	23,096	0.005*	22,114	0.050*	24,035	<0.001*
Gender	11,836	0.524	11,047	0.032*	12,943	0.893	10,694	0.010*	10,989	0.027*	10,207	0.002*

*Significant at the $p < 0.05$ level, (*) marginally significant at $p = 0.05$.

[a]*p* values and associated Mann-Whitney *U* statistics are shown.

[b]GPA (mean ± SEM) is based on smaller sample sizes (MC: 231; MC+SA: 155; male: 129; female: 176) of students who reported their start-of-semester GPA. An independent sample Mann-Whitney *U*-test was used (Mann-Whitney *U* decimals were rounded up from 0.5 to 1.0). *p* values are reported after Benjamini-Hochberg correction for multiple comparisons.

Table 6. Summary statistics of GPA and final exam performance by exam format and gender[a]

	GPA	Final exam performance category			
Exam format by gender	Mean ± SEM	FE total (%)	CR total (%)	MC total (%)	MC/CR ratio
MC male (78)	3.31 ± 0.017	65.05 ± 1.16	60.54 ± 1.54	66.05 ± 1.16	1.129 ± 0.02
MC female (117)	3.34 ± 0.034	62.56 ± 1.0	62.18 ± 1.22	62.64 ± 1.04	1.036 ± 0.21
MC + SA male (63)	3.16 ± 0.066	69.63 ± 1.40	68.17 ± 1.51	69.94 ± 1.4	1.037 ± 0.17
MC + SA female (65)	3.32 ± 0.053	64.18 ± 0.78	68.23 ± 1.5	66.83 ± 1.22	1.003 ± 0.025

[a]GPA (mean ± SEM) is based on smaller sample sizes of students who reported their start-of-semester GPA. Final exam performance (mean ± SEM) is reported as percent of total for the different performance categories: FE total (final exam score), CR total (3 CR questions), MC total (all 90 MC questions), MC/CR ratio (MC%/CR% = 1.0 if students do equally well on MC and CR).

Table 7. Differences in GPA and final exam performance by exam format and gender[a]

			Final exam performance category					
	GPA		FE total		CR total		MC total	
Exam format by gender	Mann-Whitney *U*	*p*	Mann-Whitney *U*	*p*	Mann-Whitney *U*	*p*	Mann-Whitney *U*	*p*
MC by gender	4,001	0.751	4,030	0.167	48,203	0.437	3,325	0.001*
MC + SA by gender	1,867	0.655	1,750	0.156	2,025	0.914	1,558	0.020*

*Significant at the $p < 0.05$ level.

[a]An independent sample Mann-Whitney *U*-test was used (Mann-Whitney *U* decimals were rounded up from 0.5 to 1.0).

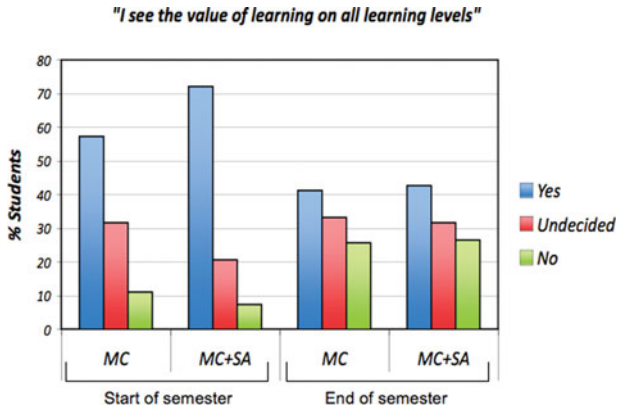

Figure 3. Student attitudes toward higher-level thinking skills. Students were asked to rate the statement "I see the value of learning on all learning levels" on a 5-point Likert scale before the first exam and before the final exam. Yes, agree or strongly agree; No, disagree or strongly disagree.

male students tended to perform better on the MC questions than the female students in the same section, but this was not the case for the CR questions.

Exam Format, GPA, Gender, and Final Exam Performance. GPA significantly influenced student performance on the final exam (ANCOVA $F = 85.0$, $p < 0.001$), explaining an estimated 22.1% of the total variance. After accounting for GPA (evaluated at 3.28), both gender ($F = 8.27$, $p = 0.004$) and exam format ($F = 29.33$, $p < 0.001$) significantly influenced final exam scores, with gender explaining 2.7% of the total variance, and exam format explaining 8.9%. The estimated marginal means of the final exam scores (after accounting for GPA) were 63.38 ± 0.704 for students in the MC-only class, compared with 69.43 ± 0.860 for students in the MC + SA class (and 68.0 ± 0.83 for male and 64.81 ± 0.731 for female students; Figure 2).

Exam Format and Student Attitudes toward Higher-Level Thinking. At the beginning of the semester (students knew their exam format but had not taken any exams yet), students in the MC + SA section tended to "see the value of learning on all learning levels" significantly more than the students in the MC section ($\chi^2 = 12.131$, $p(2) = 0.0023$). In the MC section, 57.3% of students agreed or strongly agreed that they saw the value of learning on all learning levels (Figure 3), and 11.1% of students disagreed or strongly disagreed with the statement. In the MC + SA section, 72.1% of students agreed or strongly agreed that they saw the value of learning on all learning levels, and 7.4% of students disagreed or strongly disagreed. Students in both sections changed their rating significantly between the beginning and the end of the semester (MC: Wilcoxon $Z = 1385.5$, $p < 0.001$; MC + SA: Wilcoxon $Z = 646$, $p < 0.001$). At the end of the semester, students in both sections responded very similarly ($\chi^2 = 0.196$, $p(2) = 0.907$), with an overall lower value rating for learning on all learning levels. To test the hypothesis that this may be a response to being challenged—and struggling—with critical-thinking questions in the exams of this class, I analyzed whether this change in value was related to a change in how students rated their own ability in the higher-level (critical-thinking) assessment tasks after getting feedback on exams. Across both classes, students who rated their ability in performing these higher-level tasks more highly at the end of the semester (after taking four exams) compared with the beginning (before taking any exam), also tended to rate the value of learning on all learning levels more highly than they did at the beginning of the semester. Similarly, students who rated their own ability in the higher-level (critical-thinking) tasks lower after taking four exams, also tended to rate the value of learning on all learning levels lower at the end of the semester. This resulted in a significant positive correlation between student perceptions of their own critical-thinking ability and how they valued learning on all learning levels (Spearman's rho = 0.263, $p < 0.001$, $N = 335$).

Exam Format and Student Evaluations. Even though students in the MC + SA section learned significantly more than students in the MC section, they did not like being assessed with CR questions. In the anonymous end-of-semester class evaluations, the students in the MC + SA section rated the fairness of grading in the course much lower than did the students in the MC-only section (Figure 4). In their written comments, the MC + SA students attributed their low ratings to the fact that they had to answer SA questions, while their friends in the other section did not (the different grading scales in the two sections did not make a difference to them). Students in both sections commented on the emphasis on higher-level thinking in the introductory biology class in this study. For example, some students noted that the instructor should "just teach biology" rather than emphasize higher-level thinking skills.

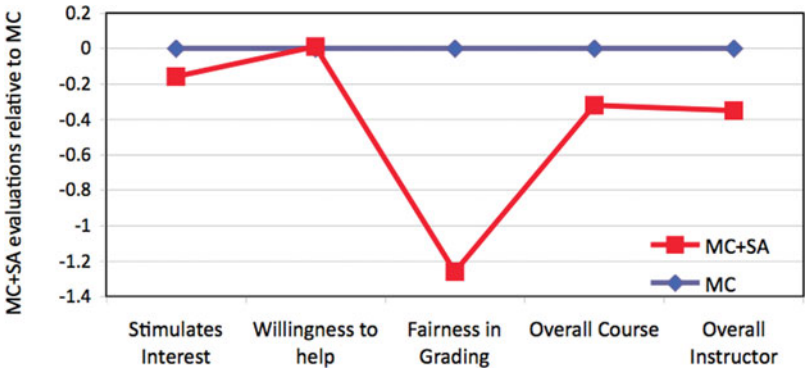

Figure 4. Student evaluations at the end of the semester. The average student evaluation scores from the MC + SA class are shown relative to the MC class (baseline).

DISCUSSION

The MC-Only Exam Format Poses an Obstacle for Critical Thinking

The purpose of this study was to assess whether an MC-only exam format might hinder the development of higher-level (critical) thinking skills in introductory science students. The answer is a convincing *yes*. The MC-only exam format seemed to undermine the instructor's efforts to convince students of the importance of critical-thinking skills, even though 25–30% of the MC questions assessed higher-level thinking. Simply knowing that they would be assessed with SA questions in addition to MC questions, significantly more students in the MC + SA section (72% vs. 57%) reported that they saw the value of learning on all learning levels at the beginning of the semester (before taking any exam). This perception was associated with a different approach to studying and a significantly better performance on the final exam. This illustrates the powerful role of perceptions of assessment in the learning process (Scouller, 1998; Watters and Watters, 2007). It is well known that students have different expectations for MC and CR exams, and as a result study differently in preparation for these exams (e.g., Scouller, 1998). However, in most previous studies that compared actual performance on MC versus CR exams, MC questions were treated as a homogeneous entity without consideration of the question level (see also Simkin and Kuechler, 2005; Kuechler and Simkin, 2010), and as a result, these studies may have compared performance on different cognitive levels rather than performance due solely to the format of the questions. In contrast, in the present study, each exam (both formats) was designed to include 25–30% higher-level thinking questions, and the students were made aware of that before taking any exams.

Interestingly, students in the MC + SA section did not study more than the students in the MC-only section. Students in both sections spent on average considerably less time studying (3 h per non–exam week) than was recommended by the instructor (2 h per hour class time or 6 h/wk). This is in line with national data: college students spend on average a total of 15 h/wk on studying, or about 7% of their time in a 5-d week (Arum and Roksa, 2011). These data also reflect national trends of declining study times in college students (Babcock and Marks, 2011): full-time college students in 1961 allocated on average 24.4 h/wk to studying, while in 2003 students spent on average 14.4 h/wk (10 h fewer).

Instead of studying more, the students in the MC + SA section used their study time more effectively for practicing higher-level thinking. Students in both sections reported a similar number of cognitively passive (surface) learning behaviors (~3.5) during studying (Figure 1), and the average number of reported cognitively active (deep) learning behaviors increased in both sections in response to their exams. This shows that students will respond with more active learning if challenged, even in the MC-only format. However, the students in the MC + SA section consistently reported more cognitively active learning behaviors in non–exam weeks (Figure 1) than the students in the MC-only section, and this difference in study behavior translated into significantly better student performance on the cumulative final exam. The somewhat puzzling decrease in study time before the cumulative final exam (Figure 1A) could be explained if most students (incorrectly and against instructions) assumed that the final exam would mostly consist of repeated questions from previous exams and were planning on memorizing the old exam questions during the week of the final exam, and/or if an extraordinary amount of semester papers and/or lab reports from other classes was due during that week, and students waited until the last minute to work on these at the cost of their final exam preparation. The significant drop in self-reported active learning behaviors during the last class week (Figure 1B) supports a shift to memorization.

The MC + SA students significantly outperformed the MC students on all final exam measures (MC and CR items, Tables 4 and 5). It could be argued that the MC students did not take the SA question too seriously, because it was an extra credit question; however, the MC + SA students significantly outperformed the MC students in the other question types (fill-in table, fill-in flowchart, and MC) as well. Most importantly, the significantly better performance of the MC + SA students on the MC questions was mostly due to a significantly better performance on the higher-level (critical-thinking) MC questions. This further supports the hypothesis that the MC-only exam format indeed discourages the practice of critical-thinking skills in introductory science classes, while the addition of CR questions encourages it.

While the change to a mixed exam format in introductory science classes requires a commitment by colleges and universities to provide adequate grading support, this investment would be a cost-effective strategy to significantly improve the critical-thinking skills of college students.

Exam Format and Student Evaluations

The clear student preference for assessment with MC questions (and student perception of MC questions being easier to answer and thus less effort to prepare for) is reflected in the assessment literature (Simkin and Kuechler, 2005). Due to the mixed exam format, mistakes in reasoning were more obvious for the students in the MC + SA section and likely contributed to their less-favorable student evaluations of both the class and the instructor at the end of the semester (Kearney and Plax, 1992; Keeley *et al.*, 1995). But even though many students in the MC + SA section disliked the experience, they learned significantly more, including critical-thinking skills, than the students in the MC-only section. This illustrates the limited use of student evaluations as a measure of actual student learning, and suggests that student ratings should not be overinterpreted, especially if students are asked to practice new thinking skills (McKeachy, 1997).

Overcoming Resistance

Student resistance to learning seems to be a common occurrence in college classrooms (Burroughs *et al.*, 1989; Kearney and Plax, 1992). For example, the comments of students in both sections that the instructor should "just teach biology," rather than emphasize thinking skills, seems to be a typical expression of student resistance to a critical-thinking challenge (e.g., Keeley *et al.*, 1995). This resistance (defined by Keeley *et al.*, 1995 as any student behavior that hinders their development into critical thinkers) was also expressed in students spending on average 50% less time studying than was recommended by the instructor, insistence on using mainly

cognitively passive study strategies, and downgrading the value of learning on all learning levels when struggling on exams.

As pointed out by Karpicke and coworkers (Karpicke *et al.*, 2009), some students may be under the illusion of competence and believe that they know the material better than they actually do when they rely purely on their subjective learning experience (e.g., their fluency of processing information during rereading and other passive study strategies). As students adjusted their own competency ratings with feedback from exams, students who downgraded their higher-level thinking skills tended to like the idea of learning on all levels less, and students who reported an increase in their higher-level thinking skills tended to value learning on all learning levels more.

Given the increased learning gains with the mixed exam format, an important question for both instructors and students is how to overcome student misgivings (e.g., Kearney and Plax, 1992) about the learning process to further maximize learning gains. In the present study, possible sources for student resistance included: 1) different exam formats in different sections, 2) expectation to practice unfamiliar thinking skills, and 3) overestimation of own critical-thinking ability. To reduce these influences, ideally all introductory science classes should implement a mixed exam format. This would not only improve student learning, but would also reduce student resistance associated with the perception of unfairness in grading due to different exam formats. In addition, all college classes should emphasize higher-level (critical) thinking skills (AAAS, 1990, 1993, 2010; Boyer Commission, 1998; NRC, 2003). This would greatly reduce student resistance to critical thinking in individual college classes. However, this is presently not the case (Crowe *et al.*, 2008; Arum and Roksa, 2011), possibly due to lingering faculty resistance toward teaching critical thinking (Haas and Keeley, 1998), and unfamiliarity of faculty with how to teach critical-thinking skills (DeAngelo *et al.*, 2009). Finally, the resistance component due to discomfort associated with facing one's own limitations (e.g., when failing to reason out an answer on an exam) could be reduced if students were trained to construct written answers with proper reasoning. In the current study, exam format accounted for 9% of the variance on the final exam performance. This occurred without practice opportunities for constructing arguments and reasoning out answers in a written format (e.g., through graded homework assignments). By adding such opportunities (requiring additional teaching assistant support) the critical-thinking gains would be expected to be even higher, while student resistance to critical thinking should be reduced. With more practice opportunities and individual feedback, students should gain competency faster, and more students should end the semester with a (realistically) higher rating of their critical-thinking skills and a more positive attitude toward higher-level learning. Ideally, combining these approaches would refocus student energies away from resisting toward practicing their higher-level thinking.

Gender Bias

A potentially troubling issue for any instructor is the possibility that exam format per se could create a performance bias beyond student achievement. In the present study, male students performed significantly better on the MC questions of the final exam than female students. An important question is whether this is an accurate measure of student achievement or whether this is due to a bias in assessment format. Research has shown that society-specific gender stereotypes predict sex differences in science performance (Nosek, 2009), and these differences in the approach to science are hard to change due to student (and school) focus on grades over engagement with the material (Carlone, 2004). In 2008, even though male and female U.S. twelfth-graders did not differ significantly in their science scores, male scores tended to be higher than female scores (NCES, 2009), and more male students (26%) scored above the proficiency level than female students (19%). Male students also tended to have completed more science courses (biology, chemistry, and physics) in high school (NCES, 2009), which has been shown to be a good predictor for science success in college (Muller *et al.*, 2001; Arum and Roksa, 2011).

As a consequence, at least some of the gender difference in the MC questions on the final exam in this study seems to be based on differences in achievement. However, if entirely due to achievement differences, the MC differences should also be reflected in the other question formats. In the present study, male students tended to perform better than female students on both assessment formats on the final exam, but they performed relatively better on the MC questions than on the CR questions, resulting in significant gender differences for MC questions only. This suggests that there may be at least some inherent bias toward male students in the MC question format (e.g., through differences in "testwiseness" [Zimmerman and Williams, 2003] and/or male students being more willing to guess than female students [Ben-Shakhar and Sinai, 2005]). Whatever the reason, the change from MC-only exam formats in introductory science classes to mixed exam formats would not only increase student learning and higher-level thinking in general, but would also remove a potential handicap for female students in introductory science classes and possibly encourage their pursuit of a career in the STEM disciplines.

ACKNOWLEDGMENTS

This work was supported by a University of Georgia System Board of Regents STEM grant, which funded the hiring and training of a graduate student teaching assistant for grading the exams in the MC+SA section. This study was conducted under the guidelines of IRB # 2007-10197. I thank Tom Koballa for advice in the planning phase of this study, Julie Palmer for inspiring the essay exercise, and the UGA Science Education Research Group for feedback on an earlier version of the manuscript. This is a publication of the UGA Science Education Research Group.

REFERENCES

Allen DE, Duch BJ, Groh SE (1996). The power of problem-based learning in teaching introductory science courses. New Dir Teach Learn *68*, 43–52.

American Association for the Advancement of Science (AAAS) (1990). Science for All Americans, New York: Oxford University Press.

AAAS (1993). Benchmarks for Science Literacy, New York: Oxford University Press.

AAAS (2010). Vision and Change. A Call to Action, Washington, DC.

Anderson LW, Krathwohl DR (eds.) (2001). A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, New York: Addison Wesley Longman.

Arum R, Roksa J (2011). Academically Adrift. Limited Learning on College Campuses, Chicago: University of Chicago Press.

Babcock P, Marks M (2011). The falling time costs of college: evidence from half a century of time use data. Rev Econ Stat *93*, 468–478.

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B *57*, 289–300.

Ben-Shakhar G, Sinai Y (2005). Gender differences in multiple-choice tests: the role of differential guessing tendencies. J Educ Meas *28*, 23–35.

Biggs JB (1973). Study behavior and performance in objective and essay formats. Australian J Educ *17*, 157–167.

Bloom BS (1956). Taxonomy of Educational Objectives: The Classification of Educational Goals, Chicago: Susan Fauer.

Bolger N, Kellaghan T (1990). Method of measurement and gender differences in scholastic achievement. J Educ Meas *27*, 165–174.

Boyer Commission on Educating Undergraduates in the Research University (1998). Reinventing Undergraduate Education: A Blueprint for American Research Universities, Stony Brook, NY.

Burroughs NF, Kearney P, Plax TG (1989). Compliance-resistance in the college classroom. Communication Educ *38*, 214–229.

Caldwell J (2007). Clickers in the large classroom: current research and best-practice tips. CBE Life Sci Educ *6*, 9–20.

Carlone HB (2004). The cultural production of science in reform-based physics: girls' access, participation, and resistance. J Res Sci Teach *41*, 392–414.

Crouch CH, Mazur E (2001). Peer instruction: ten years of experience and results. Am J Phys *69*, 970–977.

Crowe A, Dirks C, Wenderoth MP (2008). Biology in Bloom: implementing Bloom's taxonomy to enhance student learning in biology. CBE Life Sci Educ *7*, 368–381.

DeAngelo L, Hurtado S, Pryor JH, Kelly KR, Santos JL (2009). The American College Teacher: National Norms for the 2007–2008 HERI Faculty Survey. Research Brief, Los Angeles: Higher Education Research Institute, University of California, Los Angeles.

DeMars DE (1998). Gender differences in mathematics and science on a high school proficiency exam: the role of response format. Appl Meas Educ *11*, 279–299.

Dufresne RJ, Leonard WJ, Gerace WJ (2002). Making sense of students' answers to multiple-choice questions. Phys Teach *40*, 174–180.

Eberlein T, Kampmeier J, Minderhout V, Moog RS, Platt T, Varma-Nelson P, White HB (2008). Pedagogies of engagement in science. A comparison of PBL, POGIL, and PLTL. Biochem Mol Biol Educ *36*, 262–273.

Ebert-May D, Brewer C (1997). Innovative in large lectures. Teaching for active learning. Bioscience *47*, 601–607.

Gosser DK, Jr, Roth V (1998). The Workshop Chemistry Project: peer-led team learning. J Chem Educ *75*, 185–187.

Haas PF, Keeley SM (1998). Coping with faculty resistance to teaching critical thinking. Coll Teach *46*, 63–67.

Handelsman J *et al.* (2004). Scientific teaching. Science *304*, 521–522.

Handelsman J, Miller S, Pfund C (2007). Scientific Teaching, New York: W.H. Freeman.

Karpicke JD, Butler AC, Roediger HL (2009). Metacognitive strategies in student learning: do students practice retrieval when they study on their own? Memory *17*, 471–479.

Kearney P, Plax TG (1992). In: Student Resistance to Control Power in the Classroom: Communication, Control and Concern, ed. VP Richmond and JC MacCroskey, Hillsdale, NJ: Lawrence Erlbaum, 85–100.

Keeley SM, Shemberg KM, Cowell BS, Zinnbauer BJ (1995). Coping with student resistance to critical thinking. What the psychotherapy literature can tell us. Coll Teach *43*, 140–145.

Kuechler WL, Simkin MG (2010). Why is performance on multiple-choice tests and constructed-response tests not more closely related? Theory and an empirical test. Decision Sci J Innovative Educ *8*, 55–73.

Martinez ME (1999). Cognition and the question of test item format. Educ Psychol *34*, 207–218.

McKeachy WJ (1997). Student ratings: the validity of use. Am Psychol *52*, 1218–1225.

Momsen J, Long TA, Wyse SA, Ebert-May D (2010). Just the facts? Introductory undergraduate biology courses focus on low-level cognitive skills. CBE Life Sci Educ *9*, 435–440.

Moog RS, Spencer JN (2008). POGIL: An overview. In: Process-Oriented Guided Inquiry Learning (POGIL), ed. RS Moog and JN Spencer. American Chemical Society, ACS Symposium Series, Vol. 994, 1–13.

Muller PA, Stage FK, Kinzie J (2001). Science achievement growth trajectories: understanding factors related to gender and racial-ethnic differences in precollege science achievement. Am Educ Res J *38*, 981–1012.

National Center for Education Statistics (2009). The Nation's Report Card 2009, Washington, DC: Institute of Education Sciences of the U.S. Department of Education. www.nationsreportcard .gov/science_2009/science_2009_report (accessed 12 May 2011).

National Research Council (2003). BIO2010: Transforming Undergraduate Education for Future Research Biologists. Committee on Undergraduate Biology Education to Prepare Research Scientists for the 21st Century, Washington, DC: National Academies Press. www.nap.edu/catalog/10497.html (accessed 12 May 2011).

Nosek BA (2009). National differences in gender—science stereotypes predict national sex differences in science and math achievement. Proc Natl Acad Sci USA *106*, 10593–10597.

Paul RW, Elder L, Bartell T (1997). California Teacher Preparation for Instruction in Critical Thinking: Research Findings and Policy Recommendations, Sacramento, CA: Foundation for Critical Thinking.

Pfund C *et al.* (2009). Summer institute to improve university science teaching. Science *324*, 470–471.

Prince M (2004). Does active learning work? J Eng Educ *93*, 223–231.

Scouller K (1998). The influence of assessment method on students' learning approaches: multiple choice question examination versus assignment essay. Higher Educ *35*, 453–472.

Simkin MG, Kuechler WL (2005). Multiple choice tests and student understanding: what is the connection? Decis Sci J Innovative Educ *3*, 73–97.

Smith MK, Wood WB, Adama WK, Wieman C, Knight JK, Guild N, Su TT (2009). Why peer discussion improves student performance on in-class concept questions. Science *323*, 122–124.

SPSS (2011). Statistical Analysis, Version 19.0, Chicago: SPSS.

Stanger-Hall KF, Lang S, Maas M (2010). Facilitating learning in large lecture classes: testing the teaching teams approach to peer learning. CBE Life Sci Educ *9*, 489–503.

Stanger-Hall KF, Shockley FW, Wilson RE (2011). Teaching students how to study: a workshop on information processing and self-testing helps students learn. CBE Life Sci Educ *10*, 187–198.

Tanner K, Allen D (2005). Approaches to biology teaching and learning: understanding the wrong answers-teaching toward conceptual change. Cell Biol Educ *4*, 112–117.

Watters DJ, Watters JJ (2007). Approaches to learning by students in the biological sciences: implications for teaching. Int J Sci Educ *29*, 19–43.

Weimer ME (2002). Learner-Centered Teaching. Five Key Changes to Practice, San Francisco: Jossey-Bass.

Zeidner M (1987). Essay versus multiple-choice type classroom exams: the student perspective. J Educ Res *80*, 352–358.

Zheng AY, Lawhorn JK, Lumley T, Freeman S (2008). Assessment-application of Bloom's taxonomy debunks the "MCAT myth." Science *319*, 414–415.

Zimmerman DW, Williams RH (2003). A new look at the influence of guessing on the reliability of multiple choice tests. Appl Psychol Meas *27*, 357–371.