Course Elements that Impact Performance and Behavior in Undergraduate Science Classes
by

Sara Elizabeth Odom

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
August 8, 2020

Keywords: active learning, engagement, galvanic skin response, gender equity, student performance

Approved by

Cissy Ballen, Chair, Assistant Professor of Biological Sciences
Karen McNeal, Associate Professor of Geosciences
Mary Mendonca, Alumni Professor of Biological Sciences

Abstract

Using evidence-based approaches to teaching can help instructors and institutions improve student experiences in the classroom. By understanding the ways that different classroom environments and activities impact students, instructors can design curricula that maximize positive student outcomes and provide equitable educational experiences. While a variety of teaching strategies and classroom changes have been proposed to improve student engagement and performance, there remains the question of how effective these changes are, and who benefits from them.

In chapter one, I investigated gender disparities in undergraduate life science classes. Using a meta-analytic approach, I analyzed performance differences between men and women's exam scores, course grades, and concept inventories in a variety of different classroom settings. I found a statistically significant gender gap with a small effect size, with men overperforming relative to women within life science classes. However, the degree of difference was affected by class size, pedagogy, assessment type, and pedagogy, with the most dramatic gender gaps found in exam scores.

In chapter two, I investigated student engagement within lecture-based introductory science classes. I used galvanic skin response readings as a biometric measure of student engagement and compared responses to various class activities. I found that students were most engaged during clicker questions.

Together, this research identifies effective instructional approaches, including a reduced reliance on high-stakes exams, and the integration of active learning elements, such as clicker questions. Research into the forces that drive gender disparities in science courses will inform evidence-based teaching practices that promote equitable student outcomes.

Acknowledgments

First, I would like to thank Dr. Cissy Ballen for being such an amazing mentor. I don't believe I would have finished this degree working with anyone but you. You were always there to offer support, encouragement, and help whenever I needed it (and I needed it a lot). Thank you for agreeing to bring me on even in the middle of all the chaos of getting your lab set up here at Auburn. I am honored to have been your first graduate student.

Additionally, I would like to thank the rest of the Ballen lab for their support and feedback. A special thanks goes out to the undergraduate assistants who helped me process all of my data. I never would have finished sorting through all of those meta-analysis papers without your help. I would also like to recognize the collaborators who helped collect data that I used in my research, particularly Zoe Koth and the members of the Equity and Diversity in Undergraduate STEM (EDU-STEM) research coordination network.

Finally, I would like to thank my family and friends. Thank you to my parents for teaching me the importance of a good education and supporting me in all my educational endeavors. And thank you especially to my best friend in the whole world, Renée, for being here with me for the entire journey. There's no one else I would have rather traversed the ups and downs of these last six years with, and wherever life takes us next, I hope we are never too far apart in distance or in spirit.

Research was funded through an NSF-DBI grant awarded to C. J. Ballen and financial support from Dr. Sehoya Cotner. Equipment was provided by Dr. Carl Weiman.

I'd like to dedicate this thesis to my beloved Cassiopeia. Grad school was hard, but it was worth it to finally bring home my grad school cat.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AIC	Akaike information criterion
COPUS	Classroom Observation Protocol for Undergraduate STEM
GSR	Galvanic Skin Response
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
STEM	Science, Technology, Engineering, and Math

CHAPTER ONE

Gender performance gaps in undergraduate life science classrooms

Abstract

Because of the important role that undergraduate grades play in the decision of whether or not to remain in a discipline, understanding performance disparities on the basis of gender may help us to address the attrition of women in higher levels of science education. To investigate broad patterns of gender-based performance gaps, I conducted a meta-analysis reviewing data from 21 published and unpublished studies collected in a total of 169 undergraduate life science classes. I investigated whether attributes of the learning environment contribute to performance disparities. My analysis found a small but significant gender gap in life science classes, with women underperforming relative to men. Several factors that moderated performance differences, such as class size, pedagogy, and assessment type, which was the most influential factor. Specifically, I show that high stakes exams had the largest negative impact on performance differences. My results challenge the frequent use of high-stakes multiple choice exams as the primary assessment method within life science classes and concludes with recommendations for instructional practices that promote women's performance in life sciences.

.

Introduction

Extensive research on the experiences of women in science, technology, engineering, and mathematics (STEM) fields have revealed several common patterns of inequalities that influence the reduced retention of women in STEM (Eddy & Brownell, 2016). Such challenges include but are not limited to gender stereotypes about STEM careers (DiDonato & 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, 2009). While these are patterns that can and should be addressed from a variety of angles, one factor of particular interest to educators is gender differences in student performance within educational settings.

Gender performance gaps in science are well documented in a variety of STEM classes (Brooks & Mercincavage, 1991; Creech & Sweeder, 2012; Grandy, 1994; Hansen & Birol, 2014; Lauer et al., 2013; Matz et al., 2017; McCullough, 2013; Peters, 2013; Rauschenberger & Sweeder, 2010; Sonnert & Fox, 2012; Tai & Sadler, 2001), including studies that control for measures of incoming student ability (Eddy & Brownell, 2016). To understand the underlying sources of performance gaps between students, several models have been proposed. One explanation, the *student deficit model*, proposes that differences in performance stem solely from the students themselves. It assumes an unbiased learning environment, and thus, if students underperform, it is the result of deficits in their own abilities, incoming preparations, or motivations (Solorzano, 1992; Valencia, 2012). However, a second explanation, the *course deficit model*, proposes that class environment also plays an important role in student performance. Rather than assume that classrooms provide a fair learning environment for all students, the course deficit model proposes that some classroom factors may create a bias against certain groups of students, thus leading to performance disparities in underrepresented groups of students (Cotner & Ballen, 2017). This course deficit model is the theoretical framework that I follow in this study.

Academic performance in STEM coursework has lasting repercussions on future STEM careers. For example, Wang et al. (2015) found that 12th grade math scores—on which girls underperformed relative to men—mediated a student's selection of STEM occupations in their

early- to mid-30s. In other cases, the impact is more immediate. For example, many undergraduate students start out in introductory courses that serve as prerequisites required to continue in that major. If women receive lower grades in these initial 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 (Harris et al., 2020; Rask & Tiefenthaler, 2008). Thus, research that addresses underlying mechanisms that drive observed gaps in performance has the potential to enhance the persistence of women in STEM.

To understand these underlying mechanisms requires first investigating ways that institutional or instructional practices impact student performance. Previous research has investigated a number of non-mutually exclusive course elements hypothesized to impact gender performance gaps. One such element in class size, as many introductory classes are taught in large classrooms (Matz et al., 2017), despite evidence that suggests increasing class size may negatively affect women's performance (Ballen et al., 2018; Ho & Kelman, 2014) and participation (Ballen et al., 2019). Another element is pedagogical approach. Significant evidence shows that active learning improves student outcomes in STEM classes (Freeman et al., 2014), and it may offer disproportional benefits for other groups often underrepresented in STEM, such as racial minorities (Ballen, Wieman, et al., 2017; Casper et al., 2019; Eddy & Hogan, 2014) and first generation students (Eddy & Hogan, 2014). However, when it comes to gender gaps, the effectiveness of active learning at reducing that gap has been mixed. While some studies claim reduced gender gaps in active learning classes (Lorenzo et al., 2006), other studies have been unable to reproduce the same effect (Ballen, Wieman, et al., 2017; Madsen et al., 2013; Pollock et al., 2007). Finally, assessment strategy has been proposed to have an impact on gender gaps. Especially in large introductory classes, student performance is often assessed

primarily through the use of timed, often multiple choice exams (Matz et al., 2017), but research has shown that this is often not a meaningful measure of learning (Crowther et al., 2020) and may specifically disadvantage women (Ballen, Salehi, et al., 2017), leading to performance gaps that are greater on high stakes exams than they are on other assessment types, such as overall grade point average, or their performance on lower stakes exams (Cotner & Ballen, 2017; Kling et al., 2013; Stanger-Hall, 2012).

I used a meta-analysis approach to observe overarching trends by analyzing data from a wide selection of published and unpublished data (Glass, 1976). Focusing on life science undergraduate classes, I analyzed student scores from a large number of diverse classes and institutions, to identify factors that impact gender equity. I address the following questions:

1. Is there a performance gap between men and women in undergraduate life science classes?
2. What classroom factors predict potential gender performance gaps?

Methods

Study identification

I identified studies following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) protocol (Moher et al., 2009). I performed a database search on February 27, 2019, of three online education research affiliated databases: ERIC, Education Research Complete, and PsychINFO, with search results limited to journal articles, theses, and dissertations. I used the following search terms, limited to SU descriptors: (biology OR STEM OR science OR medical OR chemistry) AND (education OR achievement OR test OR performance OR outcomes OR examinations 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.

I used the following criteria to select data for inclusion:

(a.) The data was collected in undergraduate level courses at colleges and universities in the United States.

(b.) The data came from a course within the field of life science. Data could be aggregated across multiple sections of the same course but could not be combined across multiple different subjects.

(c.) The data included exam scores (average score on one or more exams), course grades (the final grade that students received in a course), or concept inventory (CI) scores disaggregated by gender.

I screened studies first by reading the abstract, and if a study could not be disqualified based on this description, screened the full study text to ensure it met the study criteria. When a study published in the last 10 years met the first two criteria but not the 3rd, and the text suggested that the appropriate data was collected but not included in the publication, I emailed authors to request additional data. The 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 I needed for this study. In total, 18 published studies met all of the required criteria for inclusion. Additionally, I also included data from three unpublished sources provided by collaborators. In total, these 21 studies included data from 169 classes and over 43,000 students. Because some studies provided data from multiple types of assessment, my final data set included a total of 246 pairs of scores.

Data collection

From each class, I collected sample size, mean scores, and standard deviation for men and women for whichever of the three specified assessment types (exam scores, final grades, or concept inventory) that were available. If standard deviation was not included (9.76% of studies), I imputed them based on the average standard deviation of the other scores in each assessment category (Furukawa et al., 2006). To account for the possibility that these studies had larger standard deviations than the average, I ran a sensitivity analysis by using a larger standard deviation (75th percentile). Because this did not change any of the outcomes, I present only the results calculated using average standard deviation. For three studies, gender differences in scores were only available in the form of z-scores. Additionally, I collected the following information as it was available: institution name and/or type (minority-serving, community college, or research intensive), course title (ex. "Introduction to Biology"), broader topic (biology or chemistry), intended student audience (life science majors or non-majors), number of sections (one or multiple sections), class size, instructor(s) gender, pedagogy (qualitative description given by the author, see Appendix 1), and course level (introductory/lower division or upper division) (Figure 1). Courses were described as introductory courses when they were described as an introductory level course, or if the class title included the terms "introductory" or "principles". Courses were described as upper level if they required prerequisites, or if the study specified that they were typically taken by upper level students. Because pedagogy was originally a qualitative description, in order to analyze it within the model, I categorized the descriptions given into one of three categories: lecture-based, interactive, and active learning, based on criteria outlined in Table 1.

Because of a lack of data in some of these fixed effects, I could not consider all effects in the mixed effects analysis. Certain analyses only included classes for which I had complete information about the predictive factors under consideration.

Statistical analyses

I ran all statistical analyses using R version 3.6.2 (R Core Team, 2019) within R studio version 1.2.5033 (RStudio Team, 2019). I 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 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 classes, I quantified gender gaps by calculating a standardized mean difference for each class in the form of Hedges' *g* (Hedges, 1981):

$$g = \frac{(Mean\ men's\ score) - (Mean\ women's\ score)}{Pooled\ weighted\ Standard\ Deviation} \times \frac{N-3}{N-2.25} \times \sqrt{\frac{N-2}{N}}$$

For Hedges' *g* calculated from z-scores rather than means, I used the following formula:

$$g = (Mean\ women's\ z\ score) - (Mean\ men's\ z\ score) \times \frac{N-3}{N-2.25} \times \sqrt{\frac{N-2}{N}}$$

I set up these calculations so that a positive Hedges' g indicates that women scored higher than men, while a negative Hedges' g indications that men scored higher than women. The degree of difference is based on the absolute value of the effect size, with 0.2 typically

considered the threshold for a small difference, 0.5 the threshold of a medium difference, and 0.8 the threshold for a large difference between groups (Cohen, 1977).

I used a random effects model to calculate the overall effect size based on the Hedges' *g* estimates and sampling variances of all of the grade comparisons, using Hedges estimator to account for heterogeneity. I used this model to check 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).

I used a mixed effects model to measure the impact of class factors on gender gap. I selected models based on Akaike information criterion (AIC) (Arnold, 2010; Theobald, 2018), considering the following as potential fixed effects: class size, assessment type, pedagogy category, course level, and broad topic. University and subject were included as nested random effects (Konstantopoulos, 2011). Because two of the fixed effects contained three factors each, I performed post hoc pairwise comparisons on assessment type and pedagogy within the model, using Tukey and Holm adjustments, to compare each of the factors against each other.

Finally, I compared exam scores and course grades within each pedagogy by creating a dummy variable that combined assessment type and pedagogy, refit the model using this variable in place of assessment type and pedagogy (other fixed and random effects included as before), and removed the intercept to allow all levels of the factor to be included in the summary. Then, I ran a post hoc pairwise comparison using Holm adjustments.

I also plotted the overall Hedges' *g* of each of these groups, calculated using random effects models of each subgroup, against each other to visually show comparisons.

Results

I found that across all studies (n=246), men overperformed relative to women in life science classes (*p*-value=0.0251); however, the difference was relatively small (Hedges' g= -0.1300). This model had a high degree of heterogeneity (I^2=97.00%). I found a negligible impact of publication bias in this data set. While some points fell outside of the expected distribution cone in the funnel plot (Figure 2), the distribution of data was relatively symmetrical. Furthermore, a trim and fill analysis did not add any additional points, meaning that there were not any identified gaps in the data distribution. My fail-safe *n* calculation predicted that 7768 "missed" studies would need to exist to invalidate the study's conclusions. Based on these results, I proceeded with the remaining analyses without any publication bias correction.

Mixed model selection identified several models within ΔAIC <2 (Table 2). I used the model with the lowest AIC value for the remaining comparisons. This model included assessment type, pedagogy, class level, and class size as fixed effects, and university and subject as random effects (Table 3). Class size significantly affected gender gaps (*p*-value<0.0001), with women's relative performance dropping as class size increased. I also found significant impacts of assessment type and pedagogy, but because these factors had more than two levels, I investigated their differences using pairwise comparisons, discussed below. Class level (introductory versus upper level classes) was present in the selected model but had a nonsignificant effect on the gender gap in life science classes (*p*-value=0.1029).

Pairwise comparisons showed the following effects of each factor:

Assessment type: I examined three different assessment types: concept inventories, exam scores, and course grade. Exam scores showed the greatest gender gaps (Figure 3.A). Gender gaps observed in course grades were significantly smaller than gaps seen in exam scores,

with the standard deviation between women and men increasing by 0.31630 when considering exam scores instead of course grades (Table 4). Concept inventory scores were not significantly different from either exam scores or course grades.

Pedagogy: I examined three categories of pedagogy: lecture-based, interactive, and active learning. Active learning classes had the greatest gender gaps (Figure 3.B). The standard deviations of gender difference in active learning classes were greater by 0.33117 compared to lecture-based classes, and 0.33116 compared to interactive classes (Table 4). Scores in lecture-based and interactive courses were not significantly different from each other (*p*-value=0.729191).

Assessment type within pedagogy: When broken down by both assessment type and pedagogy, there is a significant difference between exam scores and assessment type within each pedagogy (Table 5). Exam scores within lecture-based and active learning classes both show significant gender gaps, with women underperforming, that are removed when observing course grades (Figure 4).

Discussion

Overall, I found a minor, but significant gender gap favoring men within life sciences classes. However, because of the high degree of heterogeneity within the data, it makes more sense to consider gender gaps in the context of different class factors. I identified three course elements—class size, assessment type, and pedagogy—that can either exacerbate or reduce gender gaps. Below, I discuss the implications of each.

Class size

My results add to the chorus of studies calling for a decrease in class size to promote student learning and performance. Based on this model, an increase in class size from 50 to 250 students increases gender gaps by 0.5 standard deviations, the equivalent of approximately half a letter grade based on a normally distributed grading scheme. Prior studies note the association of smaller classes with increased student performance (Achilles, 2012; Ballen et al., 2018), satisfaction with class experience (Cuseo, 2007), and equity (Baker et al., 2016; Ballen et al., 2019). However, large classes remain common in undergraduate institutes, especially in introductory level courses (Matz et al., 2017). While institutional demands don't always make it feasible for small class sizes (Saiz, 2014), instructors should at least be aware of this effect, and consider the integration of class activities that can be implemented within large classrooms (Eichler & Peeples, 2016; Kim et al., 2015; Lowry et al., 2006).

Assessment type

I found that the greatest driver of gender gaps in life sciences was assessment type. Due to the low number of classes in this study, it is difficult to make inferences about the specific effect of concept inventories, but there was a distinct difference seen between course grades and exam scores. Women underperformed on exam scores, but not in overall course grades. While it is common for courses—especially large, introductory courses—to rely heavily on timed, multiple choice exams to assess students (Koester et al., 2016), this approach may not always provide an accurate reflection of student's knowledge (Crowther et al., 2020). High-stakes exams are limited in their ability to assess critical thinking skills (Martinez, 1999), and there is doubt about the ability for multiple choice questions to accurately assess student knowledge and

understanding (Dufresne et al., 2002). Furthermore, women are disproportionately affected by test anxiety, leading to lower exam scores (Ballen, Salehi, et al., 2017; Salehi et al., 2019). However, there is evidence that reducing the weight of exams on final scores can reduce performance gaps on the exams themselves (Cotner & Ballen, 2017). Additionally, instructors can promote equity by clearly outlining learning objectives and aligning exam questions and homework questions (Feldman, 2018), and by integrating affirmation exercises before exams (Harris et al., 2019; Miyake et al., 2010). Instructors can lower the sense of risk in exams by allowing students to re-take exams (Nijenkamp et al., 2016; Sullivan, 2017), or avoiding multiple-choice exams altogether (Stanger-Hall, 2012). By utilizing field tested solutions, instructors can adjust their assessments to better assess student ability and make their classes more equitable.

Pedagogy

Active learning is rapidly gaining popularity in undergraduate classrooms, and for good reason: plenty of research shows its advantages in regard to improving student grades (Freeman et al., 2014; Smith et al., 2009). However, out of the three pedagogy categories defined in this study, active learning showed the greatest gender gaps. There are a few considerations that may help us understand this continued disparity

Selective benefits of active learning: Underrepresented groups are not homogenous. While previous research has shown that active learning can reduce gaps for some minority and underrepresented groups (Theobald et al., 2020), not all groups see the same benefit (Eddy & Hogan, 2014; Schwartz et al., 2016). Active learning often involves students working in groups, many times with strangers, a situation that some students may be uncomfortable with (Cohen et

al., 2019), and a student's identity can shape how they feel about certain activities (Eddy & Hogan, 2014), how comfortable they are speaking up and sharing their opinions in discussions (Henning et al., 2019), and how likely they are to participate in activities (Aguillon et al., 2020). Active learning emphasizes the importance of student engagement; if a student is not engaged, they will not reap the benefits, therefore it is important to make sure the active learning environment promotes inclusion so that all students are able to fully participate.

Variety in definition and application: Freeman et al. (2014) used the responses of 338 biology seminar audience members to define active learning as such: "Active learning 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, pgs. 8413-8414). While this definition broadly describes the philosophy behind active learning, in practice, "active learning" is used to describe a wide variety of different institutional practices. Although some studies have assessed the effect of specific strategies, such as audience response questions (Caldwell, 2007; Knight et al., 2013; Smith et al., 2009), group discussions (Miller & Tanner, 2015), case studies (Allen & Tanner, 2005; Miller & Tanner, 2015), and flipped classrooms (Tucker, 2012; van Vliet et al., 2015), among others, future work will benefit from a broader understanding of what works, and for whom. Furthermore, I selected pedagogy categories in a way that considered the intensity of active learning implementation, but I was limited by the language used to describe the classes I included. When descriptions were available, they ranged from highly specific descriptions of the class period, to simple designations (i.e. "this was an active learning class" or "traditional lecture class"), and it is important to recognize that the categories I selected are not clear cut and do not fully reflect the range an nuance of what occurs inside each classroom. Finally, an instructor's

experience with and understanding of how to implement active learning also affects the effectiveness of strategies used (Andrews et al., 2011), meaning that a strategy that works in some classrooms might not always show the same effects in other classrooms.

Underlying influence of assessment strategy: While active learning classes as a whole in this study showed a significant gap, when I broke that down by assessment type, it revealed that the gender gap only occurred in exam scores. Thus, the most likely explanation for continued gender gaps in active learning classes is the continued use of high-stakes assessments. Despite an integration of other activities into their course work, when it comes to student assessment, many active learning classes still rely heavily on exam grades with tests that do not vary much from those seen in lecture-based courses (Crowther et al., 2020), which as noted above, may disadvantage women. Thus, it is possible that potential benefits arising from new pedagogy approaches are being obscured by outdated assessment strategies.

Considering the potential impact of these three considerations, the takeaway from this assessment should not be that active learning is ineffective, but rather, that active learning is not a perfect fix-all, and these complications need to be considered when assessing active learning. As such, instructors wishing to implement active learning should consider not only what activities to integrate, but also address how to best integrate them and how these activities could integrate into student assessment.

Limitations

One factor this analysis did not control for was incoming preparation. Due to the format and availability of the data I used, I focused on raw outcomes, without accounting for any initial differences in performance between men and women when they entered the classes.

Consistently, incoming preparation (often in the form of ACT/SAT scores or high school grade point average) is the primary predictor of a student's outcome in a class (Lopez et al., 2014; Rodriguez et al., 2018). Thus, systemic differences in incoming preparation make it difficult to address the extend of inequality in the classroom.

Because this study was a meta-analysis, built off of previously collected data, my investigations are limited by the data provided. The factors I chose to investigate were chosen based on both previous hypotheses of their effects, and on the general availability of adequate descriptions in the educational research studies I included. Often, descriptions of certain elements were limited to studies specifically investigating that effect, and some factors that I originally wished to investigate had to be abandoned due to limited data. Likely, there are other factors that merit additional investigation. Because of the myriad ways that a classroom environment can affect student outcomes, it is important for education studies to provide a thorough description of the context the data was collected in.

Finally, I should recognize that as our understanding of gender and gender expression changes (Ainsworth, 2015; Rasmussen, 2009), the traditional gender binary utilized by this analysis is not fully representative of every student's experience. However, it is a binary that has been heavily relied upon in prior studies, and as such, this analysis follows the model laid out in the studies I included, meaning that I can't really address the details of how gender identity outside of the traditional gender binary affects students. Furthermore, I should recognize that gender is not the only identity-related factor that affects student performance. Many other elements of identity, such as race (Ballen, Wieman, et al., 2017; Beichner et al., 2007), socioeconomic class (Haak et al., 2011), sexuality (Henning et al., 2019), and personal beliefs (Henning et al., 2019) can effect a student's experiences in a class, and it is incredibly likely that

these factors could interact with gender expectations in ways that lead to pattern within certain subgroups that differ from what is seen overall. While my analysis lacks the data to investigate this, it merits future study, and beyond that, illuminates the importance of creating an inclusive classroom environment.

Final remarks

Unfortunately, gender inequality still exists in undergraduate life science classrooms. However, there are ways that instructors can mitigate this effect. This analysis indicated multiple ways that instructors and administers can work to reduce gender gaps in undergraduate life science classes, specifically, by reducing class sizes when possibly, decrease reliance on high-stakes exams, and critically assess the benefits of active learning approaches. Future work should explore a more nuanced assessment of the ways different elements of active learning minimize or exacerbate gender gaps in order to create equitable active learning approaches. By utilizing informed, field tested solutions, instructors and administrations can create classrooms that are inclusive and equitable.

A. Lecture-based ("Lecture")	
Description	**Examples**
Instructors spent the majority of class time lecturing. Additional activities, such as asking or responding to questions, could exist, but were largely spontaneous and took up only a small portion of class time. Typically, these were self-described as "lecture" courses, with little to no mention of other class activities.	Rauschenberger and Sweeder, 2010: "The instructional model for these courses consists of a single large lecture section …does not contain many of the previously mentioned curricular changes so can provide a good reference for ''traditional'' instruction"
B. Lecture plus interactive elements ("Interactive")	
Description	**Examples**
Interactive elements, such as clickers, think/pair/share, and other "active learning"-like activities were incorporated into the class structure, but the majority of class time was still spent in lecture-based instruction.	Collaborator Survey: "70 min lec and lab twice a week. Traditional text-book (not OER). Lectures were mostly traditional with active learning incorporated."
C. Active Learning	
Description	**Examples**
Over 50% of class time was spent using peer led, project based, flipped classrooms, or groupwork based activities. May be self-described as "active learning", so long the full description given did not indicate it would better fit under one of the previous categories.	Gross, et al., 2013: "This reduced in-class time was supplemented with prerecorded "lectures" available to the students at least a week before class, which increased the online component in the flipped course compared with the standard course. … a substantial difference between course formats was the increased use of active learning in the flipped classroom. This took the form of peer–peer think–pair–share activities, clicker responses, and example problems for students to work in the once-weekly 75-min sections. In the twice-weekly 50-min sessions, team-based learning (Michaelsen et al., 2004) was used. In this format, teams of five to eight students remained allied throughout the semester. In-class activities included difficult example problems attacked by teams, individual and team readiness assessments on new material, and student explanations of problem solutions on projected whiteboards."

Table 1. Pedagogy categorizations and illustrative examples. Because different instructors can employ a wide variety of teaching strategies into their curriculum, when pedagogy was described, I categorized it into 3 categories, based roughly on the three classifications characterized in the latent profile analysis conducted by Stains et al. (2018). For a full breakdown of class descriptions within each category, see Appendix 1.

Model (random effects= university/subject)	AIC	ΔAIC	LogLik	weight
Assessment+pedagogy+intro.or.upper+class.size	**508.5**	**0.00**	**-244.641**	**0.427**
Assessment+pedagogy+Biol.or.Chem+intro.or.upper+class size	509.4	0.93	-243.964	0.269
Assessment+pedagogy+class.size	510.2	1.74	-246.635	0.179

Table 2. Model selection by AIC values. The model selected for remaining analyses is bolded.

Regression Coefficient	**Estimate ± SE**	***p* value**
Intercept	0.3051 ± 0.4091	0.4558
Class size	-0.0025 ± 0.0005	**<0.0001**
Assessment type (reference level: exams)		
Course grade	0.1399 ±0.0402	**0.0005**
Concept Inventory	-0.5697 ± 1.5657	0.7160
Pedagogy (reference level: active learning)		
Interactive	0.8709 ±0.1579	**<0.0001**
Lecture	0.1928 ± 0.1344	0.1516
Class level (reference level: introductory)		
Upper level	-0.2780 ± 0.1705	0.1029

Table 3. Model estimates. Factors with significant slopes are bolded.

Comparison	Estimate	Std. Error	z-value	Pr(>|z|)
Assessment type				
CI-Exams	0.16844	0.21326	0.790	0.859
Course-Exams	0.31630	0.03307	9.562	**< 2e-16**
Course-CI	0.14785	0.21537	0.687	0.859
Pedagogy				
Interactive-Active	0.33116	0.08392	3.946	**0.000159**
Lecture-Active	0.36117	0.04149	8.704	**< 2e-16**
Lecture-Interactive	0.03001	0.08668	0.346	0.729191

Table 4. Pairwise comparison between multileveled fixed effects. Pairs with a significant difference are bolded.

Comparison	Estimate	Std. Error	z-value	Pr(>|z|)
Exam scores - Course grades				
Lecture-based	0.16238	0.04760	3.411	**0.00129**
Interactive	-1.23550	0.19861	-6.221	**1.48e-09**
Active	0.22908	0.07506	3.052	**0.00227**

Table 5. Pairwise comparisons of assessment type within each pedagogy. Groups with significant differences between exam scores and course grades are bolded.

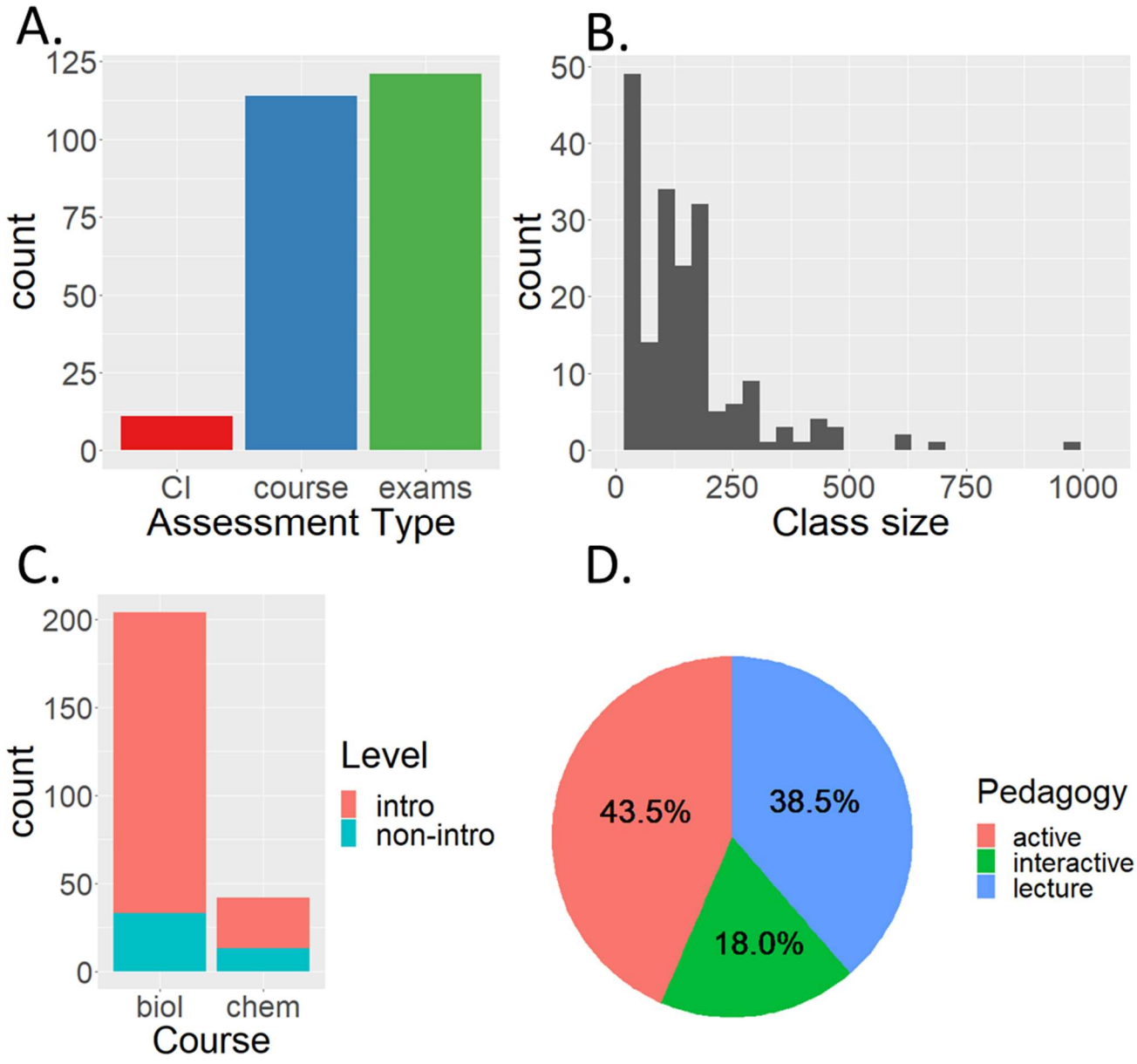

Figure 1. Descriptions of classes used in meta-analysis. (A) Number of comparisons for each assessment type. (B) Histogram of class sizes. (C) Number of classes by broad subject and level. (D) Percentage of pedagogy categories.

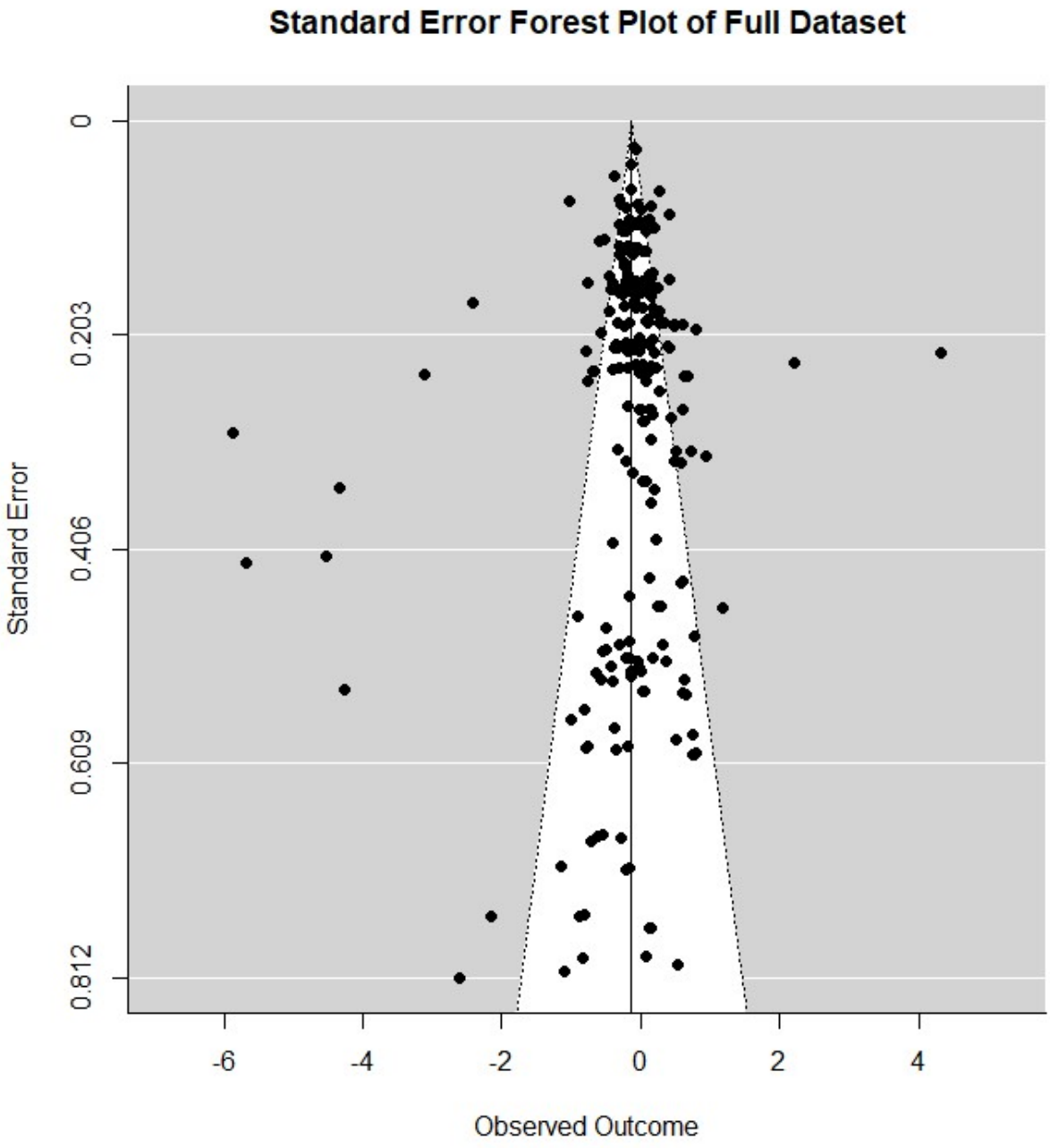

Figure 2. Funnel plot addressing publication bias. In a study with minimal publication bias, data should be symmetrically spread, with the majority of data within the indicated cone.

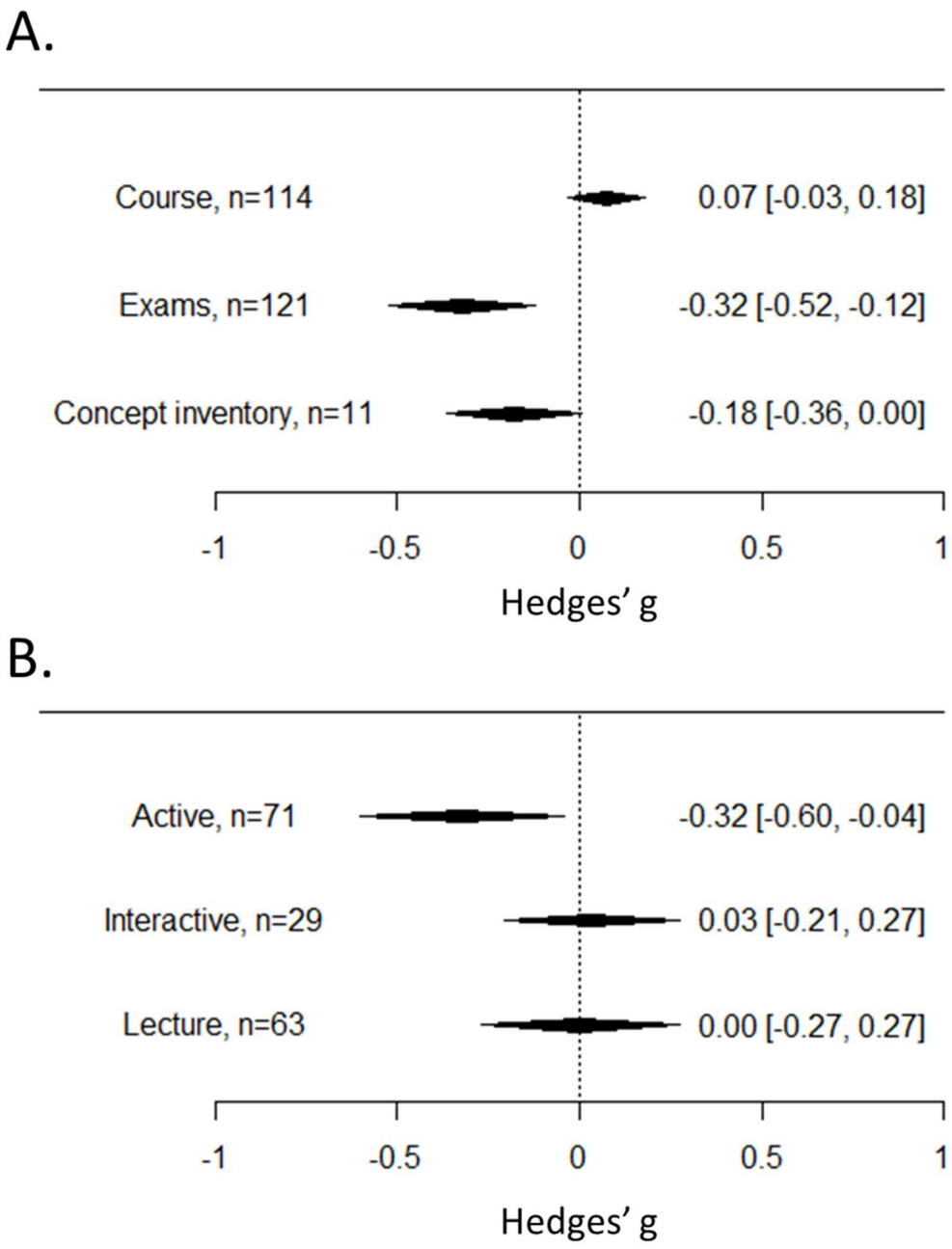

Figure 3. Gender gaps within subgroups. (A) Gaps within different assessment types. (B) Gaps within different pedagogies. A negative Hedges' g indicates men overperformed compared to women, in uncorrected (i.e., not model-based) units of standard deviation.

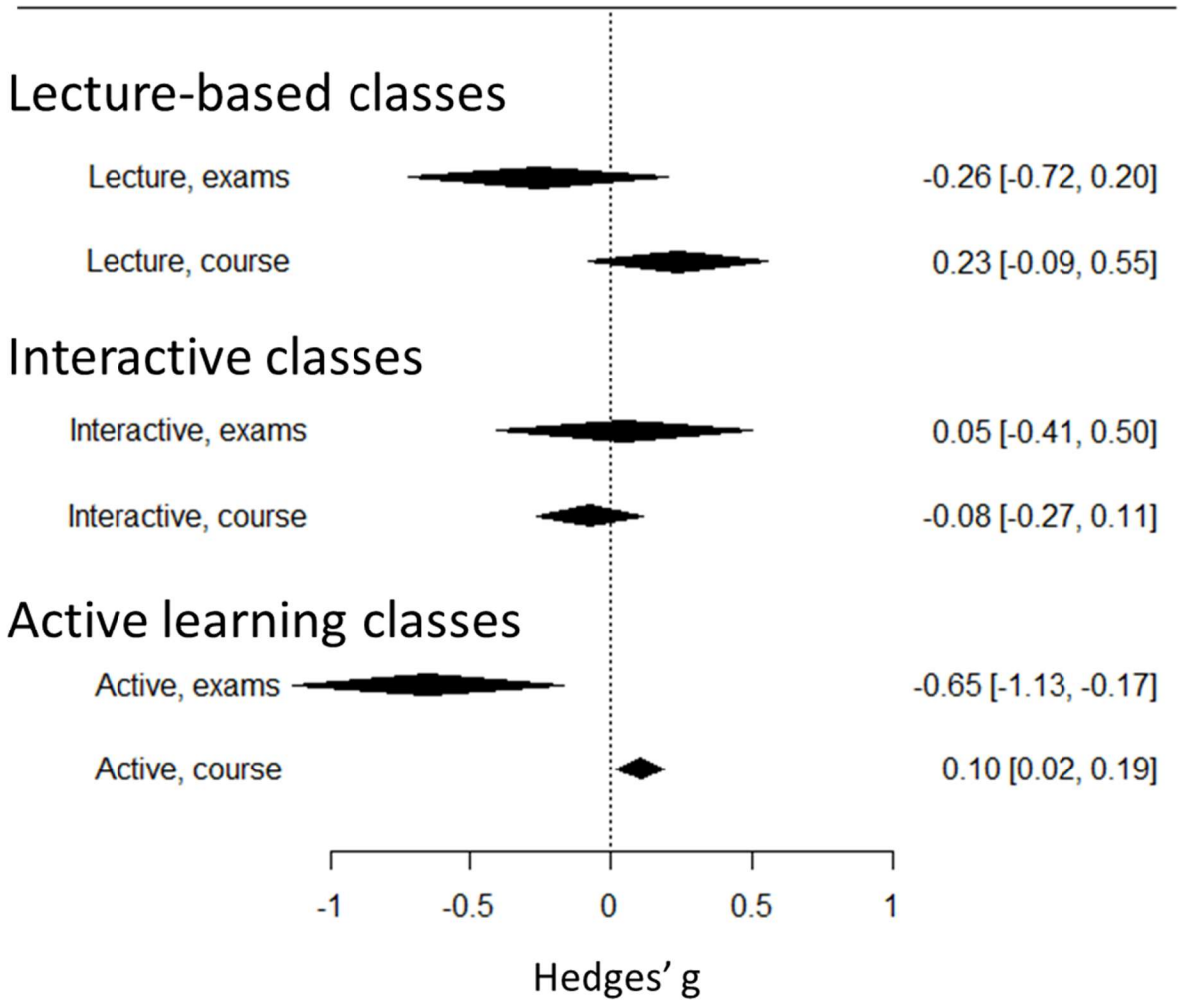

Figure 4. Gender gaps by pedagogy and assessment type. A negative Hedges' g indicates men overperformed compared to women, in uncorrected units of standard deviation.

Literature Cited

Achilles, C. M. (2012). Class-Size Policy: The STAR Experiment and Related Class-Size Studies. NCPEA Policy Brief. Volume 1, Number 2. In *NCPEA Publications*. NCPEA Publications. 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

Ainsworth, C. (2015). Sex redefined. *Nature News, 518*(7539), 288. https://doi.org/10.1038/518288a

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

Arnold, T. W. (2010). Uninformative Parameters and Model Selection Using Akaike's Information Criterion. *The Journal of Wildlife Management, 74*(6), 1175–1178. https://doi.org/10.1111/j.1937-2817.2010.tb01236.x

Baker, B. D., Farrie, D., & Sciarra, D. G. (2016). Mind the Gap: 20 Years of Progress and Retrenchment in School Funding and Achievement Gaps. *ETS Research Report Series, 2016*(1), 1–37. https://doi.org/10.1002/ets2.12098

Ballen, C. J., Aguillon, S. M., Awwad, A., Bjune, A. E., Challou, D., Drake, A. G., Driessen, M., Ellozy, A., Ferry, V. E., Goldberg, E. E., Harcombe, W., Jensen, S., Jørgensen, C., Koth,

Z., McGaugh, S., Mitry, C., Mosher, B., Mostafa, H., Petipas, R. H., … Cotner, S. (2019). Smaller Classes Promote Equitable Student Participation in STEM. *BioScience, 69*(8), 669–680. https://doi.org/10.1093/biosci/biz069

Ballen, C. J., Aguillon, S. M., Brunelli, R., Drake, A. G., Wassenberg, D., Weiss, S. L., Zamudio, K. R., & Cotner, S. (2018). Do Small Classes in Higher Education Reduce Performance Gaps in STEM? *BioScience, 68*(8), 593–600. https://doi.org/10.1093/biosci/biy056

Ballen, C. J., Salehi, S., & Cotner, S. (2017). Exams disadvantage women in introductory biology. *PLOS ONE, 12*(10), e0186419. https://doi.org/10.1371/journal.pone.0186419

Ballen, C. J., Wieman, C., Salehi, S., Searle, J. B., & Zamudio, K. R. (2017). Enhancing Diversity in Undergraduate Science: Self-Efficacy Drives Performance Gains with Active Learning. *CBE—Life Sciences Education, 16*(4), ar56. https://doi.org/10.1187/cbe.16-12-0344

Barton, K. (2020). *MuMIn: Multi-Model Inference*. R package version 1.43.17.

Beichner, R. J., Saul, J. M., Abbott, D. S., Morse, J. J., Allain, R. J., Bonham, S. W., Dancy, M. H., & Risley, J. S. (2007). *The Student-Centered Activities for Large Enrollment Undergraduate Programs (SCALE-UP) Project*. 42.

Brooks, C. I., & Mercincavage, J. E. (1991). Grades for Men and Women in College Courses Taught by Women. *Teaching of Psychology, 18*(1), 47–48. https://doi.org/10.1207/s15328023top1801_17

Caldwell, J. E. (2007). Clickers in the Large Classroom: Current Research and Best-Practice Tips. *CBE—Life Sciences Education, 6*(1), 9–20. https://doi.org/10.1187/cbe.06-12-0205

Casper, A. M., Eddy, S. L., & Freeman, S. (2019). True Grit: Passion and persistence make an innovative course design work. *PLOS Biology, 17*(7), e3000359. https://doi.org/10.1371/journal.pbio.3000359

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press.

Cohen, M., Buzinski, S. G., Armstrong-Carter, E., Clark, J., Buck, B., & Reuman, L. (2019). Think, pair, freeze: The association between social anxiety and student discomfort in the active learning environment. *Scholarship of Teaching and Learning in Psychology, 5*(4), 265–277. https://doi.org/10.1037/stl0000147

Cotner, S., & Ballen, C. J. (2017). Can mixed assessment methods make biology classes more equitable? *PLOS ONE, 12*(12), e0189610. https://doi.org/10.1371/journal.pone.0189610

Creech, L. R., & Sweeder, R. D. (2012). Analysis of Student Performance in Large-Enrollment Life Science Courses. *CBE—Life Sciences Education, 11*(4), 386–391. https://doi.org/10.1187/cbe.12-02-0019

Crowther, G., Wiggins, B., & Jenkins, L. (2020). Testing in the Age of Active Learning: Test Question Templates Help to Align Activities and Assessments. *HAPS Educator, 24*(1), 592–599. https://doi.org/10.21692/haps.2020.006

Cuseo, J. (2007). *THE EMPIRICAL CASE AGAINST LARGE CLASS SIZE: ADVERSE EFFECTS ON THE TEACHING, LEARNING, AND RETENTION OF FIRST- YEAR STUDENTS*. 26.

DiDonato, L., & Strough, J. (2013). Do College Students' Gender-typed Attitudes About Occupations Predict Their Real-World Decisions? *Sex Roles, 68*(9–10), 536–549. https://doi.org/10.1007/s11199-013-0275-2

Dufresne, R. J., Leonard, W. J., & Gerace, W. J. (2002). Marking sense of students' answers to multiple-choice questions. *The Physics Teacher, 40*(3), 174–180. https://doi.org/10.1119/1.1466554

Eddy, S. L., & Brownell, S. E. (2016). Beneath the numbers: A review of gender disparities in undergraduate education across science, technology, engineering, and math disciplines. *Physical Review Physics Education Research, 12*(2), 020106. https://doi.org/10.1103/PhysRevPhysEducRes.12.020106

Eddy, S. L., & Hogan, K. A. (2014). Getting Under the Hood: How and for Whom Does Increasing Course Structure Work? *CBE—Life Sciences Education, 13*(3), 453–468. https://doi.org/10.1187/cbe.14-03-0050

Eichler, J. F., & Peeples, J. (2016). Flipped classroom modules for large enrollment general chemistry courses: A low barrier approach to increase active learning and improve student grades. *Chemistry Education Research and Practice, 17*(1), 197–208. https://doi.org/10.1039/C5RP00159E

Feldman, J. (2018). *Grading for Equity: What It Is, Why It Matters, and How It Can Transform Schools and Classrooms*. Corwin Press.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences, 111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Furukawa, T. A., Barbui, C., Cipriani, A., Brambilla, P., & Watanabe, N. (2006). Imputing missing standard deviations in meta-analyses can provide accurate results. *Journal of Clinical Epidemiology, 59*(1), 7–10. https://doi.org/10.1016/j.jclinepi.2005.06.006

Glass, G. V. (1976). Primary, Secondary, and Meta-Analysis of Research. *Educational Researcher, 5*(10), 3–8. https://doi.org/10.3102/0013189X005010003

Goulden, M. (2009). *Patching America's Leaky Pipeline in the Sciences*. 52.

Grandy, J. (1994). Gender and Ethnic Differences Among Science and Engineering Majors: Experiences, Achievements, and Expectations. *ETS Research Report Series, 1994*(1), i–63. https://doi.org/10.1002/j.2333-8504.1994.tb01603.x

Haak, D. C., HilleRisLambers, J., Pitre, E., & Freeman, S. (2011). Increased Structure and Active Learning Reduce the Achievement Gap in Introductory Biology. *Science, 332*(6034), 1213–1216. https://doi.org/10.1126/science.1204820

Hansen, M. J., & Birol, G. (2014). Longitudinal Study of Student Attitudes in a Biology Program. *CBE—Life Sciences Education, 13*(2), 331–337. https://doi.org/10.1187/cbe.13-06-0124

Harris, R. B., Mack, M. R., Bryant, J., Theobald, E. J., & Freeman, S. (2020). Reducing achievement gaps in undergraduate general chemistry could lift underrepresented students into a "hyperpersistent zone." *Science Advances, 6*(24), eaaz5687. https://doi.org/10.1126/sciadv.aaz5687

Harris, Rebecca B., Grunspan, D. Z., Pelch, M. A., Fernandes, G., Ramirez, G., & Freeman, S. (2019). Can Test Anxiety Interventions Alleviate a Gender Gap in an Undergraduate STEM Course? *CBE—Life Sciences Education, 18*(3), ar35. https://doi.org/10.1187/cbe.18-05-0083

Hedges, L. V. (1981). Distribution Theory for Glass's Estimator of Effect size and Related Estimators: *Journal of Educational Statistics 6*(2), 107–128. https://doi.org/10.3102/10769986006002107

Henning, J. A., Ballen, C. J., Molina, S. A., & Cotner, S. (2019). Hidden Identities Shape Student Perceptions of Active Learning Environments. *Frontiers in Education, 4*, 129. https://doi.org/10.3389/feduc.2019.00129

Ho, D. E., & Kelman, M. G. (2014). Does Class Size Affect the Gender Gap? A Natural Experiment in Law. *The Journal of Legal Studies, 43*(2), 291–321. https://doi.org/10.1086/676953

Hothorn, T., Bretz, F., Ag, N. P., & Westfall, P. (2008). Simultaneous Inference in General Parametric Models. *Biometric Journal, 50*(3), 346–363.

Kim, Y., Jeong, S., Ji, Y., Lee, S., Kwon, K. H., & Jeon, J. W. (2015). Smartphone Response System Using Twitter to Enable Effective Interaction and Improve Engagement in Large Classrooms. *IEEE Transactions on Education, 58*(2), 98–103. https://doi.org/10.1109/TE.2014.2329651

Kling, K. C., Noftle, E. E., & Robins, R. W. (2013). Why Do Standardized Tests Underpredict Women's Academic Performance? The Role of Conscientiousness. *Social Psychological and Personality Science, 4*(5), 600–606. https://doi.org/10.1177/1948550612469038

Knight, J. K., Wise, S. B., & Southard, K. M. (2013). Understanding Clicker Discussions: Student Reasoning and the Impact of Instructional Cues. *CBE—Life Sciences Education, 12*(4), 645–654. https://doi.org/10.1187/cbe.13-05-0090

Koester, B. P., Grom, G., & McKay, T. A. (2016). Patterns of Gendered Performance Difference in Introductory STEM Courses. ArXiv:1608.07565 [Physics]. http://arxiv.org/abs/1608.07565

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods, 2*(1), 61–76. https://doi.org/10.1002/jrsm.35

Lauer, S., Momsen, J., Offerdahl, E., Kryjevskaia, M., Christensen, W., & Montplaisir, L. (2013). Stereotyped: Investigating Gender in Introductory Science Courses. *CBE—Life Sciences Education, 12*(1), 30–38. https://doi.org/10.1187/cbe.12-08-0133

Lopez, E. J., Shavelson, R. J., Nandagopal, K., Szu, E., & Penn, J. (2014). Factors Contributing to Problem-Solving Performance in First-Semester Organic Chemistry. *Journal of Chemical Education, 91*(7), 976–981. https://doi.org/10.1021/ed400696c

Lorenzo, M., Crouch, C. H., & Mazur, E. (2006). Reducing the gender gap in the physics classroom. *American Journal of Physics, 74*(2), 118–122. https://doi.org/10.1119/1.2162549

Lowry, P. B., Romano Jr., N. C., & Guthrie, R. (2006). Explaining and Predicting Outcomes of Large Classrooms Using Audience Response Systems. *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06), 1*, 4c–4c. https://doi.org/10.1109/HICSS.2006.173

Madsen, A., McKagan, S. B., & Sayre, E. C. (2013). Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap? *Physical Review Special Topics - Physics Education Research, 9*(2), 020121. https://doi.org/10.1103/PhysRevSTPER.9.020121

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207–218. https://doi.org/10.1207/s15326985ep3404_2

Matz, R. L., Koester, B. P., Fiorini, S., Grom, G., Shepard, L., Stangor, C. G., Weiner, B., & McKay, T. A. (2017). Patterns of Gendered Performance Differences in Large Introductory Courses at Five Research Universities. *AERA Open, 3*(4), 2332858417743754. https://doi.org/10.1177/2332858417743754

McCullough, L. (2013). *Gender, Context, and Physics Assessment. Journal of International Women's Studies, 5*(4), 20–30.

Miller, S., & Tanner, K. D. (2015). A Portal into Biology Education: An Annotated List of Commonly Encountered Terms. *CBE Life Sciences Education, 14*(2). https://doi.org/10.1187/cbe.15-03-0065

Miyake, A., Kost-Smith, L. E., Finkelstein, N. D., Pollock, S. J., Cohen, G. L., & Ito, T. A. (2010). Reducing the Gender Achievement Gap in College Science: A Classroom Study of Values Affirmation. *Science, 330*(6008), 1234–1237. https://doi.org/10.1126/science.1195996

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, T. P. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLOS Medicine, 6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences, 109*(41), 16474–16479. https://doi.org/10.1073/pnas.1211286109

Newsome, J. L. (2008). The chemistry PhD: the impact on women's retention. *RSC*.

Nijenkamp, R., Nieuwenstein, M. R., Jong, R. de, & Lorist, M. M. (2016). Do Resit Exams Promote Lower Investments of Study Time? Theory and Data from a Laboratory Study. *PLOS ONE, 11*(10), e0161708. https://doi.org/10.1371/journal.pone.0161708

Peters, M. L. (2013). EXAMINING THE RELATIONSHIPS AMONG CLASSROOM CLIMATE, SELF-EFFICACY, AND ACHIEVEMENT IN UNDERGRADUATE MATHEMATICS: A MULTI-LEVEL ANALYSIS. *International Journal of Science and Mathematics Education, 11*(2), 459–480. https://doi.org/10.1007/s10763-012-9347-y

Pollock, S. J., Finkelstein, N. D., & Kost, L. E. (2007). Reducing the gender gap in the physics classroom: How sufficient is interactive engagement? *Physical Review Special Topics - Physics Education Research, 3*(1), 010107. https://doi.org/10.1103/PhysRevSTPER.3.010107

R Core Team. (2019). R: A language and environment for statistical computing. *R Foundation for Statistical Computing.*

Rask, K., & Tiefenthaler, J. (2008). The role of grade sensitivity in explaining the gender imbalance in undergraduate economics. *Economics of Education Review, 27*(6), 676–687. https://doi.org/10.1016/j.econedurev.2007.09.010

Rasmussen, M. L. (2009). Beyond gender identity? *Gender and Education, 21*(4), 431–447. https://doi.org/10.1080/09540250802473958

Rauschenberger, M. M., & Sweeder, R. D. (2010). Gender performance differences in biochemistry. *Biochemistry and Molecular Biology Education, 38*(6), 380–384. https://doi.org/10.1002/bmb.20448

Rodriguez, M., Mundy, M.-A., Kupczynski, L., & Challoo, L. (2018). Effects of Teaching Strategies on Student Success, Persistence, and Perceptions of Course Evaluations. *Research in Higher Education Journal, 35*. https://eric.ed.gov/?id=EJ1194444

Rosenberg, M. S. (2005). The File-Drawer Problem Revisited: A General Weighted Method for Calculating Fail-Safe Numbers in Meta-Analysis. *Evolution, 59*(2), 464–468. https://doi.org/10.1111/j.0014-3820.2005.tb01004.x

RStudio Team. (2019). RStudio: Integrated Development Enviromnent for R. *RStudio, Inc*.

Saiz, M. (2014). Economies of Scale and Large Classes. *The NEA Higher Education Journal,* 149–160.

Salehi, S., Cotner, S., Azarin, S. M., Carlson, E. E., Driessen, M., Ferry, V. E., Harcombe, W., McGaugh, S., Wassenberg, D., Yonas, A., & Ballen, C. J. (2019). Gender Performance Gaps Across Different Assessment Methods and the Underlying Mechanisms: The Case of Incoming Preparation and Test Anxiety. *Frontiers in Education, 4*. https://doi.org/10.3389/feduc.2019.00107

Schwartz, D. L., Cheng, K. M., Salehi, S., & Wieman, C. (2016). The half empty question for socio-cognitive interventions. *Journal of Educational Psychology, 108*(3), 397–404. https://doi.org/10.1037/edu0000122

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

Solorzano, D. G. (1992). An Exploratory Analysis of the Effects of Race, Class, and Gender on Student and Parent Mobility Aspirations. *The Journal of Negro Education*, *61*(1), 30–44. JSTOR. https://doi.org/10.2307/2295627

Sonnert, G., & Fox, M. F. (2012). Women, Men, and Academic Performance in Science and Engineering: The Gender Difference in Undergraduate Grade Point Averages. The *Journal of Higher Education, 83*(1), 73–101. https://doi.org/10.1353/jhe.2012.0004

Stains, M., Harshman, J., Barker, M. K., Chasteen, S. V., Cole, R., DeChenne-Peters, S. E., Eagan, M. K., Esson, J. M., Knight, J. K., Laski, F. A., Levis-Fitzgerald, M., Lee, C. J., Lo, S. M., McDonnell, L. M., McKay, T. A., Michelotti, N., Musgrove, A., Palmer, M. S., Plank, K. M., … 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/10.1080/09500690010025067

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., Chambwe, N., Cintrón, D. L., Cooper, J. D., Dunster, G., Grummer, J. A., Hennessey, K., Hsiao, J., Iranon, N., Jones, L., Jordt, H., Keller, M., Lacey, M. E., Littlefield, C. E., … Freeman, S. (2020). Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math. *Proceedings of the National Academy of Sciences, 117*(12), 6476–6483. https://doi.org/10.1073/pnas.1916903117

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

Valencia, R. R. (2012). *The Evolution of Deficit Thinking: Educational Thought and Practice*. Routledge.

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., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

CHAPTER TWO

Measuring engagement via galvanic skin response across multiple in-class activities

Abstract

Incorporating different class activities may improve student engagement, and thus student learning. I analyzed galvanic skin response (GSR) data from 14 students in 5 undergraduate classes using linear mixed model analysis to determine the impact of multiple in-class activities (lecturing, clicker questions, demonstrations, and other questions) and gender on classroom engagement. Activity and the interaction between gender and activity were both significant predictors of GSR change. Clicker questions showed the greatest increase in student engagement overall, and there was no statistically significant difference between men and women's responses within each activity. Thus, common institutional practices impact student engagement in the classroom, and identifying how these practices impact students differently using biometric measures can inform evidence-based teaching.

Introduction

Student engagement is an important factor in the classroom setting. Students who are engaged with their education have better educational outcomes, including higher graduation rates, higher grades, and better critical thinking skills (Carini et al., 2006; Medicine et al., 2003). Engagement is a multifaceted concept that can cover a wide variety of definitions and aspects. Fredricks et al. (2004) proposes a theoretical framework that focuses on three major aspects of engagement: *behavioral engagement*, which focuses on student behaviors and participation; c*ognitive engagement,* which focuses on psychological efforts put towards learning and

comprehending new material; and *emotional engagement,* which focuses on positive and negative emotions elicited in response to elements of the learning experience. In this study, I focus specifically on emotional engagement.

Employing educational approaches that engage students' emotions can help overcome misconceptions and lead to better understanding (Irwin & Wynne, 2003; Pekrun et al., 2002; Pintrich et al., 1993; Weber, 2010). Physiologically, neurotransmitters related to emotional arousal are important in the process of forming memories (Hu et al., 2007). Thus, emotions should not be ignored when studying learning, and furthermore, may give us a means to measure student engagement and learning.

Often, student engagement is measured either through observation (Smith et al., 2013; Witkowski & Cornell, 2015) or through self-assessment surveys (Appleton et al., 2006; Goldberg & Ingram, 2011; Marks, 2000; Pekrun et al., 2002). However, these methods are both subjective and can be difficult to verify. For instance, observations typically rely on behavioral indicators, but can be less accurate in identifying cognitive and emotional states (Peterson et al., 1984). Self-assessments are typically administered after the event in question, and require accurate self-reflection and memory of the events, which can be difficult to verify. In response to this, biometric measures of student engagement have been developed and used.

Emotional arousal is related to the sympathetic nervous system, which controls a number of physiological processes that can be measured using biometric readings (Malmivuo et al., 1995). Once such process is sweat glands, which can change in response to arousal and attention, in turn changing skin conductance (Boucsein, 2012). As such, devices that measure galvanic skin response (GSR) allow researchers to use these measures as indicators of emotional arousal (Poh

et al., 2010). GSR has previously been used in learning environments to measure student engagement (Dragon et al., 2008; Hardy et al., 2013; McNeal et al., 2014, 2020).

In this study, I use GSR to measure the responses of undergraduate students in introductory classes to various class activities. Previous research has focused on ways to increase student engagement in the classroom, often through the use of various active learning techniques (AlKandari, 2012; Goldberg & Ingram, 2011; Witkowski & Cornell, 2015). While the classes analyzed in this study rely heavily on lecture-based instruction, and thus would not be considered full "active learning" classes, they do incorporate elements that may elicit greater engagement than uninterrupted lecturing, such as personal response systems (i.e. "clickers") and in-class demonstrations. Because of how common lecture-based classes are, especially in introductory level classes (Stains et al., 2018), understanding the way that activities that can be easily integrated into such classroom environments will give instructors a better grasp on managing student engagement.

In addition to the question of what activities elicit increased student engagement, I am also interested in the effect of gender on how students respond to different activities. Previous research has found variation in student participation, with women interacting in the classroom less frequently than men (Aguillon et al., 2020; Crombie et al., 2003). As such, to encourage equitable participation within classrooms, it is important to understand how men and women potentially engage differently with various class activities.

In this study, I address the following specific research questions:

1. How do different classroom activities affect student engagement?
2. Is there a gender difference in the response to activities?

Methods

Data collection

Data was collected at a large midwestern university, across 5 undergraduate level classes during the spring and summer semesters of 2018 (Table 6). In order to collect data about students' responses to in-class activities, student participants wore GSR monitors on multiple non-exam class periods. Monitors strapped to students' fingers on their non-dominant hands. Readings began five minutes before class start time, during which students were asked to limit activity in order to create a "non-engaged" pre-class baseline. During these same class periods, an observer trained in the Classroom Observation Protocol for Undergraduate STEM (COPUS) (Smith et al., 2013) used this protocol to document the occurrence of instructor and student activities within two minute time blocks. I used the patterns present in these observations to create 6 distinct categories that could be assigned to each GSR reading as appropriate: pre-class, lecturing, clicker questions, demonstrations, other questions, and uncategorized (Table 7).

Data processing

I organized and processed data in Excel prior to statistical analysis. Statistical analyses were performed in R version 3.6.2 (R Core Team, 2019) in RStudio version 1.2.5033 (RStudio Team, 2019). I used the tidyverse package (Wickham et al., 2019) to produce graphics and streamline code, the lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2017) packages to fit mixed models, and the multcomp package (Hothorn et al., 2008) for pairwise comparisons.

I aligned GSR measurements and activity designations based on time stamps. Any GSR readings that took place after the last time block with a categorize activity was removed, as this indicated readings that were captured after class concluded but before the student shut off the

device. For each distinct set of GSR collections (i.e. different students, days, and class), I used the GSR readings designated as "pre-class" to calculate an average baseline GSR response, then subtracted that average from all of the remaining GSR readings for that collection to create an "adjusted GSR" that I used in the remaining analyses. I calculated the mean of each group and created a boxplot of adjusted GSR broken down by activity and gender.

Statistical analyses

I excluded data with "preclass" and "uncategorized" activity designations in my statistical analyses, focusing instead on the differences between lecturing and the other three specified activities: clicker questions, demonstrations, and other questions; as well as gender comparisons within each activity.

I used a linear mixed effects model to assess the effects of activity and gender on adjusted GSR readings. I used a theory-based approach to build my model, and included activity and gender, with interactions, in order to identify variation by gender within each activity, as fixed effects. Class and student ID were set as random effects to avoid issues of non-independence arising from repeated collections of data from the same students and within the same classes. Of the activities included, lecturing was set as the basis for the intercept, allowing for the slopes generated by this model to represent changes introduced by non-lecturing activities.

To compare men and women's responses to each activity, I created a dummy interaction term combining gender and activity designations. I fit a linear mixed effects model using this interaction term as the fixed effect, with class and student ID as random effects, as in the previous model. I removed the intercept from this model in order to get a distinct slope for each gender-activity combination. I used the results from this model to run pairwise comparisons via

general linear hypotheses between estimates for men and women in each activity (ex. "men-lecturing" – "women-lecturing"). These comparisons used Bonferroni corrections to account for familywise error rates (McDonald, 2014).

Results

A total of 60175 individual GSR measures were collected. Figure 5 shows an example of student GSR readings plotted across a class period. After removal of "pre-class" and "uncategorized" readings, 51378 individual readings were included in statistical analyses (Figure 6). The average adjusted GSR for lecturing was 3014.34μS, for clicker questions was 4589.91μS, for demonstrations was 2998.90μS, and for other questions was 2718.47μS (Figure 7).

Activity was a significant predictor of GSR. Gender on its own was not significant (*p*-value-0.60670), but the interaction between gender and activity was a significant predictor (Table 8). For women, clicker questions caused an increase in GSR over lecturing, while demonstrations and other questions showed a drop in GSR. For men, clicker questions, demonstrations, and other questions all caused an increase in GSR.

Pairwise comparisons of gender within each activity showed the largest difference in response between men and women was in demonstrations, though this difference was not statistically significant within the model (*p*-value=0.0747) (Table 9).

Discussion

My results showed variation in GSR readings between lecturing and non-lecturing activities, (i.e. clicker questions, demonstrations, and other questions). Using this as a proxy for engagement, I conclude that class activity does impact student engagement. Clicker questions

elicited the highest student engagement. While graphics suggest a slightly higher response from men, there was no statistically significant impact of gender.

The use of clickers and other personal response systems have been previously utilized as an active learning technique that show improvement in student understanding (Caldwell, 2007; Knight et al., 2013; Smith et al., 2009). By requiring students to consider, and sometimes discuss, questions posed in class, students are required to mentally interact with the material, rather than passively listening to it. While my results did not show a significant difference in response between men and women, previous research has found that women may have a more favorable view of clicker questions in class (Kang et al., 2012; Niemeyer & Zewail-Foote, 2018; Wolter et al., 2011), and the use of clickers may have a disproportionate benefit on learning for women (Kang et al., 2012; King & Joshi, 2008).

Demonstrations have also been integrated into classes as an active learning strategy and have been previously linked to improved engagement (Morgan et al., 2007). While demonstrations may not always require direct student action (although some might call upon student volunteers), they are designed to be interesting and illustrative. Demonstrations have also been shown to improve conceptual understanding of topics (Anderson-Cook & Dorai-Raj, 2001; Mazzolini et al., 2012; Sokoloff & Thornton, 1997). Previous studies have specifically looked gender differences in the effectiveness of demonstrations, with some showing gender differences (Uhumuavbi & Mamudu, 2009) and others not (Adekoya et al., 2011). However, those studies focused on demonstrations in specific environments. Similarly, a significant portion of demonstration GSR readings in this study were collected from the same class. Thus, it may be difficult to make blanket statements of the true effectiveness, and potential gender bias, of demonstrations from limited examples.

Questions posed through a class that aren't associated with personal response systems, usually in the form of spontaneous questions by students or questions posed by instructors that are answered by a volunteer or selected student (what I referred to in this study as "other questions"), also play an important role in classroom engagement. While they are not specifically associated with active learning in the way that clicker questions and demonstrations have been, they nevertheless promote interaction between instructor and student. Questions can help an instructor understand elements of a topic that need clarification (Schwarz et al., 2017), and questions posed by an instructor, especially if they are higher order questions, can aid in developing and directing student comprehension (McComas & Abraham, 2004; Singer, 1978). Thus, creating an atmosphere that encourages the asking and answering of questions can create an atmosphere where students engage with the material actively rather than passively.

My results lend physiological support to the idea that integrating elements such as clickers, demonstrations, and questions can improve student engagement in the classroom. However, it is important to recognize that this effect is not uniform for all students.

Lecture heavy courses run the risk of losing students' attention during the course but incorporating different activities into the curriculum may help boost student engagement. While specific reactions varied between students, class activities, especially clicker questions, increased GSR. Thus, these activities would be a good addition to classes that are otherwise lecture heavy. Future studies may want to expand on the impact of activities that were here grouped under "uncategorized" due to not occurring often enough to designate their own category in this study, such as group work, as well as more detailed investigations into ways to make activities as equally engaging as possible.

Limitations

There are a number of limitations that should be considered within this study.

First, other factors besides classroom engagement could influence GSR. While GSR measures emotional arousal, it does not indicate if these are positive emotions, and I could not control for outside distractions and stressors that could be affecting students' focus and response to classroom activities. I did not consider classroom temperature, but increased temperature can cause students to sweat, which effects their skin conductance. And while the GSR devices used are designed to be minimally invasive, they are still noticeable, and knowing that they are being observed could change the way students focus and respond in class.

Second, GSR reading in this study were highly variable, with many outliers in each group, and thus, model estimates all had large standard errors. The only control or adjustment made to the data was to subtract pre-class average from raw GSR. While this was designed to account for some of the underlying variation between different students, it did not account for the way that individual readings could vary wildly from the mean. This variation, unaccounted for by any sort of data cleaning, may be obscuring some of the group differences.

Finally, the activity categorizations were not very specific. Categories were determined based on the presence of an event but could not determine where in that time period the event happened. As such, if an event, such as a question, occurred towards the end of the time block, all of the GSR readings that occurred during the beginning of the time period would still be analyzed as if they occurred in response to that activity. In the case of the "other questions" category, this category denotes that a question was asked, but not who asked it, and it is not unreasonable to expect different responses depending of if the question was asked by an instructor (who might be expecting a response from the student), a peer (whose question may or

may not be perceived as relevant to the observed student), or even the observed student themselves (who in that case would be taking a more active role). Additionally, I did not account for the timing of events within the full class period, which could indicate compounding effects of sequential activities, delayed responses, or simply a drop in attention over time.

Likely, a more comprehensive analysis of GSR fluctuations across classes, particularly ones that account for background noise include more specific information about class activities and participation, could reveal a more comprehensive understanding of student engagement. However, even with these shortcomings, the data still provides physiological evidence that clicker questions, demonstrations, and other questions can be used by instructors to increase student engagement.

Semester	Class	Students
Spring 2018	**General Chemistry**	3 Students
	Earth Sciences	2 Students
	Human Evolution	2 Students
Summer 2018	**Introductory Physics I**	5 Students
	Introductory Physics II	4 Students

Table 6. Classes included in GSR collection. Classes are broken down by semester, and the number of students monitored in each class is listed. Some students participated in more than one class, and not every student within a class was monitored on every day that data was collected in that class.

Activity categorization	Description
Pre-class	5-minute period before class began, used to establish baseline GSR. Designated by time rather than COPUS description.
Lecturing	Instructor was speaking and/or writing on the board, students were passively listening. No other activities occurred.
Clicker Questions	Students were presented with a question that they answered using an audience response device (such as iClickers and similar devices).
Demonstrations	A demonstration of video was shown.
Other Questions	Questions asked and/or answered by either a student or instructor, not tied with audience response devices.
Uncategorized	Any activity patterns that did not fit into prior categorizations. Mostly contained time-blocks containing the "other" activity on the COPUS checklist, but also included a few events that did not fit prior categories but did not occur enough to merit the creation of a new category.

Table 7. Activity categories and descriptions. With the exception of 'pre-class', designations were made based on which COPUS activities were present in each time block.

Regression Coefficient	Estimate ± SE	*p*-value
Intercept	2197.29 ± 1184.60	0.08269
Activity (reference level: Lecturing)		
Clicker Questions	1031.10 ± 180.72	**1.17e-08**
Demonstrations	-1119.36 ± 99.11	**<2e-16**
Other Questions	-225.73± 86.27	**0.00889**
Gender (reference level: Women)		
Male	843.72 ±1604.68	0.60670
*Activity*gender*		
Clicker Questions:Men	1336.83±254.00	**1.42e-07**
Demonstrations:Men	2932.59±139.23	**<2e-16**
Other Questions:Men	685.98±114.00	**1.79e-09**

Table 8. Mixed model results of GSR response. Factors with significant *p*-values are bolded.

Comparison	Estimate±SE	*p*-value
Men - Women		
Lecturing	843.7 ±1604.7	1.0000
Clicker Questions	2180.5 ±1619.8	0.7129
Demonstrations	3776.3 ±1605.7	0.0747
Other Questions	1529.7 ±1604.1	1.0000

Table 9. Pairwise comparison of gendered response within activity. P values are adjusted using Bonferroni correction.

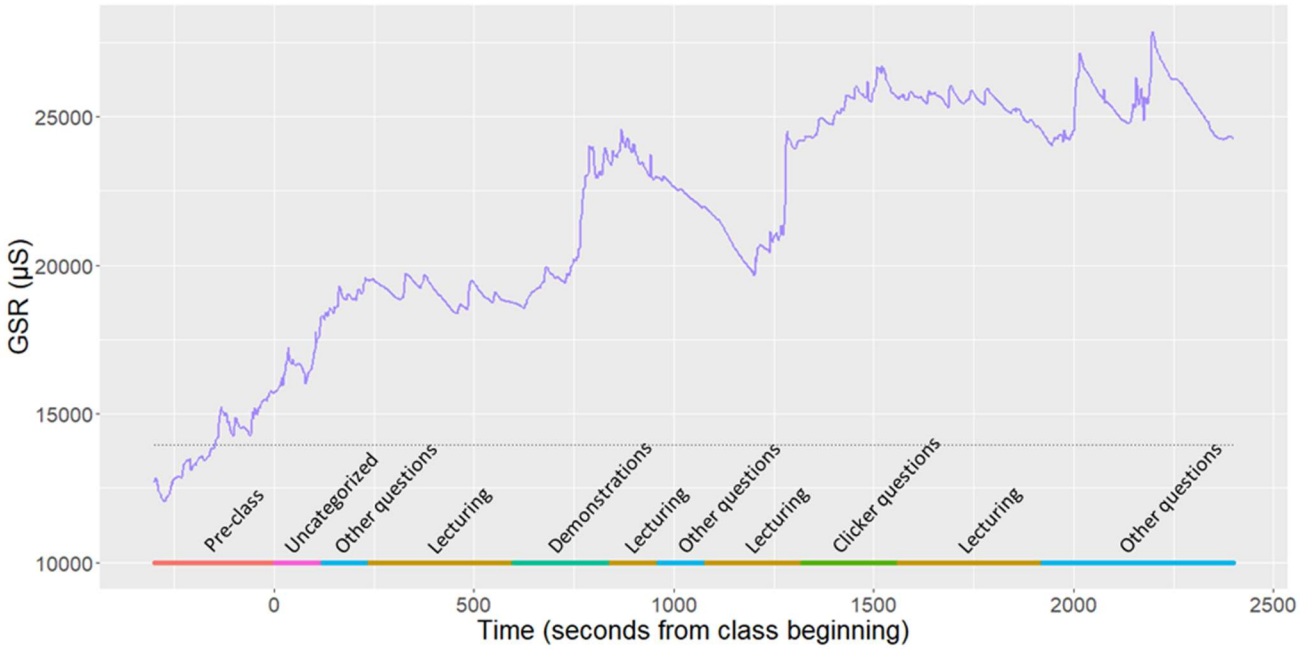

Figure 5. Example graph of a student's GSR readings during class. Class activity is noted at the bottom of the graph. The dotted line shows the pre-class baseline average, which was used to adjust GSR readings for statistical analysis.

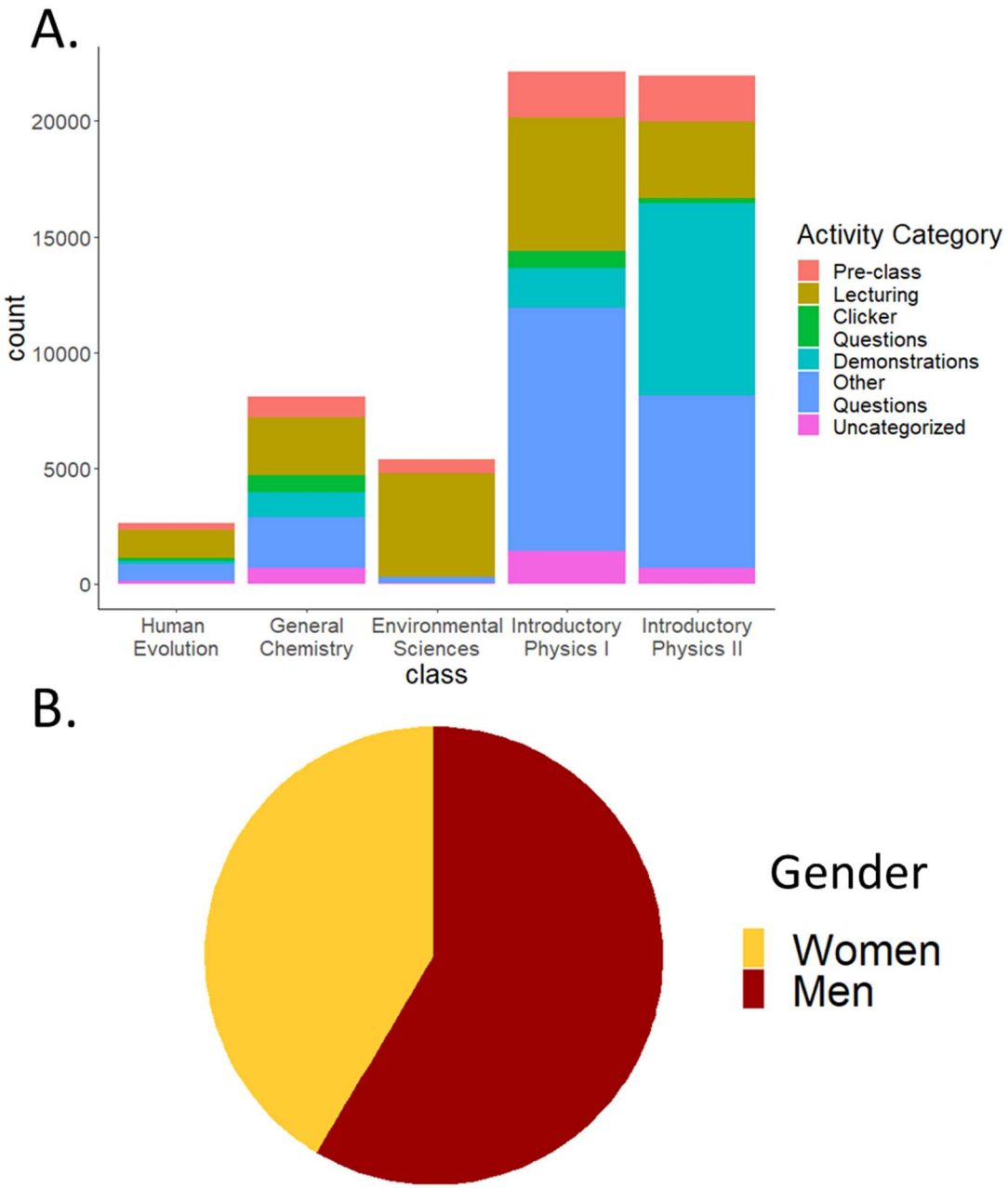

Figure 6. Context of GSR collection. (A) Number of GSR readings collected in each class, broken down by activity. GSR readings were collected every two seconds. (B) Proportion of GSR readings from women (41.5%) and men (58.5%).

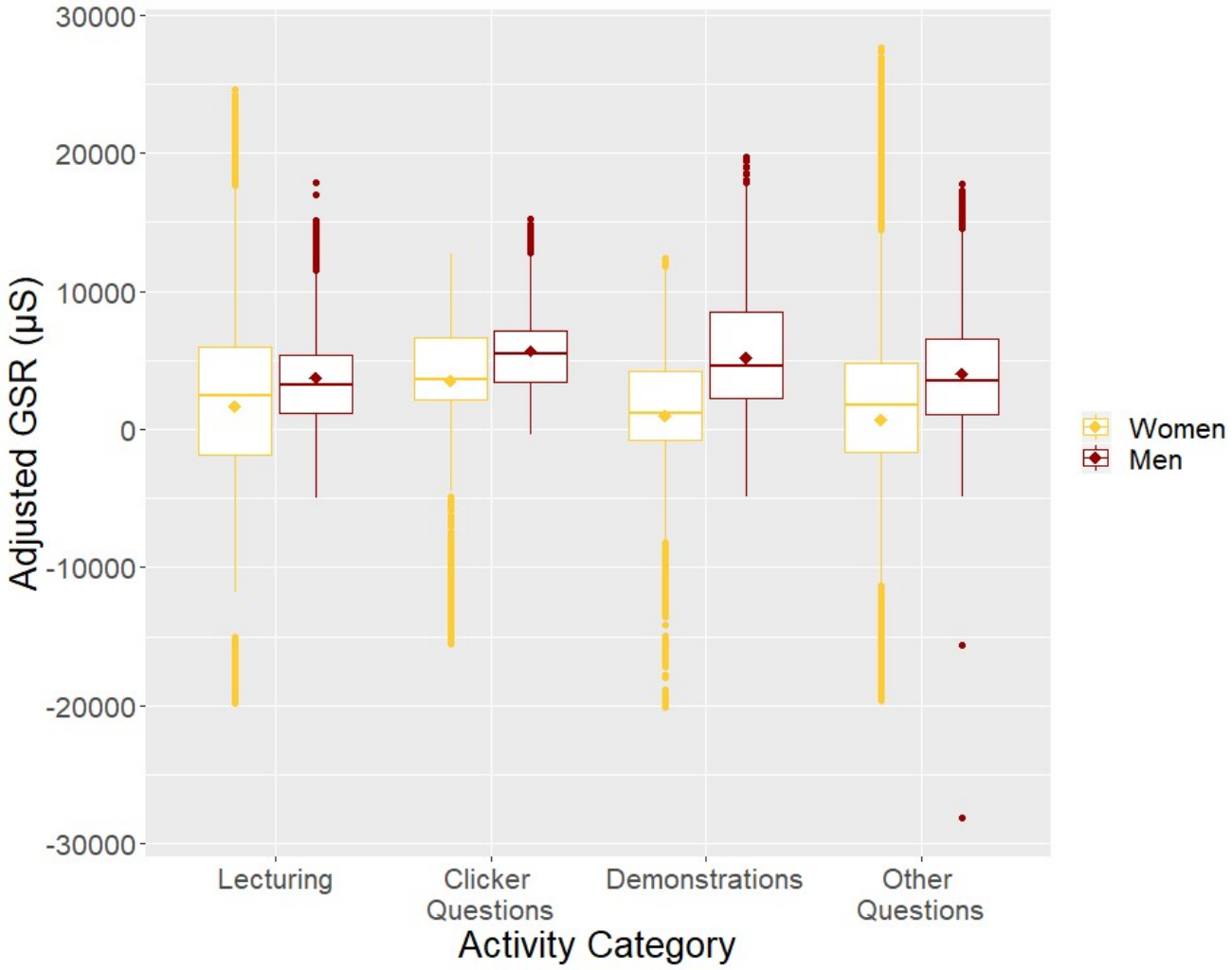

Figure 7. Adjusted GSR by activity and gender. Adjusted GSR was calculated by subtracting each session's pre-class average from raw GSR readings.

Literature Cited

Adekoya, Y. M., Ed, M., & Olatoye, R. A. (2011). Effect of Demonstration, Peer-Tutoring, and Lecture Teaching Strategies on Senior Secondary School Students' Achievement in an Aspect of Agricultural Science. *The Pacific Journal of Science and Technology, 12*(1), 320–332.

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

AlKandari, N. (2012). Students' Communication and Positive Outcomes in College Classrooms. *Education, 133*(1), 19–30.

Anderson-Cook, C. M., & Dorai-Raj, S. (2001). An Active Learning In-Class Demonstration of Good Experimental Design. *Journal of Statistics Education, 9*(1), null. https://doi.org/10.1080/10691898.2001.11910645

Appleton, J. J., Christenson, S. L., Kim, D., & Reschly, A. L. (2006). Measuring cognitive and psychological engagement: Validation of the Student Engagement Instrument. *Journal of School Psychology, 44*(5), 427–445. https://doi.org/10.1016/j.jsp.2006.04.002

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Boucsein, W. (2012). *Electrodermal Activity* (2nd ed.). Springer US. https://doi.org/10.1007/978-1-4614-1126-0

Caldwell, J. E. (2007). Clickers in the Large Classroom: Current Research and Best-Practice Tips. *CBE—Life Sciences Education, 6*(1), 9–20. https://doi.org/10.1187/cbe.06-12-0205

Carini, R. M., Kuh, G. D., & Klein, S. P. (2006). Student Engagement and Student Learning: Testing the Linkages*. *Research in Higher Education, 47*(1), 1–32. https://doi.org/10.1007/s11162-005-8150-9

Crombie, G., Pyke, S. W., Silverthorn, N., Jones, A., & Piccinin, S. (2003). Students' Perceptions of Their Classroom Participation and Instructor as a Function of Gender and Context. *The Journal of Higher Education, 74*(1), 51–76. https://doi.org/10.1080/00221546.2003.11777187

Dragon, T., Arroyo, I., Woolf, B. P., Burleson, W., el Kaliouby, R., & Eydgahi, H. (2008). Viewing Student Affect and Learning through Classroom Observation and Physical Sensors. In B. P. Woolf, E. Aïmeur, R. Nkambou, & S. Lajoie (Eds.), *Intelligent Tutoring Systems* (pp. 29–39). Springer. https://doi.org/10.1007/978-3-540-69132-7_8

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School Engagement: Potential of the Concept, State of the Evidence. *Review of Educational Research, 74*(1), 59–109. https://doi.org/10.3102/00346543074001059

Goldberg, N. A., & Ingram, K. W. (2011). Improving Student Engagement in a Lower-Division Botany Course. *Journal of the Scholarship of Teaching and Learning, 11*(2), 76–90.

Hardy, M., Wiebe, E. N., Grafsgaard, J. F., Boyer, K. E., & Lester, J. C. (2013). Physiological Responses to Events during Training: Use of Skin Conductance to Inform Future Adaptive Learning Systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 57*(1), 2101–2105. https://doi.org/10.1177/1541931213571468

Hothorn, T., Bretz, F., Ag, N. P., & Westfall, P. (2015). Simultaneous Inference in General Parametric Models. *Biometrical Journal, 50*(3), 346–363.

Hu, H., Real, E., Takamiya, K., Kang, M.-G., Ledoux, J., Huganir, R. L., & Malinow, R. (2007). Emotion Enhances Learning via Norepinephrine Regulation of AMPA-Receptor Trafficking. *Cell, 131*(1), 160–173. https://doi.org/10.1016/j.cell.2007.09.017

Irwin, A., & Wynne, B. (2003*). Misunderstanding Science?: The Public Reconstruction of Science and Technology*. Cambridge University Press.

Kang, H., Lundeberg, M., Wolter, B., delMas, R., & Herreid, C. F. (2012). Gender differences in student performance in large lecture classrooms using personal response systems ('clickers') with narrative case studies. *Learning, Media and Technology, 37*(1), 53–76. https://doi.org/10.1080/17439884.2011.556123

King, D. B., & Joshi, S. (2008). Gender Differences in the Use and Effectiveness of Personal Response Devices. *Journal of Science Education and Technology, 17*(6), 544–552. https://doi.org/10.1007/s10956-008-9121-7

Knight, J. K., Wise, S. B., & Southard, K. M. (2013). Understanding Clicker Discussions: Student Reasoning and the Impact of Instructional Cues. *CBE—Life Sciences Education, 12*(4), 645–654. https://doi.org/10.1187/cbe.13-05-0090

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82*(13). https://doi.org/10.18637/jss.v082.i13

Malmivuo, J., Malmivuo, P. of B. and H. of the R. G. I. J., Plonsey, R., & Plonsey, P. of B. E. R. (1995). *Bioelectromagnetism: Principles and Applications of Bioelectric and Biomagnetic Fields*. Oxford University Press.

Marks, H. M. (2000). Student Engagement in Instructional Activity: Patterns in the Elementary, Middle, and High School Years. *American Educational Research Journal, 37*(1), 153–184. https://doi.org/10.3102/00028312037001153

Mazzolini, A. P., Daniel, S., & Edwards, T. (2012). Using Interactive Lecture Demonstrations to Improve Conceptual Understanding of Resonance in an Electronics Course. Australasian *Journal of Engineering Education, 18*(1), 69–88. https://doi.org/10.7158/22054952.2012.11464066

McComas, W. F., & Abraham, L. (2004). ASKING MORE EFFECIVE QUESTIONS. *Rossier School of Education,* 1–16.

McDonald, J. H. (2014). *Handbook of Biological Statistics* (3rd ed.). Sparky House Publishing.

McNeal, K. S., Spry, J. M., Mitra, R., & Tipton, J. L. (2014). Measuring Student Engagement, Knowledge, and Perceptions of Climate Change in an Introductory Environmental Geology Course. *Journal of Geoscience Education, 62*(4), 655–667. https://doi.org/10.5408/13-111.1

McNeal, K. S., Zhong, M., Soltis, N., Doukopoulos, L., Johnson, E., Courtney, S., Alwan, A., & Porch, M. (2020). *Measuring undergraduate student engagement in traditional and active learning biology classrooms using skin biosensors*. Manuscript in Review.

Medicine, I. of, Council, N. R., Education, D. of B. and S. S. and, Families, B. on C., Youth, and, & Learn, C. on I. H. S. S. E. and M. to. (2003). *Engaging Schools: Fostering High School Students' Motivation to Learn*. National Academies Press.

Morgan, J., Barroso, L. R., & Simpson, N. (2007). Active demonstrations for enhancing learning. *2007 37th Annual Frontiers In Education Conference - Global Engineering: Knowledge*

Without Borders, Opportunities Without Passports, S2A-1-S2A-5. https://doi.org/10.1109/FIE.2007.4418057

Niemeyer, E. D., & Zewail-Foote, M. (2018). Investigating the Influence of Gender on Student Perceptions of the Clicker in a Small Undergraduate General Chemistry Course. *Journal of Chemical Education, 95*(2), 218–223. https://doi.org/10.1021/acs.jchemed.7b00389

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic Emotions in Students' Self-Regulated Learning and Achievement: A Program of Qualitative and Quantitative Research. *Educational Psychologist, 37*(2), 91–105. https://doi.org/10.1207/S15326985EP3702_4

Peterson, P. L., Swing, S. R., Stark, K. D., & Waas, G. A. (1984). Students' Cognitions and Time on Task During Mathematics Instruction. *American Educational Research Journal, 21*(3), 487–515. https://doi.org/10.3102/00028312021003487

Pintrich, P. R., Marx, R. W., & Boyle, R. A. (1993). Beyond Cold Conceptual Change: The Role of Motivational Beliefs and Classroom Contextual Factors in the Process of Conceptual Change. *Review of Educational Research, 63*(2), 167–199. https://doi.org/10.3102/00346543063002167

Poh, M.-Z., Swenson, N. C., & Picard, R. W. (2010). A Wearable Sensor for Unobtrusive, Long-Term Assessment of Electrodermal Activity. *IEEE Transactions on Biomedical Engineering, 57*(5), 1243–1252. https://doi.org/10.1109/TBME.2009.2038487

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

RStudio Team. (2019). *RStudio: Integrated Development Enviromnent for R*. RStudio, Inc.

Schwarz, C. V., Passmore, C., & Reiser, B. J. (2017). *Helping Students Make Sense of the World Using Next Generation Science and Engineering Practices*. NSTA Press.

Singer, H. (1978). Active Comprehension: From Answering to Asking Questions. *The Reading Teacher, 31*(8), 901–908. JSTOR.

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

Smith, Michelle K., Jones, F. H. M., Gilbert, S. L., & Wieman, C. E. (2013). The Classroom Observation Protocol for Undergraduate STEM (COPUS): A New Instrument to Characterize University STEM Classroom Practices. *CBE—Life Sciences Education, 12*(4), 618–627. https://doi.org/10.1187/cbe.13-08-0154

Sokoloff, D. R., & Thornton, R. K. (1997). Using interactive lecture demonstrations to create an active learning environment. *The Physics Teacher, 35*(340), 9.

Stains, M., Harshman, J., Barker, M. K., Chasteen, S. V., Cole, R., DeChenne-Peters, S. E., Eagan, M. K., Esson, J. M., Knight, J. K., Laski, F. A., Levis-Fitzgerald, M., Lee, C. J., Lo, S. M., McDonnell, L. M., McKay, T. A., Michelotti, N., Musgrove, A., Palmer, M. S., Plank, K. M., … Young, A. M. (2018). Anatomy of STEM teaching in North American universities. *Science, 359*(6383), 1468–1470. https://doi.org/10.1126/science.aap8892

Uhumuavbi, P. O., & Mamudu, J. A. (2009). Relative Effects of Programmed Instruction and Demonstration Methods on Students' Academic Performance in Science. *College Student Journal, 43*(2), 658–668.

Weber, E. U. (2010). What shapes perceptions of climate change? *WIREs Climate Change, 1*(3), 332–342. https://doi.org/10.1002/wcc.41

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Witkowski, P., & Cornell, T. (2015). An Investigation into Student Engagement in Higher Education Classrooms. InSight: *A Journal of Scholarly Teaching, 10*, 56–67.

Wolter, H. K., Lundeberg, M. A., Kang, H., & Herreid, C. F. (2011). Students' perceptions of using personal response systems ("clickers") with cases in science. *Journal of College Science Teaching 40*(4), 14–19.

Appendix 1. Qualitative descriptions of pedagogies

Author	Year	Title	Description
Active Learning			
Williams, Adrienne E.; Aguilar-Roca, Nancy M.; O'Dowd, Diane K.	2016	“Lecture Capture Podcasts: Differential Student Use and Performance In a Large Introductory Course”	Paper: "conducted in two sections of a large active-learning undergraduate introductory biology class". Qualtrics survey: "Active learning lectures with clickers, demonstrations and small group work. Roughly 15-25 minutes per 50 minutes was active."
Cotner, Sehoya	N/A	“RCN Dataset”	Two 90-minute classes per week, 60% active learning and 40% lecture on average. Course average determined by: exams-70%, assignments-10%, quizzes-20%. Exams were multiple choice, fill-in-blank, essay, and short answer.
Cotner, Sehoya	N/A	“RCN Dataset”	A 110 minute lecture twice a week, and a 170 minute lab once a week. Traditional textbook (not OER). The lecture was front loaded with extensive active learning.
Cotner, Sehoya	N/A	“RCN Dataset”	110 min lec twice a week, 170 min lab once a week. Tratitional text book (not OER). Lectures were mostly front loaded with extensive active learning.
Cotner, Sehoya	N/A	“RCN Dataset”	An Active-learning class, for 75 mins, two times per week, and a 2-hour lab each week.
Cotner, Sehoya	N/A	“RCN Dataset”	An Active-learning class, for 75 mins, two times per week, and a 2-hour lab each week.
Cotner, Sehoya	N/A	“RCN Dataset”	An Active-learning class, for 75 mins, two times per week, and a 2-hour lab each week.
Cotner, Sehoya	N/A	“RCN Dataset”	An Active-learning class, for 75 mins, two times per week, and a 2-hour lab each week.
Cotner, Sehoya	N/A	“RCN Dataset”	An Active-learning class, for 75 mins, two times per week, and a 2-hour lab each week.
Cotner, Sehoya	N/A	“RCN Dataset”	An Active-learning class, for 75 mins, two times per week, and a 2-hour lab each week.
Cotner, Sehoya	N/A	“RCN Dataset”	An Active-learning class, for 75 mins, two times per week, and a 2-hour lab each week.
Cotner, Sehoya	N/A	“RCN Dataset”	An Active-learning class, for 75 mins, two times per week, and a 2-hour lab each week.
Cotner, Sehoya	N/A	“RCN Dataset”	An Active-learning class, for 75 mins, two times per week, and a 2-hour lab each week.

Cotner, Sehoya	N/A	“RCN Dataset”	An Active-learning class, for 75 mins, two times per week, and a 2-hour lab each week.
Cotner, Sehoya	N/A	“RCN Dataset”	An Active-learning class, for 75 mins, two times per week, and a 2-hour lab each week.
Cotner, Sehoya	N/A	“RCN Dataset”	An Active-learning class, for 75 mins, two times per week, and a 2-hour lab each week.
Cotner, Sehoya	N/A	“RCN Dataset”	An Active-learning class, for 75 mins, two times per week, and a 2-hour lab each week.
Cotner, Sehoya	N/A	“RCN Dataset”	An Active-learning class, for 75 mins, two times per week, and a 2-hour lab each week.
Cotner, Sehoya	N/A	“RCN Dataset”	An Active-learning class, for 75 mins, two times per week, and a 2-hour lab each week.
Cotner, Sehoya	N/A	“RCN Dataset”	An Active-learning class, for 75 mins, two times per week, and a 2-hour lab each week.
Cotner, Sehoya	N/A	“RCN Dataset”	A monday mini-lecture with iClicker questions and writing prompts, Wednesday group activity with HOCS and iClicker questions, Friday student-led mini-review, formative assessment quiz.
Rodriguez, Marisela; Mundy, Marie-Anne; Kupczynski, Lori; Challoo, Linda	2018	“Effects of Teaching Strategies on Student Success, Persistence, and Perceptions of Course Evaluations”	Project-based learning allows students to work on complex problems and provides authentic experiences in order for students to find purposeful meaning to STEM concepts. Capraro (2013) defines project based learning as a teaching strategy that requires students to think critically and analytically, enhancing their higher-order thinking skills. Project-based learning involves students seeking a solution to complex problems situated within larger projects and justifying their results. Railsback (2002) also stated that project-based learning moves away from memorization and provides complex work that contains interdisciplinary disciplines and encourages cooperative learning. Project-based teaching strategies are a holistic method that is becoming more meaningful to students, especially those who have different learning styles, backgrounds, and abilities in which students are able to explore within the curriculum. [from intro--methods discuss that instructors underwent training related to the types of teaching, and sylabi were approved, but did not give any more specifics.
Rodriguez, Marisela; Mundy, Marie-Anne;	2018	“Effects of Teaching Strategies on Student Success,	Peer-led instruction involves breaking large lectures into smaller workshop sections in which peer instructors facilitate cooperative group work, thus increasing student interaction...Watkins and Mazur

Kupczynski, Lori; Challoo, Linda		Persistence, and Perceptions of Course Evaluations"	(2013) stated that by incorporating and structuring peer discussions, students have more opportunities to get to share ideas and form a collaborative discussion within the introductory science classroom. During these discussions, the instructor is able to listen to the students and students are able to engage more in the lecture, which increases the faculty-student interaction. PI creates a more exciting classroom and a positive environment is seen between faculty member and students. These constant interactions and feedback, as seen in PI courses, allows an instructor to see the weaknesses of the students which will allow him or her to better tailor their instruction according to the students' needs. [from intro--methods discuss that instructors underwent training related to the types of teaching, and syllabi were approved, but did not give any more specifics]
Rodriguez, Marisela; Mundy, Marie-Anne; Kupczynski, Lori; Challoo, Linda	2018	"Effects of Teaching Strategies on Student Success, Persistence, and Perceptions of Course Evaluations"	Project-based learning allows students to work on complex problems and provides authentic experiences in order for students to find purposeful meaning to STEM concepts. Capraro (2013) defines project based learning as a teaching strategy that requires students to think critically and analytically, enhancing their higher-order thinking skills. Project-based learning involves students seeking a solution to complex problems situated within larger projects and justifying their results. Railsback (2002) also stated that project-based learning moves away from memorization and provides complex work that contains interdisciplinary disciplines and encourages cooperative learning. Project-based teaching strategies are a holistic method that is becoming more meaningful to students, especially those who have different learning styles, backgrounds, and abilities in which students are able to explore within the curriculum. [from intro--methods discuss that instructors underwent training related to the types of teaching, and sylabi were approved, but did not give any more specifics]
Rodriguez, Marisela; Mundy, Marie-Anne; Kupczynski,	2018	"Effects of Teaching Strategies on Student Success, Persistence, and Perceptions of	Peerled instruction involves breaking large lectures into smaller workshop sections in which peer instructors facilitate cooperative group work, thus increasing student interaction...Watkins and Mazur (2013) stated that by incorporating and structuring peer discussions, students have more

Lori; Challoo, Linda		Course Evaluations"	opportunities to get to share ideas and form a collaborative discussion within the introductory science classroom. During these discussions, the instructor is able to listen to the students and students are able to engage more in the lecture, which increase the faculty-student interaction. PI creates a more exciting classroom and a positive environment is seen between faculty member and students. These constant interactions and feedback, as seen in PI courses, allows an instructor to see the weaknesses of the students which will allow him or her to better tailor their instruction according to the students' needs. [from intro--methods discuss that instructors underwent training related to the types of teaching, and sylabi were approved, but did not give any more specifics]
Cotner, Sehoya; Ballen, Cissy J.	2017	"Can Mixed Assessment Methods Make Biology Classes More Equitable?"	lecture+interactive elements
Cotner, Sehoya; Ballen, Cissy J.	2017	"Can Mixed Assessment Methods Make Biology Classes More Equitable?"	lecture+interactive elements
Cotner, Sehoya; Ballen, Cissy J.	2017	"Can Mixed Assessment Methods Make Biology Classes More Equitable?"	lecture+interactive elements
Cotner, Sehoya; Ballen, Cissy J.	2017	"Can Mixed Assessment Methods Make Biology Classes More Equitable?"	lecture+interactive elements
Cotner, Sehoya; Ballen, Cissy J.	2017	"Can Mixed Assessment Methods Make Biology Classes	lecture+interactive elements

		More Equitable?”	
Cotner, Sehoya; Ballen, Cissy J.	2017	“Can Mixed Assessment Methods Make Biology Classes More Equitable?”	lecture+interactive elements
Cotner, Sehoya; Ballen, Cissy J.	2017	“Can Mixed Assessment Methods Make Biology Classes More Equitable?”	lecture+interactive elements
Cotner, Sehoya; Ballen, Cissy J.	2017	“Can Mixed Assessment Methods Make Biology Classes More Equitable?”	lecture+interactive elements
Cotner, Sehoya; Ballen, Cissy J.	2017	“Can Mixed Assessment Methods Make Biology Classes More Equitable?”	lecture+interactive elements
Cotner, Sehoya; Ballen, Cissy J.	2017	“Can Mixed Assessment Methods Make Biology Classes More Equitable?”	lecture+interactive elements
Cotner, Sehoya; Ballen, Cissy J.	2017	“Can Mixed Assessment Methods Make Biology Classes More Equitable?”	lecture+interactive elements
Cotner, Sehoya; Ballen, Cissy J.	2017	“Can Mixed Assessment Methods Make Biology Classes More Equitable?”	lecture+interactive elements

Cotner, Sehoya; Ballen, Cissy J.	2017	"Can Mixed Assessment Methods Make Biology Classes More Equitable?"	lecture+interactive elements
Rhodes, Ashley E.	2013	"The Effect of Teacher Designed Multimedia on Student Comprehension and Retention Rates within Introductory College Science Courses"	Offered in a studio format. This format combines lecture and laboratory activities. Students work in groups of four and work through exercises as guided by a laboratory manual and a course website. Every course section is led by two full-time faculty instructors and two graduate teaching assistants; student to instructor ratio is 20:1. Over the course of a semester, seven main topics are covered: introduction to science, ecology, cell biology, genetics, energetics, plant biology, and animal biology. Students' grades are determined by performance on biweekly quizzes and seven unit exams.
Rhodes, Ashley E.	2013	"The Effect of Teacher Designed Multimedia on Student Comprehension and Retention Rates within Introductory College Science Courses"	Offered in a studio format. This format combines lecture and laboratory activities. Students work in groups of four and work through exercises as guided by a laboratory manual and a course website. Every course section is led by two full-time faculty instructors and two graduate teaching assistants; student to instructor ratio is 20:1. Over the course of a semester, seven main topics are covered: introduction to science, ecology, cell biology, genetics, energetics, plant biology, and animal biology. Students' grades are determined by performance on biweekly quizzes and seven unit exams...[+multimedia module] development of the first multimedia module ensued using the principles outlined for multimedia development by Mayer (2001, 2009). Special attention was given to Mayer's (2001, 2009) principles for reducing extraneous processing and managing essential processing...[animation+narration available through course website (looks to be the eqivalent of AU's Canvas)
Rhodes, Ashley E.	2013	"The Effect of Teacher Designed Multimedia on Student	Offered in a studio format. This format combines lecture and laboratory activities. Students work in groups of four and work through exercises as guided by a laboratory manual and a course website. Every course section is led by two full-time faculty

		Comprehension and Retention Rates within Introductory College Science Courses”	instructors and two graduate teaching assistants; student to instructor ratio is 20:1. Over the course of a semester, seven main topics are covered: introduction to science, ecology, cell biology, genetics, energetics, plant biology, and animal biology. Students’ grades are determined by performance on biweekly quizzes and seven unit exams...[+multimedia module] development of the first multimedia module ensued using the principles outlined for multimedia development by Mayer (2001, 2009). Special attention was given to Mayer’s (2001, 2009) principles for reducing extraneous processing and managing essential processing...[animation+narration available through course website (looks to be the eqivalent of AU's Canvas)...[module adjusted based on feedback from the prior semester]
Rhodes, Ashley E.	2013	“The Effect of Teacher Designed Multimedia on Student Comprehension and Retention Rates within Introductory College Science Courses”	Offered in a studio format. This format combines lecture and laboratory activities. Students work in groups of four and work through exercises as guided by a laboratory manual and a course website. Every course section is led by two full-time faculty instructors and two graduate teaching assistants; student to instructor ratio is 20:1. Over the course of a semester, seven main topics are covered: introduction to science, ecology, cell biology, genetics, energetics, plant biology, and animal biology. Students’ grades are determined by performance on biweekly quizzes and seven unit exams...[+multimedia module] development of the first multimedia module ensued using the principles outlined for multimedia development by Mayer (2001, 2009). Special attention was given to Mayer’s (2001, 2009) principles for reducing extraneous processing and managing essential processing...[animation+narration available through course website (looks to be the eqivalent of AU's Canvas)...[module adjusted based on feedback from the prior semester]"
Gross, David; Pietri, Evava S.; Anderson, Gordon; Moyano-Camihort,	2015	"Increased Preclass Preparation Underlies Student Outcome Improvement in	Example problems solved in class by the instructor in the standard course were adapted for peer–peer activities in the flipped-format course...The flipped-format course met either for one 75-min session per week or for two 50-min sessions per week. This

Karin; Graham, Mark J.		the Flipped Classroom"	reduced in-class time was supplemented with prerecorded "lectures" available to the students at least a week before class, which increased the online component in the flipped course compared with the standard course. These supplemental lectures were broken into 5- to 20-min chunks on specific topics in the OWLBook. Students were free to view the supplemental lectures or to skip them, as these lectures carried no course credit. The use of less in-class time allowed the instructor to offer more sections of the course, which allowed the class size to remain about constant despite a rapidly increasing total number of students taking the course. All other components of the standard course were present in the flipped course. Aside from the prerecorded lectures and reduced in-class time for the flipped-format course, a substantial difference between course formats was the increased use of active learning in the flipped classroom. This took the form of peer–peer think–pair–share activities, clicker responses, and example problems for students to work in the once-weekly 75-min sections. In the twice-weekly 50-min sessions, team-based learning (Michaelsen et al., 2004) was used. In this format, teams of five to eight students remained allied throughout the semester. In-class activities included difficult example problems attacked by teams, individual and team readiness assessments on new material, and student explanations of problem solutions on projected whiteboards.
Ballen, Cissy	N/A	"AU dataset"	Active

Ballen, Cissy	N/A	"AU dataset"	Active
Ballen, Cissy	N/A	"AU dataset"	Active
Ballen, Cissy	N/A	"AU dataset"	Active
Ballen, Cissy	N/A	"AU dataset"	Active
Ballen, Cissy	N/A	"AU dataset"	Active-learning. Students were assinged into groups for each of the four different exams. Each class period the students completed a worksheet based on the days lessons. The worksheets were collected at the end of each period. A third were graded, a third were given particpation points, and a third I selected 1 paper from the groups to grade. The student gruops were reshuffled for each exam section. The exams were a third short answer and remaining multiple choice. The final was 100% multiple choice
Lax, Neil; Morris, James; Kolber, Benedict J.	2017	"A Partial Flip Classroom Exercise in a Large Introductory General Biology Course Increases Performance at Multiple Levels"	Students in the partial flip group watched a 16-min lecture that was recorded by the instructor demonstrating the Meselson and Stahl experiment. Following the recording, students answered the same two multiple-choice questions as the control group along with a ‘password’ question, the answer to which was embedded in the recorded lecture. The password question was used to ensure that students actually watched the recorded lecture prior to class...For the partial flip group, students worked on a group worksheet for 20 min. This worksheet challenged students to recreate the experiment and determine conclusions based on hypothetical situations and results. During the worksheet component of the lesson, four individuals (two course instructors plus two graduate-level teaching assistants) walked around the room to aid students. To facilitate discussion between staff and students, four rows in the auditorium-style classroom were left empty to act as aisles for the staff. At the end of the worksheet, the students were given three PRS questions. Students without remotes were asked to provide answers on 3 × 5 cards....Following class, for both groups, students completed an online assignment that covered the Meselson and Stahl experiment...The partial flip group completed the two ‘novel’ questions only...in the partial flip section, both the students and professor spent

			approximately 75% of class time in active learning/teaching (Figure 1) (see page 6 for copus pie)
Interactive			
Stanich, Cynthia; Pelch, Michael A.; Theobald, Elli J.; Freeman, Scott	2018	"A New Approach To Supplementary Instruction Narrows Achievement and Affect Gaps For Underrepresented Minorities, First-Generation Students, and Women"	Active Learning
Cotner, Sehoya	N/A	"RCN Dataset"	Three 60-minute classes per week, approximately 60% lecture and 40% in-class case studies, activities, & discussions. Course average determined by: 56% exams, 21% quizzes, and 23% homework. Exams were about 2/3 multiple choice and 1/3 written-response/short answer.
Cotner, Sehoya	N/A	"RCN Dataset"	A 110 minute lecture twice a week, and a 170 minute lab once a week. Traditional textbook (not OER). Lectures were mostly traditional with limited active learning.
Cotner, Sehoya	N/A	"RCN Dataset"	A 110 minute lecture twice a week, and a 170 min lab once a week. Traditional textbook (not OER). Lectures were mostly traditional with limited active learning incorporated.
Cotner, Sehoya	N/A	"RCN Dataset"	A 110 minute lecture twice a week, and a 170 min lab once a week. Traditional textbook (not OER). Lectures were mostly traditional with limited active learning incorporated.
Cotner, Sehoya	N/A	"RCN Dataset"	A 110 minute lecture twice a week, and a 170 min lab once a week. Traditional textbook (not OER). Lectures were mostly traditional with limited active learning incorporated.
Cotner, Sehoya	N/A	"RCN Dataset"	A 110 minute lecture twice a week, and a 170 minute lab once a week. Traditional textbook (not OER). Lectures were mostly traditional with active learning incorporated.
Cotner, Sehoya	N/A	"RCN Dataset"	A 80 minute lecture twice a week and a 170 min lab once a week. Traditional textbook (not OER).

			Lectures were mostly traditional with active learning incorporated.
Cotner, Sehoya	N/A	“RCN Dataset”	A 110 minute lecture twice a week and a 170 min lab once a week. Traditional textbook (not OER). Lectures were mostly traditional with active learning incorporated.
Cotner, Sehoya	N/A	“RCN Dataset”	A 70 minute lecture twice a week and a 170 min lab once a week. Traditional textbook (not OER). Lectures were mostly traditional with active learning incorporated.
Cotner, Sehoya	N/A	“RCN Dataset”	A 70 minute lecture twice a week and a 170 min lab once a week. Traditional textbook (not OER). Lectures were mostly traditional with active learning incorporated.
Cotner, Sehoya	N/A	“RCN Dataset”	A 70 minute lecture twice a week and a 170 min lab once a week. Traditional textbook (not OER). Lectures were mostly traditional with active learning incorporated.
Cotner, Sehoya	N/A	“RCN Dataset”	A 70 minute lecture twice a week and a 170 min lab once a week. Traditional textbook (not OER). Lectures were mostly traditional with active learning incorporated.
Cotner, Sehoya	N/A	“RCN Dataset”	A 70 minute lecture twice a week and a 170 min lab once a week. Traditional textbook (not OER). Lectures were mostly traditional with active learning incorporated.
Cotner, Sehoya	N/A	“RCN Dataset”	A 70 minute lecture twice a week and a 170 min lab once a week. Traditional textbook (not OER). Lectures were mostly traditional with active learning incorporated.
Cotner, Sehoya	N/A	“RCN Dataset”	A 70 minute lecture twice a week and a 170 min lab once a week. Traditional textbook (not OER). Lectures were mostly traditional with active learning incorporated.
Cotner, Sehoya	N/A	“RCN Dataset”	A 70 minute lecture twice a week and a 170 min lab once a week. Traditional textbook (not OER). Lectures were mostly traditional with active learning incorporated.
Ballen, Cissy	N/A	AU dataset	Active
Bolt, Brian Grady	2009	"Measuring the Impact of Varied Instructional Approaches in an	Ten to fifteen minutes of class time were classified as conforming exclusively to one of three types of material delivery. The three classifications were labeled as either traditional lecture,; technology-

		Introductory Animal Science Course"	enhanced, or; web-enhanced...Although several teaching styles (and combinations of styles) are routinely employed, the focus of this study was to identify 10-15 periods of lecture time that had a clearly describable type of teaching, either traditional lecture (TL), technology-enhanced lecture (TE) and problem based, Web-enhanced learning, (WEB). Traditional lecture was defined as only the teacher coupled with a whiteboard, willing to interact with the students and respond to questions. Technology-enhanced was defined as the teacher coupled with projector, slides and various forms of multimedia, typically projected onto one of two large screens in the front of the lecture hall. In technology enhanced the instructor was willing to interact with students and respond to questions. Web-enhanced was defined as students presented with a problem and using Web resources to find solutions. The instructor was willing to interact with the class during the Web enhanced sessions but students were encouraged to search out solutions and answers on their own. It is important to note that during all types of instruction the teacher would respond to questions and interact with the class...[one class, so presumably alternated] Students were required to purchase I-clickers... Students were taught material throughout the course of the semester and after 10-15 minutes of instructional time knowledge questions were posed to the class to assess the relative level of understanding. Students responded with the ARS (IClickers) a system that allows for anonymous submission of answers, used for several purposes but most notably as a teaching tool to increase student engagement.
Lax, Neil; Morris, James; Kolber, Benedict J.	2017	"A Partial Flip Classroom Exercise in a Large Introductory General Biology Course Increases	The control group received the same content from the recorded lecture in a live lecture format from the course instructor (lecture time 20 min). At the end of the lecture, the students were given two questions, answered using personal response system (PRS) remotes (Turning Point Technologies). Students without remotes were asked to provide answers on 3

		Performance at Multiple Levels"	× 5 cards. Students received participation credit for answering the questions...Following class, for both groups, students completed an online assignment that covered the Meselson and Stahl experiment. The control group completed identical questions to the partial flip group's in-class worksheet plus an additional two 'novel' questions... Using COPUS, we found that both the instructor and students in the control section spent only about 50% of class time involved in active learning/teaching activities (Student: Ind, AnQ, SQ, Prd, OG, CG, WG; Instructor: DV, AnQ, FUp, MG, 1o1, CQ, PQ). (see page 6 for COPUS pie).
Lecture-based			
Cotner, Sehoya	N/A	"RCN Dataset"	Two 70 minute lectures, One 50 minute discussion section.
Cotner, Sehoya	N/A	"RCN Dataset"	Two 50 minute lectures, One 75 minute peer lead workshop.
Cotner, Sehoya	N/A	"RCN Dataset"	Two 50 minute lectures, One 75 minute peer lead workshop.
Cotner, Sehoya	N/A	"RCN Dataset"	Two 50 minute lectures, One 75 minute peer lead workshop.
Cotner, Sehoya	N/A	"RCN Dataset"	Two 75-minute lectures per week.
Cotner, Sehoya	N/A	"RCN Dataset"	Two 75-minute lectures per week.
Cotner, Sehoya	N/A	"RCN Dataset"	Two 50 minute lectures, One 75 minute peer lead workshop.
Cotner, Sehoya	N/A	"RCN Dataset"	Two 50 minute lectures, One 75 minute peer lead workshop.
Cotner, Sehoya	N/A	"RCN Dataset"	Two 50 minute lectures, One 75 minute peer lead workshop.
Cotner, Sehoya	N/A	"RCN Dataset"	Two 75-minute lectures per week.
Cotner, Sehoya	N/A	"RCN Dataset"	Three 60-minute classes per week, primarily lecture-based. Course average determined by: 75% exams, 20% quizzes/homework, and 5% participation. Exams were mostly multiple choice with some fill-in-blank, essay, and short answer questions.
Cotner, Sehoya	N/A	"RCN Dataset"	A 110 minute lecture twice a week, and a 170 minute lab once a week. Traditional textbook (not OER). Lectures were mostly traditional.
Cotner, Sehoya	N/A	"RCN Dataset"	A 110 minute lecture twice a week, and a 170 min lab once a week. Traditional textbook (not OER). Lectures were mostly traditional.

Cotner, Sehoya	N/A	“RCN Dataset”	A 110 minute lecture twice a week, and a 170 min lab once a week. Traditional textbook (not OER). Lectures were mostly traditional.
Cotner, Sehoya	N/A	“RCN Dataset”	A 70 minute lecture and lab twice a week. Traditional textbook (not OER). Lectures were mostly traditional.
Cotner, Sehoya	N/A	“RCN Dataset”	A 70 minute lecture and lab twice a week. Traditional textbook (not OER). Lectures were mostly traditional.
Cotner, Sehoya	N/A	“RCN Dataset”	A 70 minute lecture and lab twice a week. Traditional textbook (not OER). Lectures were mostly traditional.
Cotner, Sehoya	N/A	“RCN Dataset”	A 110 minute lecture twice a week, and a 170 min lab once a week. Traditional textbook (not OER). Lectures were mostly traditional.
Cotner, Sehoya	N/A	“RCN Dataset”	A 110 minute lecture twice a week, and a 170 min lab once a week. Traditional textbook (not OER). Lectures were mostly traditional.
Cotner, Sehoya	N/A	“RCN Dataset”	A 45 min lecture three times a week (Team Taught).
Rodriguez, Marisela; Mundy, Marie-Anne; Kupczynski, Lori; Challoo, Linda	2018	“Effects of Teaching Strategies on Student Success, Persistence, and Perceptions of Course Evaluations”	Instructor was introduced to each of the teaching methods; lecture-based, project-based, and peer-led instruction. In this session, the instructors received a formal understanding of the three teaching strategies. On the second session, discussions of the understanding of each of the teaching strategies took place and a checklist was created (by the biology instructor, chemistry instructor, and investigator) which identified the characteristics of each of the teaching strategies. This checklist helped guide the design and activities of each of thesections that the instructors were expected to teach. On the third session, the checklist that was created from the second session was used to drive the design of the activities and assessments for each of the teaching strategies. In this session, the instructors applied the checklist to the activities that they intended to implement throughout the three teaching strategies. During the fourth session, the activities that were discussed throughout the third session were applied to the syllabus of the instructors’ course. Each instructor was expected to submit three different syllabi that complemented the three teaching strategies. A

			final draft of the syllabus needed to be discussed and approved by the biology instructor, chemistry instructor, and investigator one week before classes resumed. In this way, before the fall 2015 semester began, the instructors for BIOL 1306 and CHEM 1311 understood the concepts of the three teachings strategies and also understood to only implement the teaching strategy that was assigned for the class. [no further details given as to what was on the syllabi]
Rodriguez, Marisela; Mundy, Marie-Anne; Kupczynski, Lori; Challoo, Linda	2018	"Effects of Teaching Strategies on Student Success, Persistence, and Perceptions of Course Evaluations"	Instructor was introduced to each of the teaching methods; lecture-based, project-based, and peer-led instruction. In this session, the instructors received a formal understanding of the three teaching strategies. On the second session, discussions of the understanding of each of the teaching strategies took place and a checklist was created (by the biology instructor, chemistry instructor, and investigator) which identified the characteristics of each of the teaching strategies. This checklist helped guide the design and activities of each of thesections that the instructors were expected to teach. On the third session, the checklist that was created from the second session was used to drive the design of the activities and assessments for each of the teaching strategies. In this session, the instructors applied the checklist to the activities that they intended to implement throughout the three teaching strategies. During the fourth session, the activities that were discussed throughout the third session were applied to the syllabus of the instructors' course. Each instructor was expected to submit three different syllabi that complemented the three teaching strategies. A final draft of the syllabus needed to be discussed and approved by the biology instructor, chemistry instructor, and investigator one week before classes resumed. In this way, before the fall 2015 semester began, the instructors for BIOL 1306 and CHEM 1311 understood the concepts of the three teachings strategies and also understood to only implement the teaching strategy that was assigned for the class. [no further details given as to what was on the syllabi]

Cotner, Sehoya; Ballen, Cissy J.	2017	“Can Mixed Assessment Methods Make Biology Classes More Equitable?”	lecture
Cotner, Sehoya; Ballen, Cissy J.	2017	“Can Mixed Assessment Methods Make Biology Classes More Equitable?”	lecture
Ballen, Cissy	N/A	"AU dataset"	Traditional lecture with 1 or 2 think pair share activities (1/3 graded, 2/3 participation credit). Students had access to "Supplementary Instructor" (SI) an undergrad that has taken the course. The SI provided an extra hour of instruction. There were 4 exams 75% multiple choice and given the choice of 1 of 3 short answer questions per exam. Final was 100% multiple choice. There were no SI instrucors available. I did use undergraduate lecture assistants (n = 6) in class to aid small group discsusions and answer student questions.
Ballen, Cissy	N/A	"AU dataset"	Traditional
Rauschenberger, Matthew M.; Sweeder, Ryan D.	2010	"Gender Performance Differences in Biochemistry"	The instructional model for these courses consists of a single large lecture section (e.g. n ¼ 466 students in fall 2008 for Biochem I), with prerequisites...The course does not contain many of the previously mentioned curricular changes so can provide a good reference for ‘‘traditional’’ instruction. [intro made mention of active learning approaches]... In recent years, some sections of these courses used in-class response pads, or clickers, in the hope of enhancing class participation and learning
Rauschenberger, Matthew M.; Sweeder, Ryan D.	2010	"Gender Performance Differences in Biochemistry"	The instructional model for these courses consists of a single large lecture section (e.g. n ¼ 466 students in fall 2008 for Biochem I), with prerequisites...The course does not contain many of the previously mentioned curricular changes so can provide a good reference for ‘‘traditional’’ instruction. [intro made mention of active learning approaches]... In recent years, some sections of these courses used in-class response pads,

			or clickers, in the hope of enhancing class participation and learning
Shibley, Ivan A., Jr.; Milakofsky, Louis; Bender, David S.; Patterson, Henry O.	2003	"College Chemistry and Piaget: An Analysis of Gender Difference, Cognitive Abilities, and Achievement Measures Seventeen Years Apart"	The course consisted of a two-period lecture and two-period lab each week for a fifteen-week semester.
Cotner, Sehoya	N/A	“RCN Dataset”	Hybrid course: One 1-hour lecture, one 2-hour lab, online work
Cotner, Sehoya	N/A	“RCN Dataset”	Hybrid course: One 1-hour lecture, one 2-hour lab, online work
Cotner, Sehoya	N/A	“RCN Dataset”	Hybrid course: One 1-hour lecture, one 2-hour lab, online work
Cotner, Sehoya	N/A	“RCN Dataset”	Hybrid course: One 1-hour lecture, one 2-hour lab, online work
Cotner, Sehoya	N/A	“RCN Dataset”	Two 75-min lectures per week
Cotner, Sehoya	N/A	“RCN Dataset”	Two 75-min lectures per week
Cotner, Sehoya	N/A	“RCN Dataset”	Two 75-min lectures per week

Appendix 2. Published studies included in meta-analysis

Bardi, M., Koone, T., Mewaldt, S., & O'Connor, K. (2011). Behavioral and physiological correlates of stress related to examination performance in college chemistry students. *Stress*, *14*(5), 557–566. https://doi.org/10.3109/10253890.2011.571322

Boli, J., Allen, M. L., & Payne, A. (1985). High-Ability Women and Men in Undergraduate Mathematics and Chemistry Courses. *American Educational Research Journal*, *22*(4), 605–626. https://doi.org/10.3102/00028312022004605

Bolt, B. (2009). MEASURING THE IMPACT OF VARIED INSTRUCTIONAL APPROACHES IN AN INTRODUCTORY ANIMAL SCIENCE COURSE. *All Dissertations*, *483*.

Cotner, S., & Ballen, C. J. (2017). Can mixed assessment methods make biology classes more equitable? *PLOS ONE*, *12*(12), e0189610. https://doi.org/10.1371/journal.pone.0189610

Cromley, J. G., Perez, T., Wills, T. W., Tanaka, J. C., Horvat, E. M., & Agbenyega, E. T.-B. (2013). Changes in race and sex stereotype threat among diverse STEM students: Relation to grades and retention in the majors. *Contemporary Educational Psychology*, *38*(3), 247–258. https://doi.org/10.1016/j.cedpsych.2013.04.003

Eddy, S. L., Brownell, S. E., & Wenderoth, M. P. (2014). Gender Gaps in Achievement and Participation in Multiple Introductory Biology Classrooms. *CBE—Life Sciences Education*, *13*(3), 478–492. https://doi.org/10.1187/cbe.13-10-0204

Gross, D., Pietri, E. S., Anderson, G., Moyano-Camihort, K., & Graham, M. J. (2015). Increased Preclass Preparation Underlies Student Outcome Improvement in the Flipped Classroom. *CBE Life Sciences Education*, *14*(4), ar36. https://doi.org/10.1187/cbe.15-02-0040

Lax, N., Morris, J., & Kolber, B. J. (2017). A partial flip classroom exercise in a large introductory general biology course increases performance at multiple levels. *Journal of Biological Education*, *51*(4), 412–426. https://doi.org/10.1080/00219266.2016.1257503

Niemeyer, E. D., & Zewail-Foote, M. (2018). Investigating the Influence of Gender on Student Perceptions of the Clicker in a Small Undergraduate General Chemistry Course. *Journal of Chemical Education*, *95*(2), 218–223. https://doi.org/10.1021/acs.jchemed.7b00389

Peterfreund, A. R., Rath, K. A., Xenos, S. P., & Bayliss, F. (2008). The Impact of Supplemental Instruction on Students in Stem Courses: Results from San Francisco State University. *Journal of College Student Retention: Research, Theory & Practice*, *9*(4), 487–503. https://doi.org/10.2190/CS.9.4.e

Rauschenberger, M. M., & Sweeder, R. D. (2010). Gender performance differences in biochemistry. *Biochemistry and Molecular Biology Education*, *38*(6), 380–384. https://doi.org/10.1002/bmb.20448

Rhodes, A. E. (2013). *The effect of teacher designed multimedia on student comprehension and retention rates within introductory college science courses*. https://krex.k-state.edu/dspace/handle/2097/15513

Rodriguez, M., Mundy, M.-A., Kupczynski, L., & Challoo, L. (2018). Effects of Teaching Strategies on Student Success, Persistence, and Perceptions of Course Evaluations. *Research in Higher Education Journal*, *35*. https://eric.ed.gov/?id=EJ1194444

Shibley, I. A., Milakofsky, L. M., Bender, D. S., & Patterson, H. O. (2003). College Chemistry and Piaget: An Analysis of Gender Difference, Cognitive Abilities, and Achievement Measures Seventeen Years Apart. *Journal of Chemical Education*, *80*(5), 569. https://doi.org/10.1021/ed080p569

Stanich, C. A., A. Pelch, M., J. Theobald, E., & Freeman, S. (2018). A new approach to supplementary instruction narrows achievement and affect gaps for underrepresented minorities, first-generation students, and women. *Chemistry Education Research and Practice*, *19*(3), 846–866. https://doi.org/10.1039/C8RP00044A

Sunny, C. E., Taasoobshirazi, G., Clark, L., & Marchand, G. (2017). Stereotype threat and gender differences in chemistry. *Instructional Science*, *45*(2), 157–175. https://doi.org/10.1007/s11251-016-9395-8

Williams, A. E., Aguilar-Roca, N. M., & O'Dowd, D. K. (2016). Lecture capture podcasts: Differential student use and performance in a large introductory course. *Educational Technology Research and Development*, *64*(1), 1–12. https://doi.org/10.1007/s11423-015-9406-5

Willoughby, S. D., & Metz, A. (2009). Exploring gender differences with different gain calculations in astronomy and biology. *American Journal of Physics*, *77*(7), 651–657. https://doi.org/10.1119/1.3133087

Appendix 3. R code for Chapter 1

```
# #startup ----------------------------------------------------------------
#upload packages
library(metafor) #used for meta-analyses
library(tidyverse)
setwd("C:/Users/sarao/OneDrive/Desktop/Meta-analysis") #set working directory
# ##Prepare data -----------------------------------------------------------
metadata <- read.csv("C:/Users/sarao/OneDrive/Desktop/Meta-analysis/6_11_data_table.csv",
stringsAsFactors=FALSE) #dataset 1
z_hg <- read.csv("C:/Users/sarao/OneDrive/Desktop/Meta-
analysis/5_30_hedges_data_table.csv", stringsAsFactors=FALSE) #already has hedge's g
calculated within excel file
#create class label so that classes with multiple assessments can be matched later
metadata$label = 1:(length(metadata$author)) #add a dummy variable so that I can match up
classes later
z_hg$label = ((length(metadata$author)+1):(length(metadata$author)+length(z_hg$author)))
#continue number list
#sort into different assessment types, then calculate hedge's g
#hedge's g ("SMD"=Standardized mean difference, specified as hedge's g in escalc manual)
#make 1=female, 2=male, then +=females performed better, - means males performed better
#label assessment type
z_hg$assessment = "exams" #all in z_hg were based on exam scores, so adding label here;
Hedges g already calculated in data
z_hg$imputed_SD = 0 #none of this group had an imputed SD, so noting that here
#overall course grades
#Hedge's g for course grades
#0.2 => small effect, 0.5 => medium effect, 0.8 => large effect
course <- select(metadata, n_female, course_grade_f, course_SD_f, course_SD_f_high, n_male,
course_grade_m, course_SD_m, course_SD_m_high, author, year, title, journal, university,
subject, Intro.or.nonintro, Biol.or.Chem, ped_cat, lab, class_size, label, imputed_SD)
course_hg<-escalc(measure="SMD", n1i=n_female, m1i=course_grade_f, sd1i=course_SD_f,
n2i=n_male, m2i=course_grade_m, sd2i=course_SD_m, data=course)
course_hg<-drop_na(course_hg, yi)
course_hg$assessment="course" #use this later
#exams
examAv <- select(metadata, n_female, allexams_f, allexams_SD_f, allexams_SD_f_high,
n_male, allexams_m, allexams_SD_m, allexams_SD_m_high, author, year, title, journal,
university, subject, Intro.or.nonintro, Biol.or.Chem, ped_cat, lab, class_size, label, imputed_SD)
examAv_hg<-escalc(measure="SMD", n1i=n_female, m1i=allexams_f, sd1i=allexams_SD_f,
n2i=n_male, m2i=allexams_m, sd2i=allexams_SD_m, data=examAv)
Sexam <- select(metadata, n_female, singleexam_f, singleexam_SD_f, singleexam_SD_f_high,
n_male, singleexam_m, singleexam_SD_m, singleexam_SD_m_high, author, year, title, journal,
university, subject, Intro.or.nonintro, Biol.or.Chem, ped_cat, lab, class_size, label, imputed_SD)
Sexam_hg<-escalc(measure="SMD", n1i=n_female, m1i=singleexam_f, sd1i=singleexam_SD_f,
n2i=n_male, m2i=singleexam_m, sd2i=singleexam_SD_m, data=Sexam)
```

```
#combine Hedge's g of average exams, single exams, and z-score exam averages
zexams_hg <- z_hg %>% select(n_female, n_male, author, year, title, journal, university,
subject, Intro.or.nonintro, Biol.or.Chem, ped_cat, lab, class_size, yi, vi, label, imputed_SD)
exams_hg <- bind_rows(examAv_hg, Sexam_hg, zexams_hg)
exams_hg <- drop_na(exams_hg, yi)
exams_hg$assessment="exams" #use this later
#CI
CI <-  select(metadata, n_female, CI_f, CI_SD_F, CI_SD_F_high, n_male, CI_m, CI_SD_m,
CI_SD_m_high, author, year, title, journal, university, subject, Intro.or.nonintro, Biol.or.Chem,
ped_cat, lab, class_size, label, imputed_SD)
CI_hg<-escalc(measure="SMD", n1i=n_female, m1i=CI_f, sd1i=CI_SD_F, n2i=n_male,
m2i=CI_m, sd2i=CI_SD_m, data=CI)
CI_hg<-drop_na(CI_hg, yi)
CI_hg$assessment = "CI" #use this later
###combine everything
long_hg <- bind_rows(course_hg, exams_hg, CI_hg) %>% select(yi, vi, assessment, label,
author, year, title, journal, university, subject, Intro.or.nonintro, Biol.or.Chem, ped_cat, lab,
class_size, imputed_SD)
##### dataset with high imputed SD
#course and quiz
course_hg_high<-escalc(measure="SMD", n1i=n_female, m1i=course_grade_f,
sd1i=course_SD_f_high, n2i=n_male, m2i=course_grade_m, sd2i=course_SD_m_high,
data=course)
course_hg_high<-drop_na(course_hg_high, yi)
course_hg_high$assessment="course" #use this later
long_hg_high <- bind_rows(course_hg_high, exams_hg, CI_hg) %>% select(yi, vi, assessment,
label, author, year, title, journal, university, subject, Intro.or.nonintro, Biol.or.Chem, ped_cat,
lab, class_size, imputed_SD)
#clean up environment
remove(CI, CI_hg, course, course_hg, course_hg_high, examAv, examAv_hg, exams_hg,
Sexam, Sexam_hg, zexams_hg)

# summary graphs ------------------------------------------------------------
#assessment type bar graph
ggplot(data=long_hg, aes(x=assessment, fill=assessment))+
  geom_bar()+
  xlab("Assessment Type")+
  theme(text = element_text(size = 40))+
  scale_fill_brewer(palette="Set1")+
  theme(legend.position = "none")
#class size histogram
ggplot(data = long_hg, aes(x=class_size))+
  geom_histogram()+
  xlab("Class size")+
  theme(text = element_text(size = 40))+
  xlim(0, 1050)
```

```
summary(long_hg$class_size)
#Intro/upper and Biol/Chem bar graph
ggplot(data=long_hg, aes(fill=Intro.or.nonintro, x=Biol.or.Chem))+
  geom_bar()+
  theme(text = element_text(size = 40))+
  #geom_bar(position = position_dodge())+
  xlab("Course")+
  labs(fill="Level")
#pedagogy pie chart
ped_data <- long_hg%>% filter(ped_cat!= "")
#161 total (72.9% of total classes)
count(ped_data, ped_cat)
#active: 43.5%
#interactive: 18.0%
#lecture: 38.5%
blank_theme <- theme_minimal()+
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.border = element_blank(),
    panel.grid=element_blank(),
    axis.ticks = element_blank(),
    plot.title=element_text(size=14, face="bold")
  )
ggplot(data=(long_hg%>% filter(ped_cat!= "")), aes(x=factor(1), fill=ped_cat))+
  geom_bar(width = 1)+
  coord_polar("y")+
  blank_theme +
  theme(axis.text.x=element_blank())+
  geom_text(x=1, y=30, label="38.5%", size=10)+
  geom_text(x=1.1, y=75, label="18.0%", size=10)+
  geom_text(x=1, y=130, label="43.5%", size=10)+
  scale_fill_discrete(name = "Pedagogy")+
  theme(text = element_text(size = 30))

# full assessments and funnel plots ------------------------------------
results_overall=rma(yi, vi, data=long_hg , method="HE")
forest(results_overall)
#model with high standard deviation
results_highSD=rma(yi, vi, data=long_hg_high , method="HE")
forest(results_highSD)
#publication bias check
#funnel plot
funnel(results_overall, main="Standard Error Forest Plot of Full Dataset")
#trim and fill
taf <- trimfill(results_overall)
```

```
funnel(taf, legend=TRUE)
#fail safe n
fsn(yi, vi, data=long_hg, type="Rosenberg", alpha=.05, digits=4)

# regression analysis ----------------------------------------------------
#set up dataset
data_0 <- long_hg %>% select(yi, vi, label, assessment, title, ped_cat, Biol.or.Chem,
Intro.or.nonintro, class_size, university, subject) %>%
  filter(ped_cat=="active"| ped_cat=="interactive"| ped_cat=="lecture") %>% na.omit()
#recode for proper order:
data_0 <- data_0 %>% mutate(assessment=recode(assessment, "exams" = "aexams"))
library(MuMIn)
eval(metafor:::.MuMIn)
#random effect: university and subject
full.university.subject <- rma.mv(yi, vi, mods = ~assessment +
                            ped_cat + Biol.or.Chem + Intro.or.nonintro +
                            class_size, random = ~ 1 | university/subject, data=data_0) #not sure if
order matters here
res.university.subject <- dredge(full.university.subject, trace=2)
subset(res.university.subject, delta <= 2, recalc.weights=FALSE)

res.30 <- rma.mv(yi, vi, mods = ~assessment +
                ped_cat + Intro.or.nonintro +
                class_size, random = ~ 1 | university/subject, data=data_0) #####best model

####pairwise comparisons
library(multcomp)
#assessment type
res <- rma.mv(yi, vi, mods = ~factor(assessment) +
              ped_cat + Biol.or.Chem + Intro.or.nonintro +
              class_size -1 #-1 removes intercept, so now all factors are included
            , data=data_0)
summary(glht(res, linfct=cbind(contrMat(c("aexams"=1, "CI"=1, "course"=1), type = "Tukey"),
0,0,0,0,0)), test=adjusted("holm"))

#pedagogy
res <- rma.mv(yi, vi, mods = ~factor(ped_cat) + #swapped pedagogy and assessment order so it
would expand the right one
              assessment + Biol.or.Chem + Intro.or.nonintro +
              class_size -1 #-1 removes intercept, so now all factors are included
            , data=data_0)
summary(glht(res, linfct=cbind(contrMat(c("active"=1, "interactive"=1, "lecture"=1), type =
"Tukey"), 0,0,0,0,0)), test=adjusted("holm"))
#assessment within pedagogy
data_0$dummy = interaction(data_0$ped_cat, data_0$assessment, drop = T)
res.dummy <- rma.mv(yi, vi, mods = ~dummy + Intro.or.nonintro +
```

```
                class_size -1, random = ~ 1 | university/subject, data=data_0) #####best model

summary(glht(res.dummy, linfct=rbind(c(-1,0,0,0,0,1,0,0,0,0), #active
                                     c(0,-1,0,0,0,0,1,0,0,0), #interactive
                                     c(0,0,-1,0,0,0,0,1,0,0) #lecturing
)), test=adjusted("holm"))

# graphs of subgroups ----------------------------------------------------
#assessment type
course.res <- long_hg %>% filter(assessment=="course") %>% rma(yi, vi, data=. ,
method="HE")
exams.res <- long_hg %>% filter(assessment=="exams") %>% rma(yi, vi, data=. ,
method="HE")
CI.res <- long_hg %>% filter(assessment=="CI") %>% rma(yi, vi, data=. , method="HE")
#count
long_hg %>% filter(assessment=="course") %>% count()
long_hg %>% filter(assessment=="exams") %>% count()
long_hg %>% filter(assessment=="CI") %>% count()
forest(c(0,0,0), c(0,0,0), xlim=c(-2, 1),alim=c(-1, 1), rows = 20, xlab = "Hedge's g") #create a
"blank" template
#add polygons of groups and
addpoly(course.res, row=3)
text(-1, 3, paste("Course, n=114"))
addpoly(exams.res, row=2)
text(-1, 2, paste("Exams, n=121"))
addpoly(CI.res, row=1)
text(-1, 1, paste("Concept inventory, n=11"))
#add title
text(0,4.5, paste("Overall Hedge's g  by Assessment Type"))
#pedagogy
active.res <- long_hg %>% filter(ped_cat=="active") %>% rma(yi, vi, data=. , method="HE")
interactive.res <- long_hg %>% filter(ped_cat=="interactive") %>% rma(yi, vi, data=. ,
method="HE")
lecture.res <- long_hg %>% filter(ped_cat=="lecture") %>% rma(yi, vi, data=. , method="HE")
#count
long_hg %>% filter(ped_cat=="active") %>% count()
long_hg %>% filter(ped_cat=="interactive") %>% count()
long_hg %>% filter(ped_cat=="lecture") %>% count()
forest(c(0,0,0), c(0,0,0), xlim=c(-2, 1),alim=c(-1, 1), rows = 20, xlab = "Hedge's g") #create a
"blank" template
#add polygons of groups and
addpoly(active.res, row=3)
text(-1, 3, paste("Active, n=71"))
addpoly(interactive.res, row=2)
text(-1, 2, paste("Interactive, n=29"))
addpoly(lecture.res, row=1)
```

```
text(-1, 1, paste("Lecture, n=63"))
#add title
text(0,4.5, paste("Overall Hedge's g  by Pedagogy"))
#pedagogy x assesessment
active.course <- long_hg %>% filter(ped_cat=="active", assessment=="course") %>% rma(yi, vi, data=. , method="HE")
active.exams <- long_hg %>% filter(ped_cat=="active", assessment=="exams") %>% rma(yi, vi, data=. , method="HE")
interactive.course <- long_hg %>% filter(ped_cat=="interactive", assessment=="course") %>% rma(yi, vi, data=. , method="HE")
interactive.exams <- long_hg %>% filter(ped_cat=="interactive", assessment=="exams") %>% rma(yi, vi, data=. , method="HE")
lecture.course <- long_hg %>% filter(ped_cat=="lecture", assessment=="course") %>% rma(yi, vi, data=. , method="HE")
lecture.exams <- long_hg %>% filter(ped_cat=="lecture", assessment=="exams") %>% rma(yi, vi, data=. , method="HE")
forest(c(0,0,0,0,0,0,0,0,0,0), c(0,0,0,0,0,0,0,0,0,0), xlim=c(-3, 2),alim=c(-1, 1), rows = 25, xlab = "Hedge's g") #create a "blank" template
#add polygons of groups and
addpoly(active.course, row=1)
text(-1.5, 1, paste("Active, course"))
addpoly(active.exams, row=2)
text(-1.5, 2, paste("Active, exams"))
addpoly(interactive.course, row=4)
text(-1.5, 4, paste("Interactive, course"))
addpoly(interactive.exams, row=5)
text(-1.5, 5, paste("Interactive, exams"))
addpoly(lecture.course, row=7)
text(-1.5, 7, paste("Lecture, course"))
addpoly(lecture.exams, row=8)
text(-1.5, 8, paste("Lecture, exams"))
#add title
text(0,7.5, paste("assessment x pedagogy"))
```

Appendix 4. R code for Chapter 2

```
# ##setup ----------------------------------------------------------------
setwd("C:/Users/sarao/OneDrive/Desktop/school related files/GSR Project/class_graphs")
library(tidyverse)
#import data
#with dataset used in STAT class
GSR <- read.csv("C:/Users/sarao/OneDrive/Desktop/school related files/GSR
        Project/GSR2instructornopost2.csv", stringsAsFactors=TRUE) %>%
  mutate(Minutes=Seconds/60)# %>% mutate(activity=recode(activity, "Preclass" =
        "aPreclass"))
#this one already has activity designations, preclass averages (checked using R)
#redo variation because it was wrong in original
GSR$variation <- GSR$GSR - GSR$PreAverage #now the errors should be fixed
#remove contested data
#create new variable that pastes class and date
GSR$combo <- paste0(GSR$class, GSR$date)
GSR_no <- GSR %>% filter(combo != "CHEM4/25/2018") %>% filter(combo !=
        "Esci4/30/2018")
GSR_no_uncategorized <- GSR_no %>% filter(activity != "Uncategorized") #remove
        uncategorized
# model  ----------------------------------------------------------
library(lme4)
library(lmerTest)
#remove baseline:
GSR_activities <- GSR_no_uncategorized %>% filter(activity != "aPreclass")
#reordered factors in original doc
GSR.model=lmer(variation~activity*gender+(1|ID)+(1|class), data=GSR_activities, REML =
        FALSE)
summary(GSR.model)
#pairwise comparisons
library(multcomp)
#activity x gender
GSR_activities$IntFac <- interaction(GSR_activities$gender, GSR_activities$activity, drop = T)
GSR.int <- lmer(variation~IntFac+(1|ID)+ (1|class)-1, data=GSR_activities, REML = FALSE)
summary(GSR.int)
#test with limited comparisons
summary(glht(GSR.int, linfct=rbind(c(-1,1,0,0,0,0,0,0), #lecturing
                                   c(0,0,-1,1,0,0,0,0), #clickerQ
                                   c(0,0,0,0,-1,1,0,0), #Demo
                                   c(0,0,0,0,0,0,-1,1) #otherQ
)), test=adjusted("bonferroni"))
# graphs ----------------------------------------------------------------
#general descriptions:
#gender breakdown
```

```
blank_theme <- theme_minimal()+
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.border = element_blank(),
    panel.grid=element_blank(),
    axis.ticks = element_blank(),
    plot.title=element_text(size=14, face="bold")
  )
ggplot(data=GSR_no, aes(x=factor(1), fill=gender))+
  geom_bar(width = 1)+
  coord_polar("y")+
  blank_theme +
  theme(axis.text.x=element_blank())+
  ggtitle("Gender Distribution")+
  theme(plot.title = element_text(hjust = 0.5))+
  #theme(legend.position="top")+
  scale_fill_manual(values=c("#FFCC33", "#990000"), labels=c("Women", "Men"))+
  theme(legend.title = element_blank())+
  theme(text = element_text(size = 40))
#  scale_fill_discrete(labels=c("Women", "Men"))
#percentages
filter(GSR_no, gender=="M")%>% count() #58.6%
filter(GSR_no, gender=="F")%>% count() #41.4%

#activity breakdown
ggplot(GSR_no, aes(x=class, fill=activity))+
  geom_bar()+
  theme_classic()+
  scale_fill_brewer(palette="Paired")+
  labs(fill="Activity Category")+
  scale_fill_discrete(labels=c("Pre-class", "Lecturing", "Clicker \nQuestions", "Demonstrations",
          "Other \nQuestions", "Uncategorized"))+
  scale_x_discrete(labels=c("Human \nEvolution", "General\nChemistry",
          "Environmental\nSciences", "Introductory\nPhysics I", "Introductory\nPhysics II"))+
  theme(text = element_text(size = 20))
####comparisons of GSR
ggplot(GSR_activities, aes(x=activity, y=variation, color=gender))+
  geom_boxplot()+
  stat_summary(fun.y=mean, geom="point", shape=18, size=3, position =
          position_dodge(width=0.75)) +
 # theme_classic()+
  scale_color_manual(values=c("#FFCC33", "#990000"),labels=c("Women", "Men"))+
  theme(legend.title = element_blank())+
  theme(text = element_text(size = 20))+
```

```
  scale_x_discrete(labels=c("Lecturing", "Clicker \nQuestions", "Demonstrations", "Other
      \nQuestions"))+
  labs(x="Activity Category", y="Adjusted GSR (µS)")
  ggtitle("GSR variation from baseline average")
  ggplot(GSR_no, aes(x=activity, y=variation))+
  geom_boxplot()
#general summary statistics:
GSR_activities %>% group_by(activity) %>% summarise(average=mean(variation),
      options(pillar.sigfig=6))
GSR_activities %>% group_by(activity, gender) %>% summarise(average=mean(variation),
      options(pillar.sigfig=6))
# example GSR class graph ------------------------------------------------
subgroup <- GSR %>% filter(class=="CHEM") %>% filter(date=="4/18/2018") %>%
      filter(ID=="PR")
ggplot(subgroup, aes(x=Seconds, y=GSR, color=ID))+
  geom_point(aes(y=10000, color=activity))+
  geom_line(size=1)+
  geom_line(aes(y=13960),linetype="dotted", color="black")+
  labs(x="Time (seconds from class beginning)", y="GSR (µS)")+
  theme(text = element_text(size = 20))
```