CBE—Life Sciences Education
Vol. 15, 1–16, Summer 2016

Article

Cognitive Difficulty and Format of Exams Predicts Gender and Socioeconomic Gaps in Exam Performance of Students in Introductory Biology Courses

Christian D. Wright,[†‡] Sarah L. Eddy,[§‡] Mary Pat Wenderoth,[‖] Elizabeth Abshire,[‖] Margaret Blankenbiller,[‖] and Sara E. Brownell[†*]

[†]School of Life Sciences, Arizona State University, Tempe, AZ 85287; [§]Texas Institute for Discovery Education in Science, University of Texas at Austin, Austin, TX 78705; [‖]Department of Biology, University of Washington, Seattle, WA 98195

Submitted December 3, 2015; Revised February 10, 2016; Accepted: February 16, 2016
Monitoring Editor: Debra Tomanek

Recent reform efforts in undergraduate biology have recommended transforming course exams to test at more cognitively challenging levels, which may mean including more cognitively challenging and more constructed-response questions on assessments. However, changing the characteristics of exams could result in bias against historically underserved groups. In this study, we examined whether and to what extent the characteristics of instructor-generated tests impact the exam performance of male and female and middle/high- and low-socioeconomic status (SES) students enrolled in introductory biology courses. We collected exam scores for 4810 students from 87 unique exams taken across 3 yr of the introductory biology series at a large research university. We determined the median Bloom's level and the percentage of constructed-response questions for each exam. Despite controlling for prior academic ability in our models, we found that males and middle/high-SES students were disproportionately favored as the Bloom's level of exams increased. Additionally, middle/high-SES students were favored as the proportion of constructed-response questions on exams increased. Given that we controlled for prior academic ability, our findings do not likely reflect differences in academic ability level. We discuss possible explanations for our findings and how they might impact how we assess our students.

INTRODUCTION

Student performance on course exams is one of the primary ways that introductory biology instructors evaluate student understanding and determine course grades, which in turn impacts a student's ability to pass a course, overall science grade point average (GPA), and, ultimately, persistence as a biology major. Although we expect to see a range of student performance based on academic ability, systematic differences in how different populations of students perform on exams potentially contributes to the unequal retention of different demographic groups in biology.

For most introductory biology courses, instructors write their own course exams and decide the format of the exam (e.g., multiple choice or short answers), the topics assessed (e.g., photosynthesis or phylogenies), and the level to test student understanding (e.g., memorization of a definition or interpretation of an experiment). Although exams can be characterized in many different dimensions, in this paper, we will focus on cognitive level (Bloom *et al.*, 1956; Anderson *et al.*, 2001) and the format of exam questions. Although there are few studies exploring biology instructor decision making on developing course exams, there is evidence that

CBE Life Sci Educ June 1, 2016 15:ar23
DOI:10.1187/cbe.15-12-0246
[‡]These authors contributed equally.

*Address correspondence to: Sara E. Brownell (sara.brownell@asu.edu).

instructors are not using cognitively challenging exam questions (Momsen *et al*., 2010, 2013) and that a large number of college instructors report they are using multiple-choice tests (DeAngelo *et al*., 2009).

In an effort to promote deeper student conceptual understanding, recent reform efforts in undergraduate biology have recommended transforming our exams to test at more cognitively challenging levels (e.g., American Association for the Advancement of Science, 2011). One way of evaluating the cognitive level of exams and their questions is to categorize each question according to Bloom's taxonomy of cognitive domains (Bloom *et al*., 1956; Anderson *et al*., 2001; Crowe *et al*., 2008). There is evidence that higher Bloom's-level exams have the potential to move students from superficial to deep conceptual understanding (Black and Wiliam, 1998; Stanger-Hall, 2012; Jensen *et al*., 2014). For example, Jensen and colleagues (2014) found that college biology students adapt their learning to the level of their exams. When students were tested at only the memorization level on high-stakes assessments, even when given the opportunity to practice cognitively more challenging questions in class, they did not develop higher-order cognitive skills. Students developed these higher-order skills only when they completed high-stakes assessments that reinforced the high cognitive level of classroom practice. Similarly, McDaniel and colleagues (2013) found that middle school science students scored higher on more cognitively challenging exam questions when previously quizzed with this level of questions compared with when they were either quizzed with low-level questions or not quizzed at all.

Further, the format of exam questions also influences how students engage with course material. When students expect a test to contain constructed-response questions (e.g., questions for which students must generate the response: short answer, essay, graphing/drawing), they tend to take a deeper approach to learning and take notes that concentrate more on main ideas and core concepts (Rickards and Friedman, 1978; Thomas and Bain, 1984; Entwistle and Entwistle, 1991). Conversely, students tend to approach learning superficially when they expect tests to have restricted-response questions (e.g., questions for which students choose from a set of answers: multiple choice, true/false) or questions that test lower-order thinking.

Thus, it appears that more cognitively challenging questions on exams and more constructed-response questions on exams have the potential to move students toward developing deeper conceptual understanding. However, incorporating more cognitively challenging and constructed-response questions may have unintended consequences for historically underserved groups in science, technology, engineering, and mathematics classrooms. A number of K–12 studies have documented that, even after controlling for a measure of academic ability, changing the characteristics of exams differentially impacted the achievement of students of different genders and races/ethnicities. Increasing the cognitive complexity of assessments favored the performance of male over female students at the middle and high school level (Carlton and Harris, 1992; Harris and Carlton, 1993; Lindberg *et al*., 2010; but see Bastick, 2002). Increasing the cognitive level of exams also favored white high school students over students of other racial/ethnic/nationality identities (Carlton and Harris, 1992). Exam format can also have a differential impact on students, as many studies have demonstrated that males tend to perform better on restricted-response questions and women tend to perform better on constructed-response questions on general science and math assessments administered at the K–12 level (Mazzeo *et al*., 1993; DeMars, 1998, 2000; Beller and Gafni, 2000; Lindberg *et al*., 2010; but see DeMars, 1998; Beller and Gafni, 2000; Neuschmidt *et al*., 2008; Le, 2009; Lindberg *et al*., 2010). Additionally, restricted-response questions tend to favor the performance of white students, while constructed-response questions tend to favor the performance of students of other racial/ethnic/national identities at the K–12 level (Taylor and Lee, 2011; but see DeMars, 2000).

There have also been college-level studies exploring the impact of exam characteristics on student performance in mathematics (Ryan and Chiu, 2001; Lindberg *et al*., 2010), atmospheric and oceanic sciences (Weaver and Raptis, 2001), and biology (Migliaccio and Sheikh, 2009; Stanger-Hall, 2012). Generally, these studies have shown somewhat inconsistent patterns regarding the impact of the cognitive level of questions on the performance of males and females. Some studies have found that the performance of male students on math and biology assessments was favored only on questions testing higher cognitive thinking (Ryan and Chiu, 2001; Migliaccio and Sheikh, 2009), whereas Stanger-Hall (2012) found that male students are favored on biology questions testing both lower and higher levels of cognitive thinking. However, Lindberg *et al.* (2010) showed that the performance of male students was favored on less cognitively challenging questions in math, but that there was a slight performance gap that favored females on more cognitively challenging questions. Similar inconsistencies have been observed with respect to format. Ryan and Chiu (2001) found that math word problems differentially favored male students. Conversely, Weaver and Raptis (2001) showed that, on atmospheric and oceanic sciences exams, male students performed better on restricted-response questions, with females doing better on constructed-response questions. Stanger-Hall (2012) observed no differences in performance between males and females on constructed-response questions in biology, but males were favored on restricted-response questions. These inconsistencies may be due to small sample sizes, and they reflect the need for more large-scale studies. Further, to our knowledge, no studies have documented how the format and cognitive-challenge level of assessment questions impacts the performance of students from different socioeconomic backgrounds.

To address this gap in the literature, we take the first step of conducting a large-scale analysis of multiple introductory biology courses to see whether and to what extent characteristics of instructor-generated summative assessments administered in introductory biology classrooms differentially impact students of different genders and different socioeconomic backgrounds.

> Prediction 1: We predict that increasing the Bloom's level of instructor-generated undergraduate biology exams will disproportionately favor the performance of males and students from middle/high-socioeconomic status (SES) backgrounds over females and students from low-SES backgrounds, respectively.

Prediction 2: Increasing the percentage of constructed-response questions on instructor-generated undergraduate biology exams will disproportionately favor the performance of female and low-SES students over male and middle/high-SES students, respectively.

METHODS

We collected data from 26 instructors teaching introductory biology classes across a time span of 3 yr. We then characterized the Bloom's level, difficulty, and format of each exam (described in more detail below in the section titled *The Exams*). Our study has a quasi-random design, as students select the classes they attend. The lack of truly random study design means that there is the potential for student exam performance to be influenced by factors that we have not measured and/or that are outside our variables of interest. To minimize the confounding variables in our study, we included variables to account for the differences between 1) the 25 classes in our sample due to instructors and content and 2) the students in our sample. We included these variables as fixed and random effect variables in a linear mixed-effects model. Below in the sections titled *The Classes*, *The Students*, *The Exams*, and *General Description of Statistical Analyses*, we outline the variation seen at each of these levels and the variables we used to control for variation at these levels.

The Classes

We examined 25 individual classes of the three-course introductory biology sequence for majors over a 3-yr period at a large public R1 university on the quarter system. Fifteen of these classes (60%) were cotaught by different instructors, each teaching for a 5-wk period. Therefore, we were able to collect exams from 26 different instructors. The three courses are intended to introduce students to the breadth of biology: the first course in the series introduces ecology and evolution, the second course focuses on molecular and cellular biology, and the third course explores plant and animal physiology. The introductory biology series must be completed in the aforementioned order, with enrollment in each subsequent course contingent on completing the previous course(s) in the introductory biology series. The individual classes in our data set ranged in size from 159 to more than 900 students. A study that recently investigated the teaching strategies of the classrooms used in our study documented that instructional practices varied between instructors, ranging from almost entirely traditional lecture to more student-centered interactive methods (Eddy *et al.*, 2015).

Accounting for Differences among Instructors and Courses. Variation in the course environment due to differences among instructors in our data set has the potential to influence student performance. For example, differences in the gender of the instructor can impact the achievement gap between male and female students in a course (Carrell *et al.*, 2010; Eddy *et al.*, 2014). In addition, differences in the instructional practices that instructors use have also been demonstrated to impact achievement (Wenglinsky, 2002; Freeman *et al.*, 2011, 2014; Haak *et al.*, 2011; Eddy and Hogan, 2014). To account for variation in the course environment due to differences among instructors that are not accounted for by the exam characteristics, we included a random-effect term for individual instructors in our model. We also included course as a fixed effect in our model to account for differences in the topics taught in each of the courses.

The Students

Students who enrolled in the introductory series were predominantly sophomore biology majors. Across students in our sample, 58% (n = 2790) were female, while 42% were male (n = 2020), but the proportion of females to males varied among classes. Only students who identified their gender were included in the analyses. In addition, 44% (n = 2094) of the students identified as white/Caucasian, 37% (n = 1765) as Asian, 2% (n = 120) as black/African American, 1% (n = 43) as Hawaiian/Pacific Islander, 5% (n = 255) as Hispanic/Latin@, 7% (n = 327) as international students, and 1% (n = 56) as Native American. Approximately 3% (n = 150) in our sample did not list a racial/ethnic/national identity. We should note that the gender and racial/ethnic/national categories represent university-designated groups and do not fully reflect the spectrum and complexities of social identities. The institution also identified students who came from educationally or economically disadvantaged backgrounds, and these students were eligible to participate in the Educational Opportunity Program (EOP). The impact(s) of family income, level of parental education, and or social/environmental barriers on academic success were taken into consideration when the institution determined whether a student qualified for the program. In our study, 17% (n = 817) of students were classified as eligible for the program, while 83% (n = 3993) of students were classified as not eligible for the program. For our analyses, eligibility for the EOP program served as a proxy for SES, and we will refer to students eligible for the EOP program as students from a low-SES background (similar to Freeman *et al.*, 2007).

In addition to demographic variables, we also collected a measure of general ability in college: cumulative GPA at time of entry into the introductory biology series. Before entering the introductory biology series, most of the students took ~45 total credits during their freshman year. Sophomore biology majors in our study were required to complete the following courses before entering into the three-course introductory biology series: three courses of general chemistry, three courses of calculus, and one English composition course, with the remaining credits being filled by general education credits (e.g., sociology, drama, geography, art). Cumulative GPA has been shown to strongly predict the grade of students at the end of introductory courses for this population (Freeman *et al.*, 2007, 2011). The mean cumulative GPA for students in the series was 3.21 (on a 4.0 scale), the median cumulative GPA was 3.27, and the GPA of all students included in this study ranged from 0.55 to 4.0.

Accounting for Differences among Students. To determine which demographic variables to include in our models, we examined our data to determine whether and to what extent our demographic variables of interest (gender, SES status, and race/ethnicity/nationality) were correlated with one another. We found that students who were from low-SES backgrounds also tended to be underserved minorities (3% of all white/Caucasian students, 19% of all Asian

students, 0% of all international students, 90% of all black/African-American students, 84% of all Hawaiian/Pacific Islander students, 88% of all Hispanic/Latin@ students, and 79% of all Native American students were identified as low-SES students; Table A, Supplemental Material). Furthermore, 19% of all females and 14% of all males were identified as low-SES students (Table A, Supplemental Material). Because underserved minorities tended to be low-SES students and because the sample size for students identifying as certain racial/ethnic/national identities was small, we chose not to test for the impact of race/ethnicity/nationality in our study. Rather, we included interaction terms for both gender and SES status with exam characteristics in our models.

To account for differences in student preparedness and academic ability among students in our models, we included two covariates: a fixed-effect term for cumulative college GPA as a proxy for differences in academic ability and a random-effect term for student identity to account for overall differences among students not accounted for by cumulative GPA or other variables in our models. We chose to use cumulative incoming GPA to control for prior academic ability in our models, because previous work (Freeman *et al.*, 2007) and our own preliminary analyses indicate that cumulative incoming GPA is the strongest predictor of performance in these biology courses. In addition, including student identity as a random effect in our model allowed us to account for repeated measures on the same students (each student is represented at least four times in the data set: once for each of the four exams they took in a term, although some students took multiple courses in the introductory series), avoiding potential issues of pseudoreplication. Additionally, including student as a random-effects term in our models allowed us to account for differences in students' prior experiences not included in our models. For example, work has shown that first-year undergraduate biology students who completed first-year biology courses taught using learner-centered teaching strategies showed higher performance on content knowledge assessments taken their senior year compared with students who completed more traditional, unrevised courses (Derting and Ebert-May, 2010). A random-effects term for student allowed us to account for these and other differences in students' prior experiences not included in our analyses.

Given the covariates included in our model, our model outputs are describing the differences between men and women or between students from lower- or middle/high-SES backgrounds who entered the biology series with equal academic ability based on their prior courses and who are experiencing the same course environment (i.e., enrolled in the same course with the same instructor).

The Exams

For each question on each exam, we determined the Bloom's level, the format, and the difficulty of the question. Because the exam score data for students were at the level of the whole exam rather than individual questions, we pooled the exam characteristic data across exam questions for each exam to produce an exam-level measure of each of these three variables. Below in the sections titled *Determining Exam Characteristics* and *Controlling for Additional Exam Characteristics*, we outline the specific methods used to determine each exam characteristic.

Determining Exam Characteristics: Weighted Bloom's Index. Using the technique described by Crowe *et al.* (2008), two raters, each of whom had a bachelor's degree in biology and had served as a teaching assistant in introductory biology classes at the college level, independently determined the Bloom's level of all of the exam questions. This process began with a third person collecting exams from all the instructors and creating a randomized list of all the exam questions. This guaranteed that the raters were blind to which instructor wrote each question. Raters normed on a set of 50 questions and received expert feedback and advice on Blooming from two of the authors of Crowe *et al.* (2008). Observers then individually assigned a Bloom's level of knowledge, comprehension, application, analysis, synthesis, or evaluation to each question. On completion of their individual scoring, they discussed the scores, and when they disagreed, they came to consensus on the Bloom's level of each question. After consensus was reached, the categories for Bloom's level were collapsed from six to three levels: high (synthesis and evaluation), medium (application and analysis), and low (knowledge and comprehension). For any question with multiple subparts (e.g., a question with part a and part b), we assigned the median Bloom's level across all the subparts.

To create an aggregate Bloom's score for each exam, we followed the methods for creating a weighted Bloom's index, described in Freeman *et al.* (2011):

$$\left(\frac{\sum_1^n p \times B}{T \times 3}\right) \times 100 = \text{weighted Bloom's median}$$

where n is the number of questions, p is points per question, B = Bloom's rank (1, 2, or 3 for low, medium, and high Bloom's level, respectively) for that question, T is the total points possible, and 3 is the maximum possible Bloom's score. Ultimately, for each exam, the weighted Bloom's level for each test was converted to a score on a scale of 0.33–1, with 0.33 being the lowest possible weighted Bloom's level (i.e., what an exam with only low-level questions would earn) and 1 being the highest Bloom's level possible (i.e., what an exam that had all high-level questions would earn). We then calculated the median weighted Bloom's level for each exam.

Determining Exam Characteristics: Constructed-Response versus Restricted-Response Questions. Two raters recorded the format of each item, identifying each question as constructed-response (e.g. short answer, fill in the blank, essay, graphing, or drawing questions) or restricted-response (e.g. multiple choice, true/false, multiple true/false, or matching). Observers came to consensus on question format. Percent of questions that were constructed-response (percent CR) on each exam was then determined. Ultimately, for each exam, the percentage of questions on each test that were constructed-response was converted to a scale of 0–1, with 0 being a test that consists of entirely restricted-response questions and 1 being a test that consists entirely of constructed-response questions.

Controlling for Additional Exam Characteristics: Weighted Difficulty Index. Although we were primarily interested in Bloom's level and exam format, we needed to control for additional exam characteristics that might be

correlated with these variables of interest. One such variable could be question difficulty. Question difficulty is a gauge of how easy or hard a question may be. In general, both low- and high-performing students tend to correctly answer questions that are considered to be "easy," while "harder" questions are defined as those questions that only high-performing individuals tend to answer correctly (de Ayala, 2009). Increased difficulty can be positively correlated with questions that test higher Bloom's level of thinking, although it is possible to require students to memorize an obscure fact that makes a question difficult (e.g., Freeman *et al.*, 2011; but see Momsen *et al.*, 2013). Given the potential correlation between difficulty and Bloom's level, we determined the difficulty of each of the exams used in our analysis.

To determine the characteristics related to item difficulty, we consulted the research literature as well as instructors ($n = 3$) and teaching assistants ($n = 6$) who had multiple terms of experience teaching in the introductory biology series. In these intro courses, teaching assistants grade the exams and have a strong contextual background to make inferences about the types of questions that are challenging for students. In consultation with these sources, we developed a list of item characteristics that, in the experience of instructors and teaching assistants, generally lead to lower student performance on exam items in the introductory biology series at our institution. The intent of the difficulty measure was to capture elements of a question not included in Bloom's level that could impact performance, including features like the reading load of a question and how challenging the topic was for students in general. Thus, we eliminated any characteristics that seemed related to cognitive processes captured by Bloom's taxonomy. The final list can be found in Table B in the Supplemental Material.

In addition to compiling a list of characteristics for raters to consider as they looked at the exam questions, we adapted the methods of Freeman *et al.* (2011) for scoring the difficulty of an item: we asked teaching assistants to determine the percent of the class who would get the item correct (for questions scored as right/wrong) or the number of points the average student would score on an item. We modified this to a three-point scale (easy, medium, hard) by looking at the distribution of student performance on one term of exams for which we had item-level performance data. We divided the distribution of item scores into thirds. The hardest third of items (a score of 3) were those items that 60% or fewer students correctly answered or, for items scored with partial credit, items for which students earned less than half the total possible points. The easiest third of the items (a score of 1) were considered those items that 80% or more of the students correctly answered or, for items scored with partial credit, items for which students earned 75% or more of the total possible points. See Table B in the Supplemental Material for the final difficulty tool.

We recognize that this difficulty metric is not a validated instrument. It was not intended to be used across institutions but instead was developed from experiences in these particular class contexts using methods used in other studies with our specific students.

Four raters determined the difficulty of exam questions across the three courses in the introductory series. Following the suggestions of Freeman *et al.* (2011), we selected raters that had been teaching assistants for multiple iterations of the classes they were scoring. They had graded multiple exams and had a strong contextual background to make inferences about the types of questions that are challenging for students. As with Bloom's levels, raters scored question difficulty using a randomized list of all of the questions so they were blind to the instructor who wrote each question. Raters came to consensus on the difficulty level for each question. If a question had multiple subparts, then the median difficulty of the subparts was used.

To create an aggregate difficulty measure for each exam, we determined a weighted difficulty index by using methods similar to Freeman *et al.* (2011) and the equation described above, with difficulty level replacing Bloom's rank. Ultimately, for each exam, the weighted difficulty for each test was converted to a score on a scale of 0.33–1, with 0.33 being the lowest possible difficulty (i.e., the raters predicted 80% or more of students would get it correct) and 1 being the highest difficulty possible (i.e., the raters predicted < 60% of students would get it correct). We then calculated the median weighted difficulty for each exam.

Determining Which Exam Characteristics to Include in Our Model. To develop the fixed effects of our baseline models, we first examined the correlation matrix between the three exam characteristic measures to determine which variables, if any, were correlated with one another (Table C, Supplemental Material). We found that the percent CR of an exam was moderately to strongly (0.47) correlated with the median weighted Bloom's level of an exam. Given the level of correlation between Bloom's level and percent CR, we chose to run separate regression analyses on these two exam characteristics. However, we found only small correlations between Bloom's level and exam difficulty as well as between percent CR and exam difficulty. We therefore chose to include difficulty as a covariate in our baseline models. Thus, when we describe our model outputs, we are describing differences between males and females or between students from lower or middle/high-SES backgrounds who took exams with the same level of difficulty.

In addition to difficulty, we also accounted for when an exam was given during a particular course. We used exam number as a proxy for time in the course, as student performance on an exam may be influenced by the duration of their learning in a course, and thus was included as a fixed effect in our analyses.

General Description of Statistical Analyses

Response Variable. The response variable for our analyses was overall student performance on each exam, which was measured as a percent of exam points earned. Because our percentage score data were not continuous (scores were limited to 0–100%, or 0–1 when converted to proportional data), we transformed our data using an arcsine transformation, which consists of taking the arcsine of the square root of the exam scores. Percentage data, like proportions, have a binomial distribution rather than a normal distribution, with the largest deviations of normality occurring for scores less than 30% or greater than 70%. Arcsine transformations, which are commonly applied to percentage and proportional data, produce data that have a nearly normal underlying distribution (Zar, 2010), thus meeting

the assumptions of the linear regression analyses that we performed.

Regression Analyses. We used a hierarchical modeling approach, because students who experienced courses taught by the same instructor and took exams constructed by the same instructor likely had scores that were more comparable to the scores of other students whose courses were taught by and whose exams were constructed by the same instructor than to exam scores of students whose courses were taught by and whose exams were constructed by a different instructor, even within the same course. Additionally, we have fixed-effect terms at both the student level (gender identity, SES, and cumulative college GPA coming into the course) and the instructor/course/exam level (course, exam characteristics, exam number). Because of the hierarchical nature of the data set, we used multilevel modeling for our statistical analyses, which is a common approach used in a wide array of fields (e.g., Paterson and Goldstein, 1991; Kreft and de Leeuw, 2002; Raudenbush and Bryk, 2002; Zuur *et al*., 2009; Eddy *et al*., 2014). See Zuur *et al.* (2009) for a detailed description of multilevel modeling to account for hierarchally nested data sets.

We then performed preliminary analyses on the impact of each of the fixed effects and potential interactions on students' arcsine-transformed exam scores. Fixed-effect terms that independently had a significant impact on our response variable were included in our baseline models. Using the information from the correlation matrices and our preliminary analyses, we generated the following full model for our analyses:

$$\text{Exam Performance} = \text{Course} + \text{Time} + \text{Cum.GPA} + \text{Gender} + \text{SES} + \text{W.Diff} + \text{Exam Characteristic} + \text{Gender} * \text{Exam Characteristic} + \text{SES} * \text{Exam Characteristic} + (1 \mid \text{Stu.ID}) + (1 \mid \text{Instr})$$

where 1) "Course" represents the three different introductory courses (a categorical variable with three levels) used in this analyses, 2) "Time" is the exam number (a categorical variable with four levels), 3) "Cum.GPA" is the cumulative college GPA upon entering the introductory biology sequence (a continuous variable ranging from 0 to 4), 4) "Gender" is the student's gender identity (constrained to a binary: male, female), 5) "SES" is represented by proxy for student SES: eligibility for the Educational Opportunities Program (constrained to a binary: middle/high-SES and low-SES), 6) "W.Diff" is the weighted median difficulty of an exam (a continuous variable ranging from 0.33 to 1), 7) "Exam Characteristics" represent either a) the weighted median Bloom's level of an exam (W.Blooms; a continuous variable ranging from 0.33 to 1) or b) the percentage of constructed-response questions on an exam (percent CR; a continuous variable ranging from 0 to 1). We also included interaction terms for both gender and SES with W.Blooms and percent CR. Finally, our models included the random-effects terms for student identity (1|Stu.ID) and instructor (1|Instr) to account for the nested nature of our data set, specifically the repeated measures on students and the fact that students are nested within instructors' courses.

We included only students who had a complete set of all of the aforementioned variables. We generated two separate full models incorporating each of the exam characteristics of interest (W.Blooms and percent CR) and separately ran a model-selection procedure using the MuMIn package (Barton, 2015) to determine the best-fitting model for each exam characteristic. Multilevel models were analyzed in R using the lme4 package (Bates *et al.*, 2014).

Model-Selection Procedure. We identified the fixed-effect variables that best explain student exam scores using a widely accepted multimodel inference approach called Akaike's information criterion (AIC; Akaike, 1973), specifically using AIC corrected for small sample sizes (AICc). AICc values were used to determine which model best fit our data given our sample size. AICc values were also used to rank models, with the lowest AICc values representing the best-fitting models. Using AICc values, we calculated differences in AICc values relative to the best model (Δ_{AICc}) and Akaike weights (ω_i). Large Δ_{AICc} values indicate that models are less likely to explain differences in the response variable, with models that have $\Delta_{AICc} > 10$ considered poor models (Burnham and Anderson, 2004). ω_i are used to compare models, as they are an approximation of the probability that a given model is the best-fitting model given the observed data (Burnham and Anderson, 2004). Thus, larger ω_i are indicative of better-fitting models. AICc analyses were performed in R using the MuMIn package (Barton, 2015).

In addition to providing AICc values, Δ_{AICc}, and ω_i for model selection, the MuMIn package also generates model-averaged regression coefficients from all of the models included in the selection procedure. The calculation of model-averaged regression coefficients takes into account uncertainty associated with determining the best-fitting model (Anderson, 2008; Garamszegi, 2011). We report the model-averaged coefficient outputs provided by the MuMIn package in our tables. Finally, the MuMIn package also calculates the relative importance of each of the fixed-effect variables included in our models using ω_i. The relative variable importance represents the likelihood that a given term is in the best model.

Important Points to Consider. It is important to note that our study has a retrospective design. Given that, we cannot use survey data or student interviews to untangle the influence of differences in ability from the influence of prior experiences and psychological factors on student exam performance. We attempt to statistically control for differences in academic ability between students by using regression models with a control variable for a student's performance in their prior college classes. Including this variable matches students in our different groups of interest (socioeconomic background and gender) by a proxy for demonstrated ability in college-level courses. Thus, any gaps observed in this paper are between students who were theoretically equally competent in their prior college classes, assuming those classes were of equal difficulty. This implies that, by at least one measure, these students have equal academic ability, and if they have differential outcomes on exams, then factors other than ability are likely influencing their performance.

RESULTS

Weighted Bloom's Index

Descriptive Information. The 87 exams collected from 26 different instructors teaching introductory courses taught over a 3-yr period had substantial variation in their median weighted Bloom's level but generally tested more moderate levels of cognitive thinking (e.g., application and analysis questions) than low levels of cognitive thinking (0.53 ± 0.085 [SD]). The sample ranged from exams that tested almost exclusively low-level thinking (0.36) to exams that tested higher-order thinking (0.71).

Model Selection. Using model selection, we found six models with reasonable support ($\Delta_{AICc} < 10$) that explained the impact of median weighted Bloom's level of an exam on the exam performance of male and female students. The top two models had the majority of the support (summed $\omega_i = 0.88$; Table 1). The best model included all of the fixed-effect terms, except course. The second-best model included all of the fixed-effects terms.

The median weighted Bloom's level of an exam had a relative variable importance of 1.0 (Table 2) and was present in all of the six best models, indicating that the weighted Bloom's index of an exam had a consistent and reliable impact on students' exam scores. Both the gender of a student and the interaction term between students' gender and the weighted Bloom's index of an exam were well supported in our models, as they were present in the top model (Table 1) and had very high relative variable importance (Table 2). The same was true for SES status of a student and the interaction term between students' SES status and the median weighted Bloom's level of an exam (Tables 1 and 2).

The cumulative incoming college GPA of students was present in all of the six top models (summed $\omega_i = 1.0$; Table 1) and had a relative variable importance of 1.0 (Table 2), indicating that the cumulative incoming college GPA had a consistent and reliable impact on students' exam scores. The incoming GPA of a student significantly and positively impacted student's exam performance ($\beta = 0.164 \pm 0.00311$ [SE], p value < 0.0001; Table 2).

Model-Averaged Regression Coefficients: Effect of Weighted Bloom's Index on Exam Performance

Weighted Bloom's Index. Because of the strong support for the inclusion of the interaction terms between gender or SES and the weighted Bloom's index, we cannot provide a universal effect for Bloom's index. Instead, we see that groups of students respond in different ways to exams with different Bloom's levels. The main effect of Bloom's level in the model describes the condition for middle/high-SES males after statistically controlling for their performance in their prior college courses, the time in the term they took the exam, the difficulty of the exam they took (median overall difficulty of all exams: 0.63), the instructor they had, and any differences among students not accounted for specifically in the model. For these students, increasing the weighted Bloom's index of an exam is predicted to decrease their exam performance ($\beta = -0.168 \pm 0.0173$ [SE], p value < 0.0001; Table 2). Framing this in terms of impact on a student's exam scores: a middle/high-SES male student with the sample median cumulative GPA of 3.27 will score 11.53% lower on an exam with the weighted Bloom's index of 0.71 (the exam in our sample with the highest Bloom's index) relative to his score on an exam with a weighted Bloom's index of 0.36 (the exam in our sample with the lowest Bloom's index).

Gender × Weighted Bloom's Index. Based on these models, the impact of weighted Bloom's index on performance is predicted to be more extreme for women than for men. The main effect of gender averaged across the model set was not significant ($\beta = -0.00295 \pm 0.0115$ [SE], p value = 0.798; Table 2 and Figure 1a), but the interaction between gender and weighted Bloom's index was significant ($\beta = -0.0418 \pm 0.0205$ [SE], p value = 0.0416; Table 2 and Figure 1a). Taken together, these terms indicate that, on exams with only low-level questions, there is no performance difference between males and females with the same demonstrated prior academic ability. However, as the weighted Bloom's index of exams increases (e.g., moving from predominantly lower Bloom's-level to higher Bloom's-level questions), the performance of female students declines more rapidly than male students of equal ability level, causing a gender-based achievement gap (Figure 1a). Specifically, the model predicts that the difference in

Table 1. Best models include the interaction between student gender identity and the median weighted Bloom's level of an exam and the interaction between SES status and the median weighted Bloom's level of an exam[a]

Rank	Model[b]	AICc	Δ_{AICc}	ω_i
1	Cum.GPA + Time + Gender + SES + W.Diff + W.Blooms + SES*W.Blooms + Gender*W.Blooms	–41429.42	0.00	0.59
2	Cum.GPA + Time + Gender + SES + W.Diff + W.Blooms + SES*W.Blooms + Gender*W.Blooms + Course	–41427.96	1.46	0.29
3	Cum.GPA + Time + Gender + SES + W.Diff + W.Blooms + SES*W.Blooms	–41424.51	4.90	0.05
4	Cum.GPA + Time + Gender + SES + W.Diff + W.Blooms + Gender*W.Blooms	–41423.36	6.06	0.03
5	Cum.GPA + Time + Gender + SES + W.Diff + W.Blooms + SES*W.Blooms + Course	–41423.05	6.36	0.02
6	Cum.GPA + Time + Gender + SES + W.Diff + W.Blooms + Gender*W.Blooms + Course	–41421.90	7.51	0.02
7	Cum.GPA + Time + Gender + SES + W.Diff + W.Blooms	–41417.46	11.96	0.00

[a]Relative ranking (from most support to least) of the six best models for predicting student exam performance using AICc model selection. Models that are informative ($\Delta_{AICc} < 10$) are shown, plus the next best model that had a $\Delta_{AICc} > 10$. The table shows only fixed-effect terms, but all models also include two random-effect terms: student and the instructor whose classes students were enrolled in.

[b]Time = exam number in a course; Cum.GPA = cumulative college GPA at start of introductory biology series; Gender = student's gender identity; SES = students' socioeconomic status; W.Diff = median weighted difficulty of an exam; W.Blooms = the median weighted Bloom's level of an exam; Course = the three courses that are part of the introductory biology sequence.

Table 2. Increasing the median weighted Bloom's level of an exam disproportionately favors male students and middle/high-SES students relative to female or low-SES students, respectively[a]

Parameter	Relative variable importance	Model averaged regression coefficient ± SE	*p* Value[b]
Intercept	NA	0.647 ± 0.0199	**<0.0001**
Cum.GPA	1.00	0.164 ± 0.00311	**<0.0001**
Course (reference level: course 1)			
Course 2	0.33	0.00945 ± 0.0199	0.634
Course 3	0.33	−0.00416 ± 0.0148	0.779
Time (reference level: time 1 (exam 1)			
Time 2 (exam 2)	1.00	0.0131 ± 0.00182	**<0.0001**
Time 3 (exam 3)	1.00	0.0388 ± 0.00255	**<0.0001**
Time 4 (exam 4)	1.00	0.0821 ± 0.00270	**<0.0001**
Student gender (reference level: male)			
Female	1.00	−0.00295 ± 0.0115	0.798
Student SES status (reference level: middle/high-SES)			
Low-SES	1.00	0.00931 ± 0.0149	0.531
Exam characteristics			
W.Diff	1.00	−0.197 ± 0.0113	**<0.0001**
W.Blooms	1.00	−0.168 ± 0.0173	**<0.0001**
Student identity × exam characteristics (reference level: male or middle/high-SES)			
Female × W.Blooms	0.92	−0.0418 ± 0.0205	**0.0416**
Low-SES × W.Blooms	0.96	−0.0628 ± 0.0263	**0.0171**

[a]The outputs were produced via model averaging of all possible models using the MuMIn package in the program R. Although not shown, the models include two random-effects terms: (1|Stu.ID) + (1|Instr).

[b]Bolded *p* values are significant.

performance between a male and female student, both of whom are middle/high-SES students, have the same GPA (3.27, the median incoming GPA for all students in our study), and took the same exam (median difficulty = 0.63), would be 1.73% on the exam with the lowest weighted Bloom's index in our data set. Increasing weighted Bloom's index of an exam to 0.50, the performance gap is predicted to increase to 2.34%, and on an exam with a weighted Bloom's index of 0.70, the gap would be in a 3.22%. Considering that 70% of our exams had a median weighted Bloom's level greater than or equal to 0.50, this results in a gender-based achievement gap across a majority of the exams that were administered to students in our data set (Figure 1a).

SES × Weighted Bloom's Index. The impact of weighted Bloom's index on students from low-SES backgrounds is more extreme than the impact on students from middle/high-SES backgrounds. As with the gender terms, the main effect of SES was not significant (β = 0.00931 ± 0.0149 [SE], *p* value = 0.531; Table 2 and Figure 1b); however, there was a significant interaction between SES status and the median weighted Bloom's level of an exam (β = −0.0628 ± 0.0263 [SE], *p* value = 0.0171; Table 2 and Figure 1b). Taken together, these terms tell us that, on exams with only low-level questions, there is no performance difference between low-SES and middle/high-SES students of equal prior academic ability. However, as the weighted Bloom's index of exams increases (e.g., moving from predominantly lower-order questions to moderate/higher-order questions), the performance of students from low-SES backgrounds declines more rapidly than that of students from middle/high-SES backgrounds, causing a SES-based achievement gap (Figure 1b) despite matched academic ability. Specifically, the difference in performance between a low- and middle/high-SES student, both of whom are male, have the same GPA (3.27) entering the same class, and took the same exam (median difficulty = 0.63), was 1.27% when students took the exam with the lowest weighted Bloom's index in our data set. The model predicts that increasing the weighted Bloom's index of an exam to 0.50 will increase the performance gap to 2.17%, and increasing the weighted Bloom's index to 0.70 would increase the SES-based gap to 3.46%. As 70% of our exams had a weighted Bloom's index greater than or equal to 0.50, our model predicts a significant achievement gap based on SES for the majority of the exams in our data set (Figure 1b).

Constructed-Response versus Restricted-Response Questions

Descriptive Information. The 87 exams collected from 26 different instructors teaching the introductory courses taught over a 3-yr period had substantial variation in their format but generally had more constructed-response questions than restricted-response questions. The average percent CR on the exams in our analysis was 0.66 ± 0.29 (SD). The exams in our sample ranged from 0% (entirely restricted-response questions) to 100% (entirely constructed-response questions).

Model Selection. We found four models with reasonable support ($\Delta_{AICc} < 10$) that explained the impact of percent CR

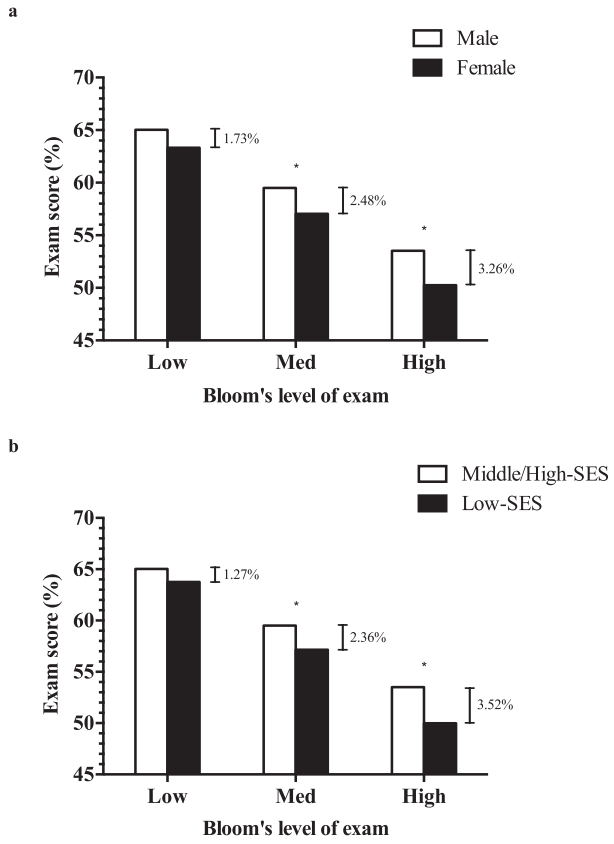

Figure 1. Increasing the median weighted Bloom's level of an exam negatively impacts all students' scores, but it disproportionately favors men more so than women and middle/high-SES students over low-SES students. The figure shows a point estimate for exam performance (percentage score) for (a) male and female students and (b) middle/high-SES and low-SES students based on the model-averaged regression coefficients. The bars are the regression-model predictors of performance for two hypothetical students with an incoming GPA of 3.27 (the median GPA for all students in our data set) who are either (a) middle/high-SES students that identify as male or female or (b) male students who are classified as middle/high-SES or low-SES students, both of whom took a moderately difficult exam with a median difficulty of 0.63 (on a scale of 0.33–1). Thus, these students differ from each other in only two ways: the median weighted Bloom's level of the exam and either (a) their gender (male, unfilled bars; females, filled bars) or (b) their SES status (middle/high-SES, unfilled bars; low-SES, filled bars). The median weighted Bloom's levels, on a scale of 0.33–1, used to calculate the low, medium, and high Bloom's-level exams were 0.36, 0.53, and 0.71, respectively. An asterisk indicates a significant differences between groups of students on a given test. Brackets with percent scores indicate the magnitude of the difference in exam scores for the two students.

on the exam performance of male and female students and students from low- and middle/high-SES backgrounds. The top two models had a majority of the support (summed ω_i = 0.86; Table 3). The best model included all fixed-effects terms, excluding the gender × percent CR interaction. The second-best model included all fixed-effects terms.

Percent CR had a very high relative variable importance (1.0; Table 4) and was present in all of the best models (Table 3), indicating this variable had a reliable effect across all of the models and was very likely to be in the best model. SES and the interaction between SES and percent CR were also very well supported in our models (present in all of the top models and with a relative variable importance of 1.0; Tables 3 and 4). Although gender had strong support (Table 4), the interaction term between gender and percent CR was not well supported (this term was present in only two of the four best models and had a low relative variable importance of 0.29; Table 4). These results suggest that different levels of percent CR on exams differentially impact students from different SES backgrounds but not male or female students.

The cumulative incoming college GPA of students was present in all of the four top models (summed $\omega_i = 1.0$; Table 3) and had a relative variable importance of 1.0 (Table 4), indicating that the cumulative incoming college GPA had a consistent and reliable impact on students' exam scores. The incoming GPA of a student significantly and positively impacted student's exam performance ($\beta = 0.165 \pm 0.00311$ [SE], p value < 0.0001; Table 4).

Model-Averaged Regression Coefficients: Effect of Percent of Constructed-Response Questions on Exam Performance

Percent CR. As with our other analysis, the strong support for the interaction term between SES and percent CR means we cannot report a universal effect of percent CR on student performance. Instead, we see that students from low-SES backgrounds respond differently to percent CR than students from middle/high-SES backgrounds. The main effect of percent CR in the models describes the condition for male students from middle/high-SES backgrounds after statistically controlling for their performance in prior college courses, the time in the term they took the exam, the difficulty of the exam (median difficulty = 0.63), the instructor, and other unmeasured differences between students. For middle/high-SES male students, there is predicted to be a positive effect of increasing the percentage of constructed-response questions on exams ($\beta = 0.0789 \pm 0.00668$ [SE], p value < 0.0001; Table 4). Put in terms of the impact on a student's exam scores, a male student from a middle/high-SES background with a cumulative GPA of 3.27 will score 14.54% lower on an exam that is purely restricted-response questions than he will on an exam that is purely constructed-response questions.

Gender × Percent CR. The significant main effect term of gender indicates that there is an achievement gap between male and female students in this model, with males outperforming females of the same academic ability ($\beta = -0.0252 \pm 0.00341$ [SE], p value < 0.0001; Table 4 and Figure 2a). The lack of a significant interaction term between gender and percent CR ($\beta = -0.000607 \pm 0.00273$ [SE], p value = 0.824; Table 4 and Figure 2a) indicates that the format of the exam questions does not differentially influence the performance of males and females of equal ability on these exams. Thus, as the percentage of constructed-response questions on an exam increases, male and female students with equivalent incoming cumulative GPAs are equally positively impacted.

SES × Percent CR. For students from low-SES backgrounds, the positive effect on performance of more constructed-response questions on an exam is less than it is for students from middle/high-SES backgrounds. The main effect of SES averaged across the model set was not significant ($\beta = -0.00503 \pm 0.00589$ [SE], p value = 0.393; Table 4 and

Table 3. Best model includes the interaction between SES status and the percentage of constructed-response questions on an exam[a]

Rank	Model[b]	AICc	Δ_{AICc}	ω_i
1	Cum.GPA + Time + Gender + SES + W.Diff + Percent CR + SES*Percent CR + Course	–41285.44	0.00	0.61
2	Cum.GPA + Time + Gender + SES + W.Diff + Percent CR + SES*Percent CR + Gender*Percent CR + Course	–41283.64	1.81	0.25
3	Cum.GPA + Time + Gender + SES + W.Diff + Percent CR + SES*Percent CR	–41281.96	3.49	0.11
4	Cum.GPA + Time + Gender + SES + W.Diff + Percent CR + SES*Percent CR + Gender*Percent CR	–41280.15	5.29	0.03
5	Cum.GPA + Time + Gender + SES + W.Diff + Percent CR + Course	–41268.62	16.82	0.00

[a]Relative ranking (from most support to least) of the four best models for predicting student exam performance using AICc model selection. Models that are informative ($\Delta_{AICc} < 10$) are shown, plus the next best model that had a $\Delta_{AICc} > 10$. The table shows only fixed-effect terms, but all models also include two random-effect terms: student and the instructor whose classes' students were enrolled in.

[b]Time = exam number in a course; Cum.GPA = cumulative college GPA at start of introductory biology series; Gender = student's gender identity; SES = students' socioeconomic status; W.Diff = median weighted difficulty of an exam; Percent CR = percentage of constructed-response question on an exam; Course = the three courses that are part of the introductory biology sequence.

Figure 2b), indicating that students of low- and middle/high-SES backgrounds perform equally well on exclusively restricted-response exams. As the percent CR increases on an exam, all students are predicted to benefit, but students from low-SES backgrounds benefit less from this changing format, despite having demonstrated similar academic ability ($\beta = -0.0278 \pm 0.00645$ [SE], p value < 0.0001; Table 4 and Figure 2b). This differential benefit with increasing percent CR on exams produces an achievement gap with students from low-SES backgrounds performing worse relative to their peers from middle/high-SES backgrounds. Specifically, the difference in performance between a low- and middle/high-SES student, both of whom identify as male, with the same GPA (3.27), entering the same class, and taking the same exam (median difficulty = 0.63) was only 0.49% when students took an exam that consisted entirely of restricted-response questions. The performance gap increases to 1.76% when the exam contains an equal mix of restricted-response and constructed-response questions and to 2.86% when the exam contained exclusively constructed-response questions. As 77% of our exams had a percent CR score greater than or equal to 0.50, this results in a consistent SES-based achievement gap due to exam question format across a majority of the exams administered to students in our data set, even though overall all students perform better on constructed-response questions (Figure 2b).

Table 4. Increasing the number of constructed-response questions on an exam disproportionately benefits middle/high-SES students, but not male students, relative to low-SES and female students, respectively[a]

Parameter	Relative variable importance	Model averaged regression coefficient ± SE	p Value[b]
Intercept	NA	0.515 ± 0.0238	**<0.0001**
Cum.GPA	1.00	0.165 ± 0.00311	**<0.0001**
Course (reference level: course 1)			
Course 2	0.85	0.0677 ± 0.0381	0.0759
Course 3	0.85	0.0159 ± 0.0247	0.521
Exam (reference level: time 1 (exam 1)			
Time 2 (exam 2)	1.00	0.0116 ± 0.00183	**<0.0001**
Time 3 (exam 3)	1.00	0.0257 ± 0.00250	**<0.0001**
Time 4 (exam 4)	1.00	0.0745 ± 0.00269	**<0.0001**
Student gender (reference level: male)			
Female	1.00	–0.0252 ± 0.00341	**<0.0001**
Student SES status (reference level: middle/high-SES)			
Low-SES	1.00	–0.00503 ± 0.00589	0.393
Exam characteristics			
W.Diff	1.00	–0.243 ± 0.0114	**<0.0001**
Percent CR	1.00	0.0789 ± 0.00668	**<0.0001**
Student identity × exam characteristics (reference level: male or middle/high-SES)			
Female × percent CR	0.29	–0.000607 ± 0.00273	0.824
Low-SES × percent CR	1.00	–0.0278 ± 0.00645	**<0.0001**

[a]The outputs were produced via model averaging of all possible models using the MuMIn package in the program R. Although not shown, the models include two random-effects terms: (1|Stu.ID) + (1|Instr).

[b]Bolded p values are significant.

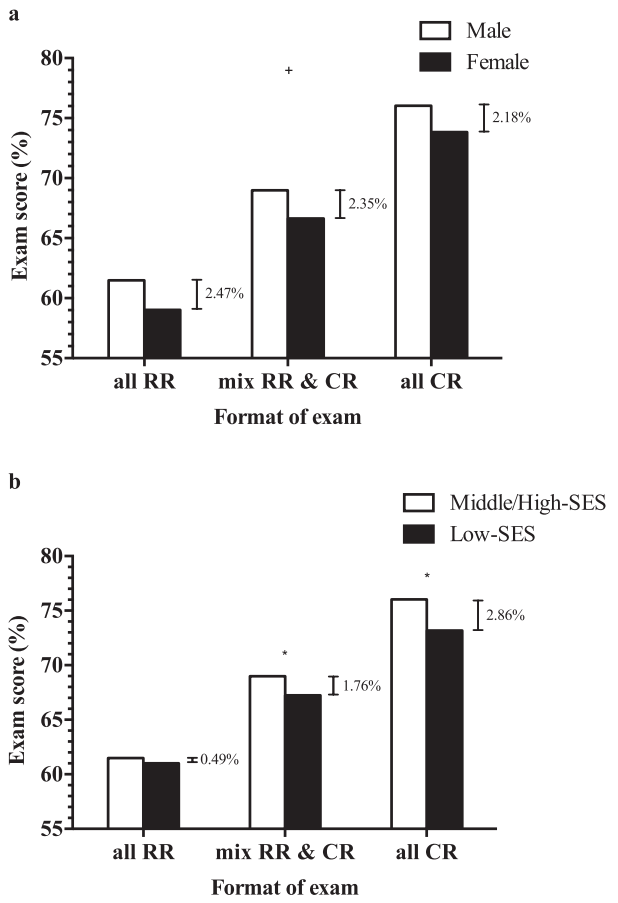

Figure 2. Increasing the number of constructed-response questions on an exam positively impacts all students' exam scores, equally benefiting male and female students yet disproportionately favoring middle/high-SES students over low-SES students. The figure shows a point estimate for exam performance (percentage score) for (a) male and female students and (b) middle/high-SES and low-SES students based on the model-averaged regression coefficients. The bars are the regression-model predictors of performance for two hypothetical students with an incoming GPA of 3.27 (the median GPA for all students in our data set) who are either (a) middle/high-SES students who identify as male or female or (b) male students who are classified as middle- to high-SES or low-SES students, both of whom took a moderately difficult exam with a median difficulty of 0.63 (on a scale of 0.33–1). Thus, these students differ from each other in only two ways: the percentage of constructed-response questions on the exam and either (a) their gender (male, unfilled bars; females, filled bars) or (b) their SES status (middle/high-SES, unfilled bars; low-SES, filled bars). The percentage of constructed-response questions, on a scale of 0–1, used to calculate the all restricted-response (RR), mixture of restricted-response and constructed-response (CR), and all constructed-response exams were 0.00, 0.50, and 1.00, respectively. The + indicates a significant overall difference between two groups of students. An asterisk indicates a significant differences between groups of students on a given test. Brackets with percent scores indicate the magnitude of the difference in exam scores for the two students.

DISCUSSION

As biology instructors continue to develop courses and assessments that promote deeper conceptual understanding in their students, it will be important to understand how the characteristics of exams may impact different populations of students. In this paper, we took the first step of exploring whether and to what extent Bloom's level and question format of instructor-generated assessments differentially impact students' exam performance using a data set that includes 25 classes, 26 instructors, 87 unique exams, and 4810 students across three introductory biology courses for majors at a large research institution. Even after controlling for a measure of student academic ability, we found that the performance of male students was favored over females as exams tested at increasingly higher levels of Bloom's taxonomy but not when there were more constructed-response questions. Additionally, the performance of middle/high-SES students was favored over low-SES students as exams tested at increasingly higher levels of Bloom's taxonomy and when exams contained increasingly more constructed-response questions.

Males Outperform Females at Higher Bloom's Levels

Our results show that males and females perform equally well on exams containing mostly low Bloom's-level questions (a difference of 1.73%) but that males outperform females on exams with increasingly higher levels of Bloom's, even after controlling for prior academic ability (a difference as high as 3.26%; Figure 1a). The differences in performance we observed in our study were fairly small, but even small performance gaps are important. Small performance gaps on any single test may accrue within a single course and across multiple courses, potentially generating GPA gaps between students who are otherwise of equal academic ability.

Previous studies examining gender differences on different Bloom's-level questions in undergraduate biology show mixed results. One study found males outperforming females on higher Bloom's-level questions in biochemistry but found that females outperformed males on knowledge-specific type (low Bloom's-level) questions (Migliaccio and Sheikh, 2009). However, another study in introductory biology classrooms found that male students outperformed female students on both low and high Bloom's-level questions (Stanger-Hall, 2012). These differences between our results and these results may be due to smaller sample sizes for these studies or different classroom and/or institutional contexts.

Additionally, our findings may provide insights into why achievement gaps between male and female students in undergraduate biology classrooms have been observed in some studies (Rauschenberger and Sweeder, 2010; Creech and Sweeder, 2012; Stanger-Hall, 2012; Eddy *et al.*, 2014) but not others (Migliaccio and Sheikh, 2009; Willoughby and Metz, 2009; Creech and Sweeder, 2012; Lauer *et al.*, 2013); perhaps the presence of achievement gaps is dependent on the characteristics of the exams used in the classes. If undergraduate biology assessments are composed mostly of "low-level" questions (Momsen *et al.*, 2010, 2013), then one might not observe a performance gap between students, whereas in exams comprising more higher-order-thinking questions (i.e., our study; Haak *et al.*, 2011; Stanger-Hall, 2012; Eddy *et al.*, 2014), a performance gap may emerge. If instructors vary in the degree to which their exams test high Bloom's levels of thinking, then it is not surprising that there are differences among studies regarding whether a performance gap exists between males and females.

No Gender Differences for Exams with More Constructed-Response Questions

Our finding that increasing the percentage of constructed-response questions on exams has no impact on the exam performance of male and female students (Figure 2a) is in contrast to what has previously been published. Previous studies examining the impact of exam format on students' performance showed that males outperformed females on restricted-response questions in an undergraduate biology classroom (Stanger-Hall, 2012), an introductory atmospheric science class (Weaver and Raptis, 2001), and an oceanic survey course (Weaver and Raptis, 2001). However, only Stanger-Hall (2012) controlled for a measure of prior academic ability. The variation in the findings between these studies and ours may again be attributed to differences in sample size. Another explanation is that perhaps other exam characteristics are correlated with constructed-response formats in some studies but not others. If these are not disaggregated, then this may explain some of the variation in observed patterns.

Middle/High-SES Students Outperform Low-SES Students at Higher Bloom's Levels and with More Constructed-Response Questions

Our results demonstrate that, as the average Bloom's level on an exam increases, the gap between students from middle/high-SES backgrounds and low-SES backgrounds increases from 1.27 to 3.52% (Figure 1b). A similar pattern was observed when the number of constructed response questions increased on exams (Figure 2b). Our study is the first to our knowledge to examine how question format (constructed response vs. restricted response) differentially impacts low- and middle/high-SES students. Given the lack of research done in this area, there is clearly a need to further explore how the performance of low- and middle/high-SES students is mediated by the Bloom's level of questions and question format across a wider array of institutions and student populations. Such studies are necessary before any broad generalizations can be made regarding the performance of low- and middle/high-SES students on biology assessments testing varying degrees of cognitive difficulty.

If Not Differences in Ability, Then What May Contribute to These Gaps?

We controlled for a measurement of prior academic ability; therefore, the differential impact of exam characteristics on students of different genders and SES backgrounds was not due to differences in ability levels. As our study was observational and retrospective, we are unable to explicitly identify the underlying mechanism(s) that may lead to the observed gaps between students. However, there is a large body of literature that suggests that student performance can be impacted by environmental factors. The environment can be broadly defined to include factors such as the social environment of the classroom, the experience of students in previous academic environments, an instructor's classroom practices, and/or how questions on exams are contextualized. In the following subsections, we outline candidate factors that could lead to the performance gaps that we observed in our study.

Stereotype Threat. Stereotype threat is a well-documented psychological phenomenon wherein an individual's concern of conforming to a stereotype about a group he or she is associated with can negatively impact his or her performance on a particular task related to that stereotype (Steele and Aronson, 1995). One of the key findings of stereotype threat is that it is most likely experienced when individuals encounter challenging situations and/or experience high frustration (Steele, 1997). Thus, stereotype threat would be more likely to be triggered on questions requiring more challenging cognitive tasks like transferring conceptual understandings to solve new, application-level or higher cognitive questions (i.e., higher Bloom's-level questions; Spencer *et al.*, 1999; Keller, 2007). Thus, the research on stereotype threat aligns with our findings for gender and SES: women and low-SES students as well as men and middle/high-SES students perform equally on low Bloom's-level exam items (low challenge, low frustration), but stereotype threat could be triggered for women and low-SES students when dealing with more cognitively challenging questions, leading to our observed disproportionate decrease in performance.

Currently, there are insufficient data to determine whether students are under threat in introductory biology classrooms. One study exploring stereotype threat for women in biology, Lauer *et al.* (2013) did not find support for the presence of stereotype threat. However, their method of documenting stereotype threat was to test one intervention to mitigate threat. This test does not rule out the possibility that students are under threat. It could mean that this intervention may not have addressed the right kind of stereotype threat for this population of women (Shapiro *et al.*, 2013). A second study surveyed female students about their experience with stereotype threat and found women in biology experience less stereotype threat than women in physics, but did not explicitly test whether women in biology experience more threat than men, as no men were surveyed in the study (Smith *et al.*, 2015). Only one study has tested whether stereotype threat is present for low-SES students in introductory biology classrooms. Harackiewicz and colleagues (2014) successfully used a values-affirmation intervention to alleviate the impact of stereotype threat on the performance of first-generation students, who are often low-SES students as well. It is possible that female and low-SES students are under threat and that this psychological phenomena could explain our findings, but further work needs to be done to assess this phenomenon in biology.

Implicit Theories of Intelligence. People tend to hold one of two beliefs about intelligence: 1) intelligence is innate and fixed at a certain level or 2) intelligence is effort-based and can grow (Dweck, 1999). The fixed mind-set tends to be more prevalent in high-achieving students and in students who are aware of stereotypes about their group in a particular field (i.e., women in math; Dweck, 2006). Students with fixed mind-sets tend to underperform relative to students with a growth mind-set, especially in the face of challenging tasks (Grant and Dweck, 2003; Blackwell *et al.*, 2007). Furthermore, it has been shown that students with a fixed mind-set are unable to recover from initially poor grades in college science classes, whereas students with growth mind-sets are able to recover from this setback (Grant and Dweck, 2003). Among students who hold a fixed mind-set, there is evidence that

males tend to outperform female students in a college science course, despite controlling for ability level (Grant and Dweck, 2003). These patterns may occur because, when faced with difficult tasks, students with fixed mind-sets tend to withdraw and denigrate their ability, whereas students with growth-based mind-sets embrace the challenge by putting in greater effort and/or trying new strategies.

If a significant number of women and low-SES students in our study population held a fixed view of intelligence, they may have become frustrated and avoided embracing the challenges of answering more cognitively challenging questions, potentially hindering their performance on exams that assess higher levels of Bloom's taxonomy. Given Grant and Dweck's (2003) findings, we would also expect an emerging gender performance gap on assessments that test increasingly higher levels of Bloom's, which is what we observed. Thus, like stereotype threat, implicit theories of intelligence could contribute to an achievement gap between groups of students as the cognitive challenge of an exam increases. This may be an interesting avenue for future research.

Students' Prior Experiences. Inequitable access to resources during students' K–12 experiences may explain our findings. Students from lower-SES backgrounds tend to come from schools with fewer resources (Oakes, 1990). These schools tend to have less experienced and/or qualified instructors (Oakes, 1990; Ingersoll, 1999) and fewer advanced and AP classes (Oakes, 1990; Handwerk *et al.*, 2008), potentially resulting in fewer opportunities for low-SES students as compared with middle/high-SES students to practice answering higher-order questions. This lack of prior practice may explain why students from lower-SES backgrounds may not perform as well on higher Bloom's-level or constructed-response questions. Interestingly, performance on the writing portion of the SAT has also been shown to correlate with students' SES (Mattern *et al.*, 2008), further supporting our assertion.

Question Context. Question context can elicit bias against students, producing achievement gaps. McCullough (2004) found that replacing questions containing stereotypically male-oriented contexts with stereotypically female-oriented contexts reduced the gender gap in performance on a physics concept inventory. Given that higher-order questions require applying conceptual understanding to a novel context, these questions are more likely than lower-order questions to contain novel scenarios. Although biology instructors might not be using sports examples to contextualize their questions to the extent that physics instructors might, it is still possible for biology instructors to construct questions that elicit bias.

In summary, these factors are but a few of the potential mechanisms that could explain why women and low-SES students perform below their male and middle/high-SES peers, even when they have demonstrated equal achievement in prior classes. These mechanisms produce the same patterns observed in our data: no achievement gap when students are exposed to low Bloom's-level questions or tests that contained only restricted-response questions but an appearance of an achievement gap when students are challenged by higher Bloom's-level or constructed-response questions. However, the degree to which these various mechanisms contribute to the gaps observed in this study is unknown, and future research should focus on examining whether and in what ways these various mechanisms may explain the trends observed in this study.

What Are Next Steps?

This study is not suggesting that we as instructors make our exams less cognitively difficult or only restricted-response format to reduce performance gaps on assessments among groups of students. Rather, we suggest instructors modify their instructional practices in ways that help to give students opportunities to practice these types of questions in a low-stakes environment before they are asked to do so on a high-stakes summative exam. For example, this may mean assigning practice exams outside of class or incorporating exam-like questions into lectures as clicker questions, similar to Freeman *et al.* (2011).

These practices may be helpful, because they may reduce the level of frustration and anxiety triggered on high-stakes exams, potentially mitigating factors such as threat and thus improving performance. Furthermore, they may provide students from disadvantaged backgrounds with the practice that may help them catch up with their middle/high-SES colleagues, again reducing performance gaps on assessments. The heavy emphasis on practice in active-learning classrooms may therefore be contributing to the reduction in achievement gaps seen between low- and middle/high-SES students (Haak *et al.*, 2011) as well as between first-generation and continuing-generation students (Eddy and Hogan, 2014) in these classes. Clearly, more work is needed to investigate the underlying reasons why active learning reduces performance gaps between groups of students.

However, these strategies may not be enough, particularly if students are experiencing something like stereotype threat. Instructors may need to use psychological interventions to help ameliorate these phenomena in an effort to promote equity on challenging exams. These interventions can include 1) reframing the assessment to address the "fairness" of the test by stating the test is not biased against a particular group (e.g., Spencer *et al.*, 1999; Good *et al.*, 2008) or that the assessment is meant to document mastery rather than compare individuals (Croizet and Dutrevis, 2004; Smeding *et al.*, 2013); 2) using values affirmation (e.g., Cohen *et al.*, 2006; Miyake *et al.*, 2010; Sherman *et al.*, 2013; Harackiewicz *et al.*, 2014) to mitigate stereotype threat; and 3) emphasizing that intelligence is fluid and malleable, can change over time, and is driven by effort rather than innate ability (e.g., Aronson *et al.*, 2002; Good *et al.*, 2003; Blackwell *et al.*, 2007). Exploring these interventions in the context of undergraduate biology exams are important areas of future research.

Study Limitations

This study was done in a particular context with a specific set of students and may not reflect conditions at other institutions or institution types. Specifically, this study was conducted at a selective R1 institution, and we encourage other instructors to look for the existence of similar patterns in their classrooms at their institutions. Additionally, our observation that increasing the number of constructed-response questions on an exam results in an increase in student's overall performance may have occurred because graders evaluating constructed-response questions may have been more lenient in their allocation of points compared with

restricted-response questions, which are binary in their point allocation.

CONCLUSIONS

Our findings illustrated that, even after controlling for academic ability, males and middle/high-SES students outperform females and low-SES students on assessments testing higher Bloom's levels of thinking and constructed-response questions favor middle/high-SES students. Simply put, these inequities are ones that, no matter how small, should not be present in our classrooms, as inequities in individual classrooms can accrue over time, potentially resulting in two students with truly equal academic abilities having different GPAs upon graduation. It is important to continue to explore the extent to which these achievement gaps exist across multiple types of institutions and identify instructional practices that can close these gaps while maintaining the rigor of our assessments.

ACKNOWLEDGMENTS

We thank Scott Freeman, Alison Crowe, Erin Shortlidge, and the members of the Brownell Lab Biology Education Research Group at Arizona State University for their feedback and comments on earlier versions of the manuscript. We thank Josh Kessack, Jack Cerchiara, Mercedes Converse, and Jennifer Mae-White Day for helping us determine the characteristics of each item; and Ben Wiggens, John Parks, and Chessa Goss for helping us with accessing course exams. Additionally, we thank Michael Angilletta and Stephen Pratt for their insights into the statistical analyses. Support for this study was provided by National Science Foundation (NSF) TUES 1118890 and NSF TUES 1322556. This research was done under approved IRB 38945, University of Washington.

REFERENCES

Akaike H (1973). Information theory as an extension of the maximum likelihood principle. In: Second International Symposium on Information Theory, ed. BN Petrov and F Csaki, Budapest, Hungary: Akademiai Kiado, 267–281.

American Association for the Advancement of Science (2011). Vision and Change in Undergraduate Biology Education: A Call to Action, Washington, DC.

Anderson DR (2008). Model Based Inference in the Life Sciences: A Primer on Evidence, New York: Springer.

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

Aronson J, Fried CB, Good C (2002). Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence. J Exp Soc Psych *38*, 113–125.

Barton K (2015). MuMIn: Multi-model Inference, R package, Version 1.13.5. http://CRAN.R-project.org/package=MuMIn (accessed 21 April 2015).

Bastick T (2002). Gender differences for 6–12th grade students over Bloom's cognitive domain. Paper presented at the Western Psychological Association, WPA Convention, Irvine, CA, April 14 – 17, 2002.

Bates D, Maechler M, Bolker B, Walker S (2014). lme4: Linear Mixed-Effects Models Using "Eigen" and S4, R Package, Version 1.1–7. http://cran.r-project.org/web/packages/lme4/index.html (accessed 21 April 2015).

Beller M, Gafni N (2000). Can item format (multiple choice vs. open-ended) account for gender differences in mathematics achievement? Sex Roles *42*, 1–21.

Black P, Wiliam D (1998). Inside the black box: raising standards through classroom assessment. Phi Delta Kappan *80*, 139–144, 146–148.

Blackwell LS, Trzesniewski KH, Dweck CS (2007). Implicit theories of intelligence predict achievement across an adolescent transition: a longitudinal study and an intervention. Child Dev *78*, 246–263.

Bloom BS, Krathwohl DR, Masia BB (1956). Taxonomy of Educational Objectives: The Classification of Educational Goals, New York: McKay.

Burnham KP, Anderson DR (2004). Multimodel inference: understanding AIC and BIC in model selection. Sociol Method Res *33*, 261–304.

Carlton ST, Harris AM (1992). Characteristics Associated with Differential Item Functioning on the Scholastic Aptitude Test: Gender and Majority/Minority Group Comparisons, Princeton, NJ: Educational Testing Service.

Carrell SE, Page ME, West JW (2010). Sex and science: how professor gender perpetuates the gender gap. Q J Econ *125*, 1101–1144.

Cohen GL, Garcia J, Apfel N, Master A (2006). Reducing the racial achievement gap: a social-psychological intervention. Science *313*, 1307–1310.

Creech LR, Sweeder RD (2012). Analysis of student performance in large-enrollment life science courses. CBE Life Sci Educ *11*, 386–391.

Croizet JC, Dutrevis M (2004). Socioeconomic status and intelligence: why test scores do not equal merit. J Poverty *8*, 91–107.

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, Korn WS (2009). The American College Teacher: National Norms for the 2007–2008 HERI Faculty Survey, Los Angeles: Higher Education Research Institution, UCLA.

de Ayala RJ (2009). The Theory and Practice of Item Response Theory, New York: Guilford.

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

DeMars CE (2000). Test stakes and item format interactions. Appl Meas Educ *13*, 55–77.

Derting TL, Ebert-May D (2010). Learner-centered inquiry in undergraduate biology: positive relationships with long-term student achievement. CBE Life Sci Educ *9*, 462–472.

Dweck CS (1999). Self-Theories: Their Role in Motivation, Personality, and Development, Philadelphia, PA: Psychology Press.

Dweck CS (2006). Mindset: The New Psychology of Success, New York: Ballantine.

Eddy SL, Brownell SE, Wenderoth MP (2014). Gender gaps in achievement and participation in multiple introductory biology classrooms. CBE Life Sci Educ *13*, 478–492.

Eddy SL, Converse M, Wenderoth MP (2015). PORTAAL: a classroom observation tool assessing evidence-based teaching practices for active learning in large science, technology, engineering, and mathematics classes. CBE Life Sci Educ *14*, ar23.

Eddy SL, Hogan KA (2014). Getting under the hood: how and for whom does increasing course structure work? CBE Life Sci Educ *13*, 453–468.

Entwistle NJ, Entwistle A (1991). Contrasting forms of understanding for degree examinations: the student experience and its implications. High Educ *22*, 205–227.

Freeman S, Eddy SL, McDonough M, Smith MK, Okoroafor N, Jordt H, Wenderoth MP (2014). Active learning increases student performance in science, engineering, and mathematics. Proc Natl Acad Sci USA *111*, 8410–8415.

Freeman S, Haak D, Wenderoth MP (2011). Increased course structure improves performance in introductory biology. CBE Life Sci Educ *10*, 175–186.

Freeman S, O'Conner E, Parks JW, Cunningham M, Hurley D, Haak D, Dirks C, Wenderoth MP (2007). Prescribed active learning increases performance in introductory biology. CBE Life Sci Educ *6*, 132–139.

Garamszegi LZ (2011). Information-theoretic approaches to statistical analysis in behavioral ecology: an introduction. Behav Ecol Sociobiol *65*, 1–11.

Good C, Aronson J, Harder JA (2008). Problems in the pipeline: stereotype threat and women's achievement in high-level math courses. J Appl Dev Psychol *29*, 17–28.

Good C, Aronson J, Inzlicht M (2003). Improving adolescents' standardized test performance: an intervention to reduce the effects of stereotype threat. J Appl Dev Psychol *24*, 645–662.

Grant H, Dweck CS (2003). Clarifying achievement goals and their impact. J Pers Soc Psychol *85*, 541–553.

Haak DC, HilleRisLambers J, Pitre E, Freeman S (2011). Increased structure and active learning reduce the achievement gap in introductory biology. Science *332*, 1213–1216.

Handwerk P, Tognatta N, Coley RJ, Gitomer DH (2008). Access to Success: Patterns of Advanced Placement Participation in U.S. High Schools, Princeton, NJ: Educational Testing Service.

Harackiewicz JM, Canning EA, Tibbetts Y, Giffen CJ, Blair SS, Rouse DI, Hyde JS (2014). Closing the social class achievement gap for first-generation students in undergraduate biology. J Educ Psychol *106*, 375–389.

Harris AM, Carlton ST (1993). Patterns of gender differences on mathematics items on the Scholastic Aptitude Test. Appl Meas Educ *6*, 137–151.

Ingersoll RM (1999). The problem of underqualified teachers in American secondary schools. Educ Res *28*, 26–37.

Jensen JL, McDaniel MA, Woodard SM, Kummer TA (2014). Teaching to the test … or testing to teach: exams requiring higher order thinking skills encourage greater conceptual understanding. Educ Psychol Rev *26*, 307–329.

Keller J (2007). Stereotype threat in classroom settings: the interactive effect of domain identification, task difficulty and stereotype threat on female students' math performance. Br J Educ Psychol *77*, 323–338.

Kreft IGG, de Leeuw J (2002). Introducing Multilevel Modeling, Thousand Oaks, CA: Sage.

Lauer S, Momsen J, Offerdahl E, Kryjevskaia M, Christensen W, Montplaisir L (2013). Stereotyped: investigating gender in introductory science courses. CBE Life Sci Educ *12*, 30–38.

Le LT (2009). Investigating gender differential item functioning across countries and test languages for PISA science items. Int J Test *9*, 122–133.

Lindberg SM, Hyde JS, Petersen JL, Linn MC (2010). New trends in gender and mathematics performance: a meta-analysis. Psychol Bull *136*, 1123–1135.

Mattern KD, Shaw EJ, Williams FE (2008). Examining the Relationship between the SAT, High School Measures of Academic Performance, and Socioeconomic Status: Turning Our Attention to the Unit of Analysis (RN-36), New York: College Board. http://research.collegeboard.org/publications/content/2012/05/examining-relationship-between-sat-high-school-measures-academic (accessed 27 April 2015).

Mazzeo J, Schmitt AP, Bleistein CA (1993). Sex-Related Performance Differences on Constructed-Response and Multiple-Choice Sections of Advanced Placement Examinations (CB Rep. No. 92-7; ETS RR No. 93-5), New York: College Entrance Examination Board. http://research.collegeboard.org/publications/content/2012/05/sex-related-performance-differences-constructed-response-and-multiple (accessed 25 April 2015).

McCullough L (2004). Gender, context, and physics assessment. J Int Womens Stud *5*, 20–30.

McDaniel MA, Thomas RC, Agarwal PK, McDermott KB, Roediger HL (2013). Quizzing in middle-school science: successful transfer performance on classroom exams. Appl Cognit Psychol *27*, 360–372.

Migliaccio B, Sheikh O (2009). Gender differences in performance in Principles of Biochemistry based on Bloom's taxonomy of question difficulty and study habits. NCSU Undergrad Res J *5*, 76–83.

Miyake A, Kost-Smith LE, Finkelstein ND, Pollock SJ, Cohen GL, Ito TA (2010). Reducing the gender achievement gap in college science: a classroom study of values affirmation. Science *330*, 1234–1237.

Momsen JL, Long TM, 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.

Momsen J, Offerdahl E, Kryjevskaia M, Montplaisir L, Anderson E, Grosz N (2013). Using assessments to investigate and compare the nature of learning in undergraduate science courses. CBE Life Sci Educ *12*, 239–249.

Neuschmidt O, Barth J, Hastedt D (2008). Trends in gender differences in mathematics and science (TIMSS 1995–2003). Stud Educ Eval *34*, 56–72.

Oakes J (1990). Multiplying Inequalities: The Effects of Race, Social Class, and Tracking on Opportunities to Learn Mathematics and Science, Santa Monica, CA: Rand Corporation.

Paterson L, Goldstein H (1991). New statistical methods for analysing social structures: an introduction to multilevel models. Br Educ Res J *17*, 387–393.

Raudenbush SW, Bryk AS (2002). Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd ed., Thousand Oaks, CA: Sage.

Rauschenberger MM, Sweeder RD (2010). Gender performance differences in biochemistry. Biochem Mol Biol Educ *38*, 380–384.

Rickards JP, Friedman F (1978). The encoding versus the external storage hypothesis in note taking. Contemp Educ Psychol *3*, 136–143.

Ryan KE, Chiu S (2001). An examination of item context effects, DIF, and gender DIF. Appl Meas Educ *14*, 73–90.

Shapiro JR, Williams AM, Hambarchyan M (2013). Are all interventions equal? A multi-threat approach to tailoring stereotype threat interventions. J Pers Soc Psychol *104*, 277–288.

Sherman DK, Hartson KA, Binning KR, Purdie-Vaughns V, Garcia J, Taborsky-Barba S, Tomassetti S, Nussbaum AD, Cohen GL (2013). Deflecting the trajectory and changing the narrative: how self-affirmation affects academic performance and motivation under identity threat. J Pers Soc Psychol *104*, 591–618.

Smeding A, Darnon C, Souchal C, Toczek-Capelle MC, Butera F (2013). Reducing the socio-economic status achievement gap at university by promoting mastery-oriented assessment. PLoS One *8*, e71678.

Smith JL, Brown ER, Thoman DB, Deemer ED (2015). Losing its expected communal value: how stereotype threat undermines women's identity as research scientists. Soc Psychol Educ *18*, 443–466.

Spencer SJ, Steele CM, Quinn DM (1999). Stereotype threat and women's math performance. J Exp Soc Psychol *35*, 4–28.

Stanger-Hall KF (2012). Multiple-choice exams: an obstacle for higher-level thinking in introductory science classes. CBE Life Sci Educ *11*, 294–306.

Steele CM (1997). A threat in the air: how stereotypes shape intellectual identity and performance. Am Psychol *52*, 613–629.

Steele CM, Aronson J (1995). Stereotype threat and the intellectual test performance of African Americans. J Pers Soc Psychol *69*, 797–811.

Taylor CS, Lee Y (2011). Ethnic DIF in reading tests with mixed item formats. Educ Assess *16*, 35–68.

Thomas PR, Bain JD (1984). Contextual dependence of learning approaches: the effects of assessments. Hum Learn *3*, 227–240.

Weaver AJ, Raptis H (2001). Gender differences in introductory atmospheric and oceanic science exams: multiple choice versus constructed response questions. J Sci Educ Technol *10*, 115–126.

Wenglinsky H (2002). How schools matter: the link between teacher classroom practices and student academic performance. Educ Policy Anal Arch *10*, 12–32.

Willoughby SD, Metz A (2009). Exploring gender differences with different gain calculators in astronomy and biology. Am J Phys *77*, 651–657.

Zar JH (2010). Biostatistical Analysis, 5th ed., Upper Saddle River, NJ: Pearson.

Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM (2009). Mixed Effect Models and Extensions in Ecology in R, New York: Springer.