Meta-analysis of Gender Performance Gaps in Undergraduate Natural Science Courses

Sara Odom,1 Halle Boso,1 Scott Bowling,1 Sara Brownell,2 Sehoya Cotner,3 Catherine Creech,4 Abby Grace Drake,5 Sarah Eddy,6 Sheritta Fagbodun,7 Sadie Hebert,7 Avis C. James,4 Jan Just,9 Justin R. St. Juliana,8 Michele Shuster,8 Seth K. Thompson,5 Richard Whittington,7 Bill D. Wills,1 Alan E. Wilson,1 Kelly R. Zamudio,6 Min Zhong,1 and Cissy J. Ballen1*

1Department of Biological Sciences, Auburn University, Auburn, AL 36849; 2School of Life Sciences, Arizona State University, Tempe, AZ 85282; 3Department of Biology Teaching and Learning, University of Minnesota – Twin Cities, Minneapolis, MN 55414; 4Department of Biology, Mt. Hood Community College, Gresham, OR 97030; 5Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, NY 14850; 6Department of Biological Sciences, Florida International University, Miami, FL 33199; 7Department of Biology, Tuskegee University, Tuskegee, AL 36088; 8Department of Biology, New Mexico State University, Las Cruces, NM 88003; 9Department of Biology, Portland Community College, Portland, OR 97217

ABSTRACT
To investigate patterns of gender-based performance gaps, we conducted a meta-analysis of published studies and unpublished data collected across 169 undergraduate biology and chemistry courses. While we did not detect an overall gender gap in performance, heterogeneity analyses suggested further analysis was warranted, so we investigated whether attributes of the learning environment impacted performance disparities on the basis of gender. Several factors moderated performance differences, including class size, assessment type, and pedagogy. Specifically, we found evidence that larger classes, reliance on exams, and undisrupted, traditional lecture were associated with lower grades for women. We discuss our results in the context of natural science courses and conclude by making recommendations for instructional practices and future research to promote gender equity.

INTRODUCTION
Extensive research on the experiences of women in science, technology, engineering, and mathematics (STEM) fields has revealed several common patterns of inequalities that reduce the retention of women in STEM (Eddy and Brownell, 2016). Such systemic challenges include gender stereotypes about STEM careers (DiDonato and Strough, 2013), poor mentorship (Newsome, 2008), unconscious bias against women (Moss-Racusin et al., 2012), and inadequate institutional support to help balance family demands (Goulden et al., 2011). Beyond these systemic challenges, institutional and pedagogical choices can also have negative impacts on metrics of performance for women. Examples include large class sizes (Ballen et al., 2018), biased in-class participation (Aguillon et al., 2020; Bailey et al., 2020), and reliance on multiple-choice exams (Stanger-Hall 2012). Due in part to these challenges, women are less likely than men to complete science-related college majors and join the STEM workforce (Chen, 2013).
Binary gender$^1$ performance gaps$^2$ in science are well documented in a variety of STEM courses (Brooks and Mercincavage, 1991; Grandy, 1994; Tai and Sadler, 2001; Rauschenberger and Sweeder, 2010; Creech and Sweeder, 2012; Sonnert and Fox, 2012; Lauer$^3$ et al., 2013; McCullough, 2013; Peters, 2013; Hansen and Birol, 2014; Matz et al., 2017), including studies that control for measures of incoming student ability (Eddy et al., 2014; Eddy and Brownell, 2016; Wright et al., 2016; Salehi et al., 2019b). In some higher education STEM studies that do not control for prior ability, there are cases in which there is no performance gap or one that favors women (Eddy and Brownell, 2016). Controlling for incoming student ability or preparation can account for differences in student performance that arise from factors correlated with demographic characteristics (e.g., gender, race/ethnicity, first-generation status; Salehi et al., 2020), so that one can compare students with similar student ability or preparation. These controlled differences are interesting to researchers, as they can point to classroom issues that create observed underperformance, defined as “not performing to ability” (Salehi et al., 2019a). The “raw” performance outcomes in STEM course work (not controlling for incoming preparation) can have lasting repercussions on future STEM careers, and this is the analytical approach we used in the current study. For example, Wang et al. (2015) found that 12th-grade math scores—on which girls underperformed relative to boys—mediated students’ selection of STEM occupations in their early to mid-30s. In other cases, the impact is immediate. Many undergraduate students start out in introductory STEM courses that serve as required prerequisites for continuing in their majors. If women receive low grades in these introductory STEM courses, then they are less likely than men with similar grades and academic preparation to retake the course, more likely to drop out, and less likely to advance (Rask and Tiefenthaler, 2008; Seymour and Hunter, 2019; Harris et al., 2020). Thus, research that addresses factors that drive observed performance gaps to minimize these inequities has the potential to enhance the persistence of women in STEM.

Understanding factors that lead to inequities requires first investigating ways that instructional practices affect student performance. Previous research has investigated a large number of non–mutually exclusive course elements hypothesized to impact gender performance gaps. For example, many introductory courses are taught in large classrooms (Matz et al., 2017), despite evidence that large courses may negatively affect women’s performance (Ho and Kelman, 2014; Ballen et al., 2018) and participation (Ballen et al., 2019; Bailey et al., 2020). Assessment strategies have also been proposed to have an impact on binary gender gaps. Especially in large introductory courses, student performance is often assessed primarily through the use of timed, multiple-choice exams (Matz et al., 2017), despite research that shows this approach is not a meaningful measure of critical thinking or learning (Martinez, 1999; Dufresne et al., 2002; Simkin and Kuechler, 2005) and may specifically disadvantage women (Ballen et al., 2017a). Performance gaps have been shown to be higher on high-stakes exams than they are on other proxies for performance, such as overall grade point average (GPA), or lower-stakes exams (Stanger-Hall, 2012; Kling et al., 2013). Finally, the instructor’s pedagogical approach in the classroom might impact performance. Substantial evidence now confirms that active learning improves student outcomes in STEM courses (Freeman et al., 2014), and it may offer disproportional benefits for other groups often underrepresented in STEM, such as underrepresented minority students (Eddy and Hogan, 2014; Ballen et al., 2017b; Casper et al., 2019; Theobald et al., 2020) and first-generation students (Eddy and Hogan, 2014). However, when it comes to gender gaps, the effectiveness of active learning has been mixed. While some studies claim reduced gender gaps in active-learning courses (Lorenzo et al., 2006), other studies have been unable to reproduce the same effect (Pollock et al., 2007; Madsen et al., 2013; Ballen et al., 2017b). Yet other studies have noted the potential for active learning to exacerbate inequities that could influence performance for students with anxiety (England et al., 2017; Cohen et al., 2019), which would disproportionately affect women (Cooper et al., 2018a; Downing et al., 2020) and could lead to gender gaps.

To test the hypothesis that gender impacts performance in natural science courses and to test the impact of moderators on relative performance outcomes, we conducted a meta-analysis by analyzing data from a wide selection of published and unpublished data (Glass, 1976). Focusing on undergraduate-level biology and chemistry (e.g., general biology, cell biology, biochemistry, general chemistry; see Supplementary Material for more information), we analyzed student scores from a large number of courses and institutions to identify factors that impact gender equity. Specifically, we address the following questions:

1. Is there a performance gap between men and women in undergraduate biology and chemistry courses?
2. What classroom factors (e.g., class size, assessment type, pedagogy) narrow historic gender gaps by promoting women’s performance?

**METHODS**

**Study Identification**

We identified studies following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) protocol (Moher et al., 2009; Supplemental Figure S1). On February 27, 2019, we performed a database search of three online education research–affiliated databases: ERIC, Education Research Complete, and PsychINFO, with search results limited to journal articles, theses, and dissertations. We used the following search terms, limited to subject descriptors: (biology OR STEM OR science OR medical OR chemistry) AND (education OR achievement OR test OR performance OR outcomes OR examinations OR medical OR chemistry)
OR student) AND (university OR college OR higher education OR adulthood) AND (sex OR gender OR female OR gap) NOT foreign countries NOT admission NOT readiness NOT high school NOT career.

We used the following inclusion criteria to determine whether the studies identified by this search would be included in the final data set:

1. Data were collected in undergraduate-level courses at colleges and universities in the United States.
2. Data came from a course within the biological and chemical sciences. Data could be aggregated across multiple sections of the same course but could not be combined across different courses.
3. Data included exam scores (average score on one or more exams), course grades (the final grade that students received in a course), or science concept inventory (CI) scores disaggregated by gender.

Data from published studies were screened and coded by authors S.O. and H.B. We screened the studies first by reading the abstracts. To increase the number and scope of classroom scores in our analysis, we carried over into full-text screenings both studies that focused on student academic performance and studies that focused on other classroom elements, in case student scores were provided as context for those studies. Studies that were not disqualified based on the abstract were downloaded and the full text screened. Studies were included in our final data set only if we could ensure that all study criteria were met. When studies suggested that data that met our criteria were collected but not included in the publication, we emailed the study’s author(s) to request additional data. Our original search identified 2822 studies. Abstract screening and exclusion of duplicate studies removed 2689 studies, leaving 133 studies for full-text evaluation. Of these, 25 studies could not be accessed, 39 were not conducted in the appropriate setting, and 51 lacked the appropriate data needed for this study (study did not provide grades, scores were not disaggregated by gender, etc.; Supplemental Figure S1). For 22 studies, the text suggested that the authors collected data that fit our criteria, but the data were not included in the published paper. In these cases, we requested data directly from the authors. We were unable to get in contact with the authors of 10 studies. For nine studies, we were able to contact authors, but they were unable to provide us with data because of privacy concerns or because they could no longer access it. Authors of three studies shared data, which we included in the final data set. In total, 18 published studies met all of the required criteria for inclusion (see Supplementary Material for full list of published studies included in the analysis). These studies included 89 different courses. Of these courses, 35 included aggregate data for multiple sections. We note that class size was not calculated based on sample size in these cases; for aggregated data, class size was either missing, or we used average class size. Additionally, we collected course grades and descriptions of 80 individual courses from institutions across the United States in conjunction with the Equity and Diversity in Undergraduate STEM Research Coordination Network (i.e., unpublished data; Thompson et al., 2020). These data were collected during the course of normal academic classes, with the intention of using them in education research studies that focus on different aspects of equity. Because the instructors who collected these data were involved in this study, we were able to directly follow up on any questions about these data and how they fit this study’s criteria. Data were provided in the form of raw grades, which authors S.O. and C.J.B. used to calculate mean scores and SD for men and women students. We had multiple comparisons from a subset of the n = 169 courses (e.g., both exam score and course grade), and so our data set included n = 246 comparisons and more than 28,000 students.

Data Collection
The research reported was determined to be exempt from Auburn University’s Institutional Review Board (protocol 19-355 EX 1908). From each course, we collected sample size, mean scores, and SD for men and women for whichever of the three specified assessment types (exam scores, final grades, or science CI) were available. If SD and other measures of variance were not included (9.76% of studies), we imputed them based on the average SD of the other scores in each assessment category (Purukawa et al., 2006). To account for the possibility that these studies had larger SDs than the average, we also ran a sensitivity analysis by using a larger SD (75th percentile). Because this did not change any of the outcomes (see Supplemental Table S4), we present only the results calculated using average SD. For three studies, gender differences in scores were only available in the form of z-scores. Additionally, we collected the following information as it was available (Supplemental Table S1): institution name and/or type (Supplemental Table S2), course title (e.g., Introduction to Biology), broader topic (biology or chemistry), intended student audience (natural science major or non-major), number of sections (one or multiple sections), class size, instructor(s) gender, pedagogy (lecture-based or active learning), assessment type (exam score, final course grade, or CI), and course level (introductory/lower division or upper division; Figure 1). We categorized introductory courses as those with a course title that included the terms “introductory” or “principles” or when the description of the course included this information. We categorized upper-level courses as those that had prerequisites or when the study specified that upper-level students typically took them. To include pedagogy in a quantitative model, we categorized descriptions provided by instructors or the literature into either lecture based or active learning (Supplemental Table S3). In “lecture” courses, the majority of course time was dedicated to instruction by the teacher, with few if any alternative activities occurring during a normal class period. “Active learning,” a broad category describing approaches designed to increase student engagement (Freeman et al., 2014; Driessen et al., 2020), included courses that incorporated interactive and student-focused activities into the course structure. Using the descriptions provided in each study, two authors (S.O. and C.J.B.) individually categorized the pedagogy of each course, with initial interrater agreement of 83.3%. The primary source of disagreement was in cases in which a course incorporated activities as part of a required laboratory component. We decided to focus only on the lecture component of the course, and following this discussion, we achieved 100% agreement.

Statistical Analyses
We ran all statistical analyses using R v. 3.6.2 (R Core Team, 2019) within R studio v. 1.2.5033 (R Studio Team, 2019). We used the metafor package (Viechtbauer, 2010) for effect size
calculations, models, and checking for publication bias; the MuMIn package (Barton, 2020) for model selection by Akaike information criterion (AIC); the multcomp package (Hothorn et al., 2008) for pairwise comparisons; and the tidyverse package (Wickham et al., 2019) to streamline coding and create some of the graphs.

To account for differences in grade distributions across different courses, we quantified gender gaps by calculating a standardized mean difference for each course in the form of Hedges’s $g$ (Hedges, 1981):

$$g = \frac{\text{Mean men’s score} - \text{Mean women’s score}}{\text{Pooled weighted SD}} \times \frac{N - 3}{N - 2.25} \times \sqrt{\frac{N - 2}{N}}$$

For Hedges’s $g$ calculated from $z$-scores rather than means, we used the following formula:

$$g = \frac{\text{Mean women’s } z\text{-score} - \text{Mean men’s } z\text{-score}}{\text{Pooled weighted SD}} \times \frac{N - 3}{N - 2.25} \times \sqrt{\frac{N - 2}{N}}$$

We set up these calculations so that a positive Hedges’s $g$ indicates that women scored higher than men, while a negative Hedges’s $g$ indicates that men scored higher than women. The degree of difference is based on the absolute value of the effect size. While interpretations of effect size impact vary depending on the context of comparisons, within education, Hedges and Hedberg (2007) suggest that Hedges’s $g$ values of 0.2 and greater can indicate differences that should be of interest to policymakers.

We used a random effects model, with university and subject as nested random effects (Konstantopoulos, 2011), to calculate the overall effect size based on the Hedges’s $g$ estimates and sampling variances of all of the grade comparisons, using the Hedges’s estimator to account for heterogeneity. Some studies provided both course grade and average exam score. In these cases, we used course grade in calculating the overall effect size (we obtained the same results when exam grades were prioritized; see Supplemental Table S4).

We checked for publication bias by generating a funnel plot, running a trim-and-fill analysis, and calculating a fail-safe $n$ using the Rosenberg method (Rosenberg, 2005).

Based on initial results, we used a mixed effects model to measure the impact of course factors on gender gaps. We selected models based on AIC (Arnold, 2010; Theobald, 2018), considering the following as potential fixed effects: class size, assessment type (science CIs, exam scores, and course grade), pedagogy category (active or traditional lecture), course level (introductory or upper level), and broad topic (biology or chemistry). University and subject were included as nested random effects (Konstantopoulos, 2011). Because we took this approach, we advise readers to interpret our moderators in an “all-else-equal” context, with the “all” consisting of our other variables. Because assessment type contained three factors, we performed post hoc pairwise comparisons on assessment type using Tukey and Holm adjustments to compare each of the factors against each other.

RESULTS

We did not identify a significant gender gap in performance across all published studies and unpublished data (Hedges’s $g = -0.2268$, $p$ value = 0.4119; Supplemental Table S4 and Supplemental Figure S2). This model had a high degree of heterogeneity ($I^2 = 97.00\%$), suggesting other factors may play a role in explaining variation in the data, which we describe in detail below.

We found a negligible impact of publication bias in this data set. While some points fell outside the expected distribution cone in the funnel plot, the distribution of data was relatively symmetrical (Figure 2). Furthermore, a trim-and-fill analysis

![FIGURE 1. Descriptive summary of classes in meta-analysis. (A) Histogram of class sizes; (B) number of comparisons for each assessment type: science CIs, course grade, exam grade; (C) classes by broad subject (biology or chemistry) and level (intro or non-intro); and (D) percentage of pedagogy categories.](image)
Across all classes, we did not detect a statistically significant gender gap within biology and chemistry courses. Due to the high degree of heterogeneity within the data, we explored a number of factors that might be associated with our outcomes. We identified three course elements that predicted gender performance differences—class size, assessment type, and pedagogy. We explored how these factors might impact the historic underrepresentation of women in STEM. Specifically, larger courses and high-stakes exams were associated with underperformance of women relative to men in natural science courses. We also found that, relative to traditional lecture, the incorporation of active-learning strategies was associated with higher performance outcomes among women. Surprisingly, we did not observe differences in gender gaps based on whether the classes in question were biology or chemistry, despite the disciplines’ differences in coverage and culture. We discuss the implications of each impactful factor in the following sections.

Class Size
Our results add to a chorus of studies calling for a decrease in class size to promote student learning and performance. Based on our model, an increase in class size from 50 to 250 students increases gender gaps by ∼0.4 SDs. Prior studies note the association of smaller courses with increased student performance (Achilles, 2012; Ballen et al., 2018), satisfaction with course experience (Cuseo, 2007), and equitable participation (Ballen et al., 2019). However, large courses remain common in undergraduate studies, especially for introductory-level courses (Matz et al., 2017). While institutional demands limit the availability of small classrooms (Saiz, 2014), instructors should be aware of this effect and implement strategies to counter some of the depersonalized, didactic, threat-promoting aspects of the large-lecture environment, such as using group work (Springer et al., 1999; Chaplin, 2009), learning assistants (Knight et al., 2015), names (Cooper et al., 2017), humor (Cooper et al. 2018b), in-class formative assessment techniques (Lowry et al., 2006; Knight et al., 2013), and strategic use of role models (Schinske et al., 2016; Yonas et al., 2020). Because it is unlikely that small classes will become smaller any time soon, future research would profit from an explicit focus on the elements of large courses (other than literal class size) that contribute to gaps in performance. Two examples include research that compares the effectiveness of active-learning strategies between small and large classrooms or tests the impact of two different assessment strategies within large classes. Additionally, descriptive work that isolates certain practices unique to and frequently used in large classes, but not in smaller classes, would build a foundational understanding of factors that may hinder or promote subsets of students.

Assessment Type
We found exams contributed to gender gaps favoring men in introductory science. Based on our model, focusing on course grades, rather than only exam scores, results in a decrease in gender gaps by ∼0.14 SDs. This supports previous research showing that, while exam scores disadvantage women, other assessments in students’ final course grades contribute to more equitable outcomes (Salehi et al., 2019b). While it is common for courses—especially large, introductory courses—to rely heavily on exams to assess students (Koester et al., 2016), this approach may not always provide an accurate reflection of students’ knowledge or critical-thinking skills (Martinez, 1999;...
It is also unlikely that a student's exam score is a reflection of that student's ability to conduct tasks proficiently as a disciplinary scientist. Furthermore, previous research in undergraduate science classrooms shows women are disproportionately affected by test anxiety, leading to lower exam scores (Ballen et al., 2017a; Salehi et al., 2019b). Instructors can promote equity by clearly outlining learning objectives and aligning exam and homework questions (Feldman, 2018) and by integrating affirmation exercises before exams (Miyake et al., 2010; Harris et al., 2019). Instructors can lower the sense of risk in exams by allowing students to retake exams (Nijenkamp et al., 2016; Sullivan, 2017), lowering the stakes of exams (Cotner and Ballen, 2017), or avoiding multiple-choice exams altogether (Stanger-Hall, 2012). The majority of course grades in our sample included exam scores in their calculation; however, course grades also

**FIGURE 3.** Predicted gender gaps across different class sizes and combinations of pedagogies (active, lecture) and assessment types (course grade, exam scores, and CIs) in units of Hedges’s $g$. Assuming grades are assigned on a bell curve, the difference of one Hedges’s $g$ is approximately the difference of one letter grade (though interpretations will vary based on the class grade distribution).
typically incorporate other types of assessments, such as participation, homework, quizzes, or in-class assignments. While some of the assessments included in this study may have incorporated one or more of the recommendations listed, their effects are outside of the scope of this analysis. Future research with an explicit focus on the impact of lowering the stakes of exams will clarify effective methods.

We found the association of CIs on gender gaps did not differ significantly from other assessment types. CIs are unique, because they probe student understanding of fundamental concepts using systematic classroom assessment techniques (Smith and Tanner, 2010). Because of sample size limitations, we caution readers as they interpret our results, and encourage future work to address the impacts of CIs on performance gaps in more depth.

**Pedagogy**

Active learning is increasingly implemented in undergraduate classrooms, and for good reason: plenty of research has demonstrated its advantages in regard to improving student grades (Smith et al., 2009; Freeman et al., 2014). We show that active-learning practices, as opposed to traditional lecture, increased women’s performance in natural science courses, with our model predicting a decrease in gender gaps by $\sim 0.26$ SDs in active-learning classes compared with traditional lecture. This relationship may hinge on one of the following factors associated with active learning: the development of self-efficacy through scaffolded interactions and consistent, low-stakes assessment (Ballen et al., 2017b); increased sense of belonging through the development of in-group relationships (Eddy et al. 2015; Eddy and Hogan, 2014); and the use of metacognition to normalize student perceptions of challenges in the course curriculum (Tanner, 2012). However, we encourage readers to interpret these results with caution due to varied implementation of active-learning practices across our categories (see **Limitations** section).

**Limitations**

One factor this analysis did not control for was incoming preparation. Due to the format and availability of the data included in the analyses, we focused on raw outcomes, without accounting for any initial differences in performance between men and women when they entered the courses. This is a limitation, because previous work identifies incoming preparation (often in the form of ACT/Scholastic Aptitude Test scores or high school GPA) as a key predictor of a student’s outcome in a course (Lopez et al., 2014; Rodriguez et al., 2018). Thus, it is difficult to address the extent of inequality in the classroom without controlling for these differences.

We identified published studies primarily through our database search of ERIC, Education Research Complete, and PsychINFO. We chose to focus on these three databases in order to identify a broad range of education papers without pulling a high volume of duplicate studies. We should acknowledge that there are other education databases and search engines that we did not explore that may have yielded additional studies. Furthermore, we did not hand search any journals or “snowball” additional papers from studies. However, we believe that our

| Model (random effects = university/subject) | AIC | $\Delta i$ | $w_i$ |
|-------------------------------------------|-----|-----------|------|
| Assessment+pedagogy+class.size            | 531.1 | 0.00  | 0.413 |
| Assessment+pedagogy+intro.or.upper+class.size | 532.3 | 1.18  | 0.228 |
| Assessment+pedagogy+biol.or.chem+class.size | 532.7 | 1.61  | 0.184 |

| Regression coefficient | Estimate ± SE | p value |
|------------------------|---------------|---------|
| Intercept              | 0.273 ± 0.416 | 0.512   |
| Class size             | −0.002 ± 0.000 | <0.001 |

| Assessment type (reference level: exams) | Estimate ± SE | p value |
|-----------------------------------------|---------------|---------|
| Course grade                            | 0.142 ± 0.040 | <0.001 |
| CI                                      | −0.661 ± 1.631 | 0.685   |

| Pedagogy (reference level: lecture) | Estimate ± SE | p value |
|------------------------------------|---------------|---------|
| Active                             | 0.262 ± 0.089 | 0.003   |

| Comparison                          | Estimate | SE  | z value | $\text{Pr}(>|z|)$ |
|-------------------------------------|----------|-----|---------|-----------------|
| Assessment type                     |          |     |         |                 |
| CI–exams                            | −0.661   | 1.631 | −0.405 | 1.000           |
| Course–exams                        | 0.142    | 0.040 | 3.546  | 0.001           |
| Course–CI                           | 0.803    | 1.631 | 0.492  | 1.000           |
Our investigations were limited due to the fundamental nature of meta-analytic methods, which are based entirely on published or previously collected data. The factors we chose to investigate were chosen based on the general availability of adequate descriptions in the educational research studies we included. Often, descriptions of certain course elements were limited to studies specifically investigating that effect, and some factors that we originally wished to investigate had to be abandoned due to limited data. For example, we were interested in how institution type may play a role, but we did not possess comprehensive data across all institution types, such as small liberal arts colleges. Additionally, high-stakes exams may be more common in larger courses, so course size may not be the problem, but rather the reliance on high-stakes exams to assess students. Unfortunately, we did not have access in the study sample to examples of large courses that used other types of assessments.

Another limitation was our broad categorization of active classrooms versus traditional lecture classrooms. Active learning is broadly defined in the literature: Freeman et al. (2014) solicited responses from 338 biology seminar academic members and defined active learning as that which “engages students in the process of learning through activities and/or discussion in class, as opposed to passively listening to an expert. It emphasizes higher-order thinking and often involves group work” (Freeman et al., 2014, pp. 8413–8414). Based on biology education literature (n = 148 articles) and feedback from biology instructors (n = 105 individuals), Driessen et al. (2020) defined active learning as “an interactive and engaging process for students that may be implemented through the employment of strategies that involve metacognition, discussion, group work, formative assessment, practicing core competencies, live-action visuals, conceptual course design, worksheets, and/or games, p. 6.” These definitions make clear that what is encompassed under the term “active learning” is extensive. It is used to describe a wide variety of different instructional practices that are infrequently detailed in scholarly publications (Driessen et al., 2020). Although some studies have assessed the effect of specific strategies, such as audience response questions (Caldwell, 2007; Smith et al., 2009; Knight et al., 2013), group discussions (Miller and Tanner, 2015), case studies (Allen and Tanner, 2005; Miller and Tanner, 2015), and flipped classrooms (Tucker, 2012; van Vliet et al., 2015; Rahman and Lewis, 2020), among others, our results add urgency to the need to move beyond coarse categorizations of active learning to more fine-grained work, as it clearly matters to marginalized groups (Thompson et al., 2020). Our research was limited by the publication or instructor descriptions of each course. When descriptions were available, they ranged from highly specific descriptions of the class period to simple designations (i.e., “this was an active-learning course” or “traditional lecture course”). We acknowledge that the categories are not precise and do not fully reflect the range and nuance of what occurs inside each classroom. And while approximately 60% of the courses we included in our analysis were considered active-learning courses, we recognize that, nationally, far fewer classrooms include active learning (Stains et al., 2018), and it is likely that published studies on active learning may bias toward instructors who are more proficient at active learning. An instructor’s experience with and understanding of how to implement active learning likely impacts its effectiveness (Andrews et al., 2011), meaning that a strategy that works in some classrooms might not always show the same effects in other classrooms.

Finally, we recognize that binary gender is noninclusive language. However, the gender binary has been heavily relied upon in prior studies, and as such, this analysis follows the model laid out in the studies we included, meaning that at this time we cannot address how gender identity outside the gender binary affects student performance in different settings. We also recognize that gender is not the only identity-related factor that affects student performance. Many other elements of identity, such as race/ethnicity (Beichner et al., 2007; Ballen et al., 2017b), socioeconomic status (Haak et al., 2011), and LGBTQ+ status (Cooper and Brownell, 2016; Henning et al., 2019) can effect a student’s experiences in a course, and it is likely that these factors could interact with gender expectations in ways that lead to patterns within certain subgroups that differ from our reports.

Final Remarks
Our results point to multiple ways that instructors and administrators can work to promote equitable outcomes in undergraduate classrooms. Particularly in introductory gateway courses, where students appraise their fit in a field based on performance outcomes relative to their peers, reducing class sizes when possible, decreasing reliance on high-stakes exams, and incorporating active-learning strategies into every lecture are possible avenues to promote equity. By using informed, data-driven solutions, instructors and institutions can create more inclusive classrooms.

ACKNOWLEDGMENTS
We are grateful to the biology education research group at Auburn University for help with data collection and valuable feedback on versions of the article: Emily Driessen, Todd Lamb, Egypt Pettway, Sara Wood, Shaday Ewell, Chloé Josefson, Abby Beatty, Tashitso Anamza, and Ash Zemenick. This work is supported by NSF DBI-1919462 awarded to S.C., S.E., and C.J.B.; DUE-2011995 awarded to C.J.B.; the Alabama Agricultural Experiment Station; the Hatch Program of the National Institute of Food and Agriculture; and the U.S. Department of Agriculture.

REFERENCES

Achilles, C. M. (2012). Class-size policy: The Star experiment and related class-size studies [NCPEA policy brief, Vol. 1, No. 2]. NCPEA Publications. Retrieved June 5, 2020, from https://eric.ed.gov/?id=ED540485

Aguillon, S. M., Siegmund, G.-F., Petipas, R. H., Drake, A. G., Cotner, S., & Ballen, C. J. (2020). Gender differences in student participation in an active-learning classroom. CBE—Life Sciences Education, 19(2), ar12. https://doi.org/10.1187/cbe.19-03-0048

Allen, D., & Tanner, K. (2005). Infusing active learning into the large-enrollment biology class: Seven strategies, from the simple to complex. Cell Biology Education, 4(4), 262–268. https://doi.org/10.1187/cbe.05-08-0113

Andrews, T. M., Leonard, M. J., Colgrove, C. A., & Kalinowski, S. T. (2011). Active learning not associated with student learning in a random sample of college biology courses. CBE—Life Sciences Education, 10(4), 394–405. https://doi.org/10.1187/cbe.11-07-0061
Meta-analysis of Gender Performance Gaps

Seymour, E., & Hunter, A. B. (2019). Talking about leaving revisited. New York: Springer.

Simkin, M. G., & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: What is the connection? Decision Sciences Journal of Innovative Education, 3(1), 73–98.

Smith, J. I., & Tanner, K. (2010). The problem of revealing how students think: Concept inventories and beyond. CBE—Life Sciences Education, 9(1), 1–5.

Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C., Knight, J. K., Guild, N., & Su, T. T. (2009). Why peer discussion improves student performance on in-class concept questions. Science, 323(5910), 122–124. https://doi.org/10.1126/science.1165919

Sonnert, G., & Fox, M. F. (2012). Women, men, and academic performance in science and engineering: The gender difference in undergraduate grade point averages. Journal of Higher Education, 83(1), 73–101. https://doi.org/10.1353/jhe.2012.0004

Springer, L., Stanne, M. E., & Donovan, S. S. (1999). Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis. Review of Educational Research, 69(1), 21–51.

Stains, M., Harshman, J., Barker, M. K., Chasteen, S. V., Cole, R., DeChenne-Peters, S. E., ... & Young, A. M. (2018). Anatomy of STEM teaching in North American universities. Science, 359(6383), 1468–1470. https://doi.org/10.1126/science.aap8892

Stanger-Hall, K. F. (2012). Multiple-choice exams: An obstacle for higher-level thinking in introductory science classes. CBE—Life Sciences Education, 11(3), 294–306. https://doi.org/10.1187/cbe.11-11-0100

Sullivan, D. (2017). Mediating test anxiety through the testing effect in asynchronous, objective, online assessments at the university level. Journal of Education and Training, 4, 107. https://doi.org/10.5296/jet.v4i2.10777

Tai, R. H., & Sadler, P. M. (2001). Gender differences in introductory undergraduate physics performance: University physics versus college physics in the USA. International Journal of Science Education, 23(10), 1017–1037. https://doi.org/10.1080/0950069010025967

Theobald, E. (2018). Students are rarely independent: When, why, and how to use random effects in discipline-based education research. CBE—Life Sciences Education, 17(3), rm2. https://doi.org/10.1187/cbe.17-12-0280

Theobald, E. J., Hill, M. J., Tran, E., Agrawal, S., Arroyo, E. N., Behling, S., ... & Freeman, S. (2020). Active learning narrows achievement gaps for under-represented students in undergraduate science, technology, engineering, and math. Proceedings of the National Academy of Sciences USA, 117(12), 6476–6483. https://doi.org/10.1073/pnas.1916903117

Thompson, S., Hebert, S., Berk, S., Brunelli, R., Creech, C., Drake, A. G., ... & Ballen, C. J. (2020). A call for data-driven networks to address equity in the context of undergraduate biology. CBE—Life Sciences Education, 19(4), rm2

Tucker, B. (2012). Online instruction at home frees class time for learning. Education Next, 12, 2

van Vliet, E. A., Winnips, J. C., & Brouwer, N. (2015). Flipped-class pedagogy enhances student metacognition and collaborative-learning strategies in higher education but effect does not persist. CBE—Life Sciences Education, 14(3). https://doi.org/10.1187/cbe.14-09-0141

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, 36(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Wang, M.-T., Degol, J., & Ye, F. (2015). Math achievement is important, but task values are critical, too: Examining the intellectual and motivational factors leading to gender disparities in STEM careers. Frontiers in Psychology, 6. https://doi.org/10.3389/fpsyg.2015.00036

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... & Yutani, H. (2019). Welcome to the Tidyverse. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686

Wright, C. D., Eddy, S. L., Wenderoth, M. P., Abshire, E., Blankenbiller, M., & Brownell, S. E. (2016). Cognitive difficulty and format of exams predicts gender and socioeconomic gaps in exam performance of students in introductory biology courses. CBE—Life Sciences Education, 15(2), ar23.

Yonas, A., Sleeth, M., & Cotner, S. (2020). In a “Scientist Spotlight” intervention, diverse student identities matter. Journal of Microbiology & Biology Education, 21(1).