AERA Open
October-December 2017, Vol. 3, No. 4, pp. 1–12
DOI: 10.1177/2332858417743754
© The Author(s) 2017. http://journals.sagepub.com/home/ero

Patterns of Gendered Performance Differences in Large Introductory Courses at Five Research Universities

Rebecca L. Matz
Michigan State University
Benjamin P. Koester
University of Michigan
Stefano Fiorini
Indiana University
Galina Grom
University of Michigan
Linda Shepard
Indiana University
Charles G. Stangor
University of Maryland
Brad Weiner
University of Minnesota
Timothy A. McKay
University of Michigan

Significant gendered performance differences are signals of systemic inequity in higher education. Understanding of these inequities has been hampered by the local nature of prior studies; consistent measures of performance disparity across many disciplines and institutions have not been available. Here, we report the first wide-ranging, multi-institution measures of gendered performance difference, examining more than a million student enrollments in hundreds of courses at five universities. After controlling for factors that relate to academic performance using optimal matching, we identify patterns of gendered performance difference that are consistent across these universities. Biology, chemistry, physics, accounting, and economics lecture courses regularly exhibit gendered performance differences that are statistically and materially significant, whereas lab courses in the same subjects do not. These results reinforce the importance of broad investigation of performance disparities across higher education. They also help focus equity research on the structure and evaluative schemes of these lecture courses.

Keywords: *achievement gap, course grade, gender studies, grade point average, higher education, laboratory, lecture, postsecondary education, science education, statistics, STEM*

In a recent review of research on gendered performance disparities in undergraduate science, technology, engineering, and mathematics (STEM) courses, Eddy and Brownell (2016) describe a confused research landscape: Some courses favor men, some favor women, and some show little bias. Their review calls specifically for systematic measurement of performance gaps across an array of disciplines and institutions, all accounting for prior academic performance, in the hope that emergent patterns might inform our understanding of "the relative contributions of different factors to performance and/or persistence in STEM." In this study, we answer this call, analyzing data on more than a million student enrollments in hundreds of courses drawn from five research-intensive public universities in the Big Ten Academic Alliance.

We find evidence of statistically significant, persistent gendered performance differences (GPDs) in some large, introductory courses, differences that are also materially significant. In particular, men earned relatively higher grades than women in biology, chemistry, physics, accounting, and economics lecture courses, even after accounting for the influence of some measures of prior academic achievements on performance. These results are remarkably consistent across all five universities, which together enroll more than 150,000 undergraduate students in any given year. These patterns confirm the importance of conducting systematic studies of performance equity, provide the impetus for extending this work to other sectors of higher education, and focus research attention on the structure and evaluative schemes of these lecture courses.

Background

Women have achieved parity with men on many indicators of educational outcomes; indeed, women now outpace men in terms of college enrollment and overall attainment of bachelor's and higher-level degrees (Snyder & Dillow, 2015). Still, significant gaps in enrollment and degree attainment remain in engineering, mathematics, computer science, and physical science disciplines (DiPrete & Buchmann, 2013). Even in the life sciences, where women now dominate numerically in terms of awarded bachelor's degrees (Mann & DiPrete, 2013), gendered differences that favor men have been identified in exam performance, participation in whole-class discussions, and who is viewed as most knowledgeable about course content (Eddy, Brownell, & Wenderoth, 2014; Grunspan et al., 2016). These gaps undermine the national priority that all students have the opportunity to fully participate in STEM fields (President's Council of Advisors on Science and Technology, 2012). Because sex is a legally protected class, disparate educational outcomes for male and female students also raise important questions of equity.

Girls and boys pursue science and mathematics courses in primary and secondary school in roughly equal proportion, but by the time they are freshmen in college, men are more likely to choose a science or mathematics major (Hill, Corbett, & St. Rose, 2010), and the underrepresentation of women in these disciplines carries all the way through to the professoriate (Urry, 2015). Research from diverse academic disciplines shows that a variety of factors affect gendered differences in STEM major selection, degree attainment, and careers (Blickenstaff, 2005). When considering performance in undergraduate STEM courses (the level of interest in this study), prior academic performance, engagement, and affective variables are all considered relevant constructs for investigating and explaining gendered differences (Eddy & Brownell, 2016). Here, we briefly review a range of important factors that influence the decisions of women and men about pursuing undergraduate STEM courses and degree programs.

Psychological and environmental factors have been shown to contribute to observed gendered gaps (Murphy, Steele, & Gross, 2007), such as the perpetuation of a "fixed mind-set" model that tends to favor men (Good, Rattan, & Dweck, 2012) as well as stereotype threat, which has been shown to reduce the performance of female students in mathematics when gendered stereotypes are invoked (H. Johnson, Barnard-Brak, Saxon, & Johnson, 2012). Microaggressions, brief and often subtle messages based on membership in a group (Sue, 2010), have been shown to act as a barrier to participation in STEM (Grossman & Porche, 2014). Unconscious bias plays a role as well. For example, biology, chemistry, and physics faculty members, regardless of their own gender, have been shown to view a male undergraduate job candidate as more competent and employable than an identical (excepting the name) female candidate (Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012). The affective dimensions of confidence and interest have also been linked to gendered differences that can impact course performance; women have reported feeling less confident than men in their calculus and engineering abilities (Ellis, Fosdick, & Rasmussen, 2016; Micari, Pazor, & Hartmann, 2007) and have provided more pessimistic self-reports of performance on a science assessment and reported subsequent diminished interest in scientific activities (Ehrlinger & Dunning, 2003). Hazari, Sonnert, Sadler, and Shanahan (2010) demonstrated that explicitly discussing gendered gaps in science positively impacted physics identity for women, which in turn strongly predicted decisions about pursuing a physics career.

The culture in undergraduate STEM courses—broadly including pedagogy, curriculum, assessment, instruction, and interaction between students and faculty—has been a major point of study with respect to gendered differences. Women in engineering programs have been shown to more frequently perceive gendered discrimination than men (Vogt, Hocevar, & Hagedorn, 2007), and some women in science courses have described being discouraged in attending large introductory classes where they felt anonymous, responding to and posing questions in class, and engaging with faculty in research (A. Johnson, 2007). Regular interaction with faculty, which is certainly easier to facilitate in smaller courses, has been shown to positively influence STEM degree completion rates for all students but especially so for women (Gayles & Ampaw, 2014).

Although experiences in college classrooms are no doubt meaningful for students, Ceci, Ginther, Kahn, and Williams (2014) argue that the roots of gendered differences in mathematically intensive fields are solely in precollege experiences that, among other outcomes, influence the likelihood of men and women pursuing different degree programs. Other studies find that showing an initial interest in STEM fields at the middle or high school level is indeed predictive of STEM degree completion but that demonstrating interest in college is still a significant factor on the pathway to a STEM degree (Maltese, Melki, & Wiebke, 2014; Maltese & Tai, 2011).

In this work, we focus on one important part of the pathway to STEM degree completion: large, foundational university courses from a range of disciplines. The gatekeeping nature of these courses is widely acknowledged—indeed, they are often collectively described as “gateway courses.” To measure GPDs, we use data likely to be available on every college campus: grades in each course compared to expectations formed from an array of prior performance information, including grades in other courses and standardized admissions test scores.

Grades have been widely criticized as poor measures of learning. Nevertheless, grades remain the only measure of academic achievement that all institutions reliably record and value. They are taken seriously by institutions, used as a threshold for passage of courses, to select students for academic awards and honors, and even to dismiss students from campus. As a result, good grades are aggressively pursued by students, sometimes to the detriment of learning (Pulfrey, Buchs, & Butera, 2011). Inflation in average grades over time has been widely reported (D. Freeman, 1999; Jewell & McPherson, 2012; Kostal, Kuncel, & Sackett, 2016; Rojstaczer & Healy, 2012). Although in some contexts this might raise concerns about their utility for comparing student performance, it is worth noting that grade inflation has been relatively small in the foundational courses we study here (Achen & Courant, 2009).

Grades constitute the only universally accessible performance feedback provided to students. Student performance in a course, particularly performance relative to that in other courses, plays an important role in shaping major and career choices (Ost, 2010). For all these reasons, understanding and ultimately addressing the GPDs we report here is essential for ensuring equitable access to participation in STEM careers.

Method

This cross-institutional study is based on administrative data, so we restricted ourselves to covariates that are readily available, complete, and similarly defined from campus to campus. Following a method described by Huberth, Chen, Tritz, and McKay (2015), we used a measure called grade point average in other courses (GPAO) because it is a powerful predictor of students’ final course grades. GPAO is the cumulative GPA for a student calculated across all semesters, including the current semester, excluding only the course enrollment being analyzed. As such, GPAO is a property of a given course enrollment but does not exist when a student has taken only a single course.

In a population largely similar to the students represented in these data, and across a similar range of STEM, social science, and humanities courses, prior research (Huberth et al., 2015; Koester, Grom, & McKay, 2016) has shown GPAO to exhibit the strongest correlation with course grades out of measures regularly collected in administrative databases (e.g., high school GPA and standardized test scores). As the most important predictor of course performance in these studies, GPAO was shown to independently account for 32% of the variance in final physics course grades. Further, traditional cumulative GPAs are known to be good predictors of college outcomes (Creech & Sweeder, 2012; S. Freeman et al., 2007; Gershenfeld, Hood, & Zhan, 2016). Koester et al. (2016) found that additional covariates recorded in administrative data, such as estimated gross income and college of admission, correlate with course grades but explain negligible additional variation in grades.

As such, GPAO helps account for many potential confounding variables that influence student achievement, reducing both systematic and random sources of error. Because each grade is considered relative only to other courses from that institution, the GPAO measure facilitates cross-institutional comparisons even when courses at different institutions may be subject to different grading practices or degrees of grade inflation. Although GPAO is clearly sensitive to the mix of courses each student takes, when comparing all students in a given course of interest, we find empirically that enrolled women and men have taken other courses with similar levels of difficulty (see online Supplementary Materials Section 2.6). Differences in GPAO for female and male students emerge from differences in performance in their other courses rather than from difference in overall grading patterns in those courses.

Six years of student record data for introductory courses were collected at each of five large, public research universities. These student-level data were locally maintained and analyzed separately at each institution using common code written in R (detailed methods are available in the online Supplementary Materials). The overall data set includes 1,122,586 course enrollments across 249 courses in 13 disciplines. The courses are primarily from STEM (i.e., biology, chemistry, engineering, mathematics, physics, and statistics) and social science (i.e., communication, economics, political science, psychology, and sociology) disciplines; accounting and writing courses are included from the business and humanities disciplines, respectively. Selection criteria for both the disciplines and comparable courses are provided in the online Supplementary Materials Section 2.4. In addition to disciplines and generalized course names, we provide classifications related to course structure by labeling each course as either *lecture, lab*, or *mixed*. Mixed courses are usually worth four or five credits and contain elements of both a lecture and a lab.

For each course, we focus on two measures: the average grade anomaly (AGA) and the GPD. AGA compares students’ performance in this course to their other courses; it is simply the difference between final course grade and GPAO averaged across all student enrollments in the course. A positive AGA for a course indicates that, overall, students’ final grades in the selected course tended to be higher than their GPAO. We call a positive AGA a *grade bonus*. In contrast, a negative course AGA indicates that, overall, students' final grades in the selected course tended to be lower than their GPAO. We call a negative AGA like this a *grade penalty*. In general, AGA measures average student performance relative to expectation.

GPD compares the AGA between women and men, that is, it measures the gendered difference in performance relative to expectation. A GPD can result from differences between men and women in final course grade, GPAO, or both. For example, a GPD that favors men could result from either (a) men and women having similar GPAOs, with men earning higher grades in the targeted course; (b) women having a higher GPAO than men, with women and men earning similar course grades; or (c) a combination of these two scenarios. Our convention in this work is that a positive GPD favors women and a negative GPD favors men.

Although GPAO is by far the strongest predictor of student performance (Huberth et al., 2015), it is possible that other factors might account for observed GPDs. To test this possibility, the GPD in each course is additionally calculated while accounting for a combination of GPAO, SAT, or ACT Mathematics and English subscores (converting when necessary) and individual course/term factors. Two methods were used: multiple linear regression and optimal matching. Each analysis offers a strength. The regression method is a familiar way to correct for confounding factors and in comparison to the matching method is more precise (i.e., has a smaller standard error) but is also less accurate because of founding assumptions (e.g., that all predictors are linearly independent). The matching method is often noisy and less precise than the regression method, but it is more accurate.

With course grade as the dependent variable, the following covariates for the regression model were selected based on the LASSO (least absolute shrinkage and selection operator) method (Hastie, Tibshirani, & Friedman, 2009) as well as the restrictions inherent in comparing multiple institutions: gender, GPAO, ACT Mathematics and English subscores, and term. Term was included as a categorical variable to account for term-to-term variation in instructors and the time of year that the course was offered, as differences between "on-" and "off-semester" student populations are common.

The matching model included the same factors as the regression model and relies on propensity scores for matching cases and controls (Hansen, 2004, 2007). Matching was performed on a term-by-term basis so long as in each term the course contained more than 50 students; nine courses would otherwise have been included in the data but did not meet this criterion. The differences in GPDs obtained by the matching and regression methods are marginal at best (see online Supplementary Figure S4 and Supplementary Table S6), with the regression and matching methods resulting in the highest and lowest GPD, respectively. Over multiple iterations of this work, we found similar results even with small changes to the baseline predictive model (i.e., to the covariates) using both methods. Analyses were performed using a custom R code (see online Supplementary Materials Section 2.4), and figures were developed using Excel and Tableau, an interactive data visualization program.

In what follows, we report results using GPDs measured by the optimal matching method. On average, this method returns the most conservative measures of GPD of the three approaches, accounting as thoroughly as possible for each student's prior academic performance.

Results

Comparison across STEM disciplines reveals two trends when each course is characterized as a lecture, lab, or mixed (Figure 1). First, the majority of lecture (74%) and mixed (93%) courses yield a grade penalty (negative AGA), and the majority of lab courses (64%) yield a grade bonus (positive AGA) for students (chi-square $p < .001$, Cramer's V = .33). Second, the lecture and mixed courses that yield grade penalties tend to favor men (negative GPD), meaning men have smaller grade penalties in these courses than women. The average GPDs across lecture and mixed courses are –.07 and –.10 grade points, respectively. The lab courses that yield grade bonuses tend to be more equitable, with an average GPD across all lab courses of .01 grade points.

Separating these data by discipline shows that both trends are apparent across biology, chemistry, engineering, and physics courses (Figure 2). It is worth noting that both AGAs and GPDs are especially large and negative for large general chemistry courses—the first STEM courses encountered by many college students. Mathematics and statistics courses exhibit somewhat different patterns. Although the majority of these courses yield a grade penalty, overall they appear to favor neither men nor women, with an average GPD across all courses of –.03 grade points. This result is not unexpected as the ACT Mathematics score covariate, although still second to GPAO, may reasonably explain more of the grade variation in mathematics and statistics courses than it does in biology, chemistry, and physics.

Comparison among non-STEM courses in these data (Figure 3) shows that the majority of introductory accounting and economics courses produce grade penalties that favor men (average GPD = –.14), exhibiting a pattern similar to STEM lecture and mixed courses. Conversely, writing courses yield grade bonuses that slightly but significantly favor women (average GPD = .06). Overall, social science courses exhibit little to no GPDs (average GPD = .01). We note again that women tend to slightly outperform men overall in college (Keiser, Sackett, Kuncel, & Brothen, 2016); the lecture courses with significant GPDs favoring men are unusual within the college landscape.

At the individual course level, the final course grade and GPAO contribute differently to the observed GPDs. In some

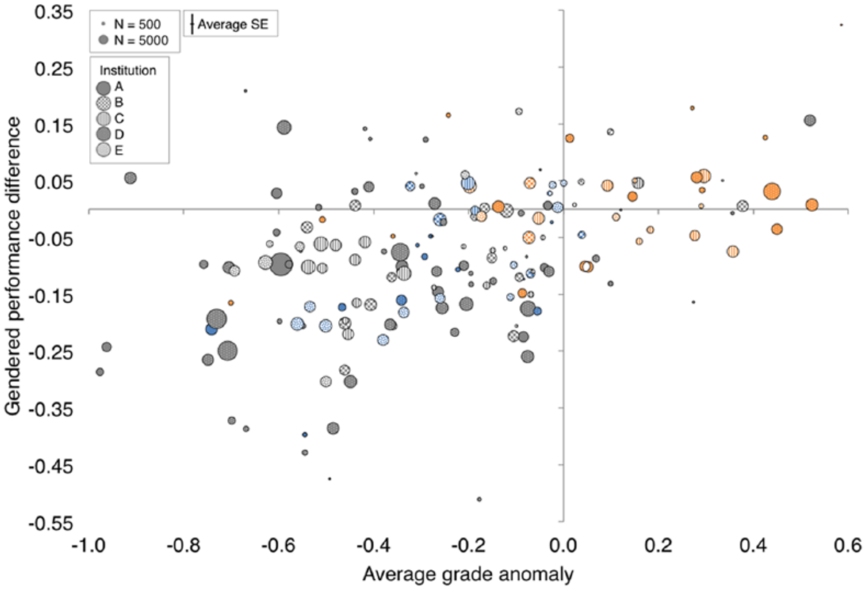

FIGURE 1. *Gendered performance differences in science, technology, engineering, and mathematics (STEM) courses. Gendered performance differences based on a matching method versus average grade anomaly for 172 introductory STEM courses across five universities, representing 677,949 course enrollments, including lectures (gray), labs (orange), and mixed courses (blue). Crosshairs indicate the average standard error on the mean.*

cases, the GPDs that favor men tend to result from women having higher GPAOs than men yet earning similar or slightly lower grades. In other cases, we find small GPAO differences but substantial final grade differences such that most of the GPD can be attributed to the difference in grades. Regardless, we find that these differences are stable over time at the individual course level, robust to changes in instructors and the time of year the course was taken (Figure 4).

Discussion

In these data, we find evidence that GPDs in many courses, although modest in size, are statistically significant and reliably present from term to term; that GPDs in biology, chemistry, and physics lecture courses tend to favor men, whereas those in corresponding lab courses tend to be more equitable; that writing and social science courses (with the exception of economics) do not tend to yield substantial GPDs; and that these results are consistent across five relatively similar universities and six academic years. These patterns mirror those observed in a precursor study local to University of Michigan even though that study accounted for high school GPA, which has a small but unique amount of power in predicting grades (Koester et al., 2016).

We do not focus on precise measurement of the magnitudes of these GPDs. Rather, we stress that, for some courses (e.g., biology, chemistry, physics, accounting, and economics lectures), they are materially significant. Scholarships, university honors, and even employment decisions rely heavily on GPA, often turning on tenth-of-a-point distinctions. Students respond individually to grade signals they receive as well, and prior research suggests that response to these signals may be gendered, compounding the potential impact of modest performance differences (Rask & Teifenthaler, 2008).

Further empirical research is required to ascertain what magnitude of GPD is meaningful to students in which contexts and to what extent the differences might accumulate throughout a student's degree program. Nevertheless, the presence of statistically and materially significant GPDs in an array of courses creates significant equity concerns for these institutions. It is also important to use parallel measures of performance equity to explore other aspects of identity and background that might intersect with gender, including race and ethnicity, first-generation status, and socioeconomic status. Although the data available for this study do not enable this analysis, we encourage others to pursue this work and provide some first insights from analyses at two institutions (see online Supplementary Materials Section 3).

AGAs themselves raise a different set of questions. That some courses are graded more harshly than others, and that these courses cluster by discipline, is well known and has been true since the adoption of letter grades (Goldman, Schmidt, Hewitt, & Fisher, 1974; King, 2015; Meyer, 1908). Still, this practice perpetuates a system in which it is normal

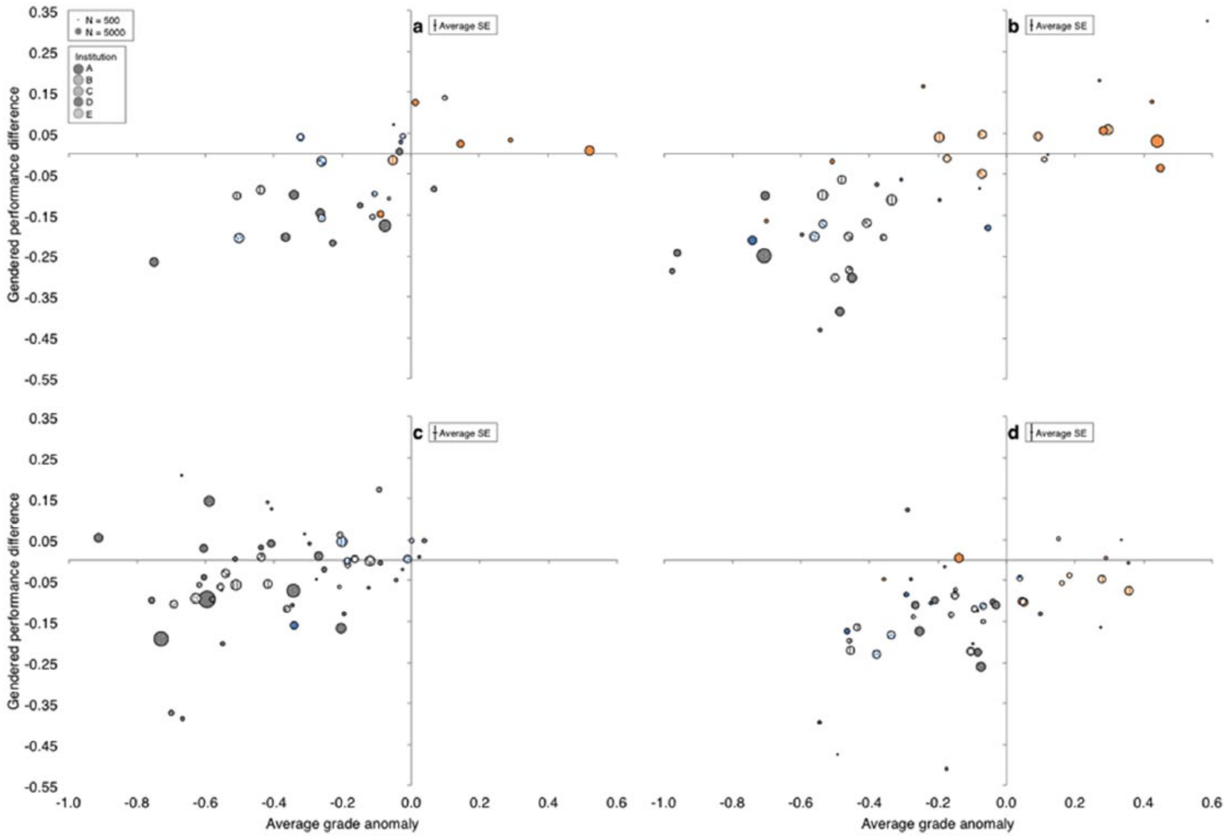

FIGURE 2. *Gendered performance differences in science, technology, engineering, and mathematics courses by discipline. Gendered performance differences based on matching versus average grade anomaly parsed by discipline for 28 (*n *= 115,066), 41 (*n *= 192,487), 53 (*n *= 222,498), and 47 (*n *= 130,908) introductory courses across five universities in (a) biology, (b) chemistry, (c) mathematics and statistics, and (d) physics, respectively, including enrollments in lectures (gray), labs (orange), and mixed courses (blue). Crosshairs indicate the average standard error on the mean. Engineering courses are not pictured due to a small sample size; these data are included in the online Supplementary Materials.*

to earn low grades in introductory STEM lecture courses, the starting point for most students who will eventually pursue a STEM major. The comparatively low grades received by students in these courses result from decisions about grading practices made by instructors rather than student ability. Indeed, there is clear evidence that students who take these low-graded courses are, by other measures, especially strong students (Koester, Fogel, Murdock, Grom, & McKay, 2017).

The GPDs we identify here in introductory biology, chemistry, physics, accounting, and economics lecture courses are surprising and clearly not predicted by students' prior performance. Although previous studies have identified GPDs in particular disciplines at particular institutions (Creech & Sweeder, 2012; Eddy et al., 2014; Kost, Pollock, & Finkelstein, 2009; Lauer et al., 2013; Rauschenberger & Sweeder, 2010), the results presented here provide the first comprehensive, cross-disciplinary picture of how consistent these trends are across an array of similar institutions.

Conflated characteristics of the courses studied here may contribute to patterns in the results. For example, large lecture courses most typically employ high-stakes, timed, and often multiple-choice exams to assess students, whereas lab courses are more often graded through written reports, projects, and lower-stakes quizzes. Although some conflicting evidence exists (Federer, Nehm, & Pearl, 2016; C. Wright et al., 2016), men tend to outperform women on multiple-choice items, and women tend to outperform men on constructed-response exercises (Garner & Engelhard, 1999; Madsen, McKagan, & Sayre, 2013; Weaver & Raptis, 2001). Particular cases in the data appear to support this claim. For instance, a reformed introductory biology course at Institution D that makes use only of constructed-response assessments shows a GPD half that of the prerequisite course in the introductory sequence, although it is also true that the reformed course draws from a subset of the student population in the traditional, prerequisite course.

Another characteristic that may be related to the patterns in GPDs is whether course work tends to be more competitive or collaborative and, relatedly, whether class sizes are large or small, respectively. Especially at large universities

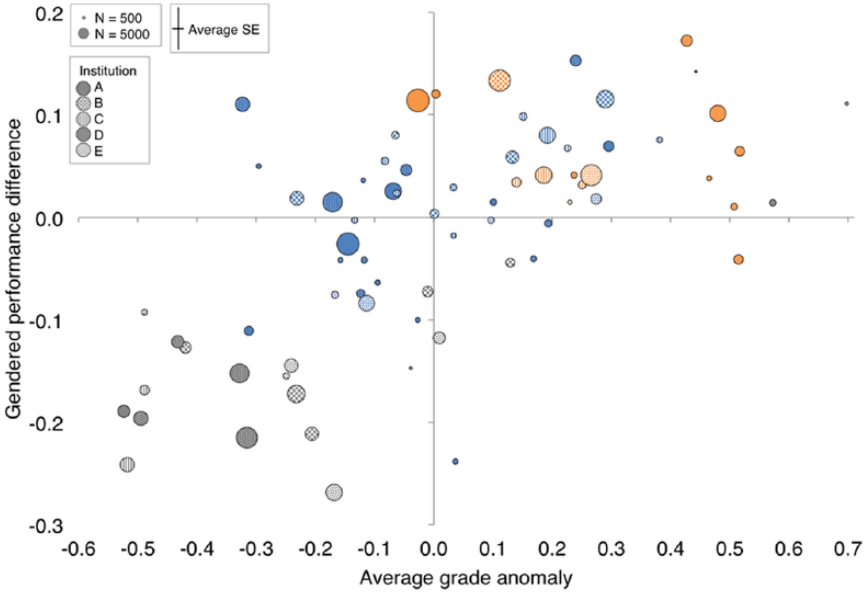

FIGURE 3. *Gendered performance differences in non–science, technology, engineering, and mathematics courses by discipline. Gendered performance differences based on matching versus average grade anomaly parsed by discipline for 22 (*n *= 160,828), 39 (*n *= 172,345), and 16 (*n *= 111,464) introductory lecture courses across five universities in accounting and economics (gray); communication, political science, psychology, and sociology (blue); and writing (orange). Crosshairs indicate the average standard error on the mean.*

like those studied here, lecture courses usually enroll hundreds of students per section, and lab courses usually enroll a few dozen students per section, making collaborative work easier to implement. Although, again, some conflicting evidence exists (Hazari et al., 2010; Micari et al., 2007; Pollock, Finkelstein, & Kost, 2007), women have often been shown to benefit from small-group work and small course sizes more than men (A. Johnson, 2007; Kokkelenberg, Dillon, & Christy, 2008; Lorenzo, Crouch, & Mazur, 2006; Rodger, Murray, & Cummings, 2007; Stump, Hilpert, Husman, Chung, & Kim, 2011), particularly in terms of student attitudes (Springer, Donovan, & Stanne, 1999).

Additionally, women have been shown to prefer collaborator as compared to leader/explainer roles (Eddy, Brownell, Thummaphan, Lan, & Wenderoth, 2015). Again, particular cases in the data appear to support this notion. For example, although each of the introductory engineering design courses included here is labeled as a lecture course in its respective course catalog, design courses usually center on conceiving and building a product with a group. It is unsurprising, then, that these engineering courses cluster with science labs in terms of yielding a grade bonus and generally favoring women. Further, with one exception, all writing courses, which are usually taught in small sections, in these data exhibit GPDs that favor women, and the two accounting courses that favor women have structural characteristics similar to lab courses. Patterns of GPD related to course structure call for research into equitable course design, raising questions of whether evaluative schemes in large lecture classes might disadvantage women as well as how best to support men in writing courses and group work situations.

The repetition of the observed performance differences on all five of these campuses reinforces the need for broader investigation of these patterns across the landscape of higher education. Are GPDs present at private research institutions; public, primarily undergraduate institutions; and community colleges? Although we reasonably expect these results would generalize to other, similar universities, we make no claims about the findings generalizing to other types of universities. These measurements of GPDs are relatively simple to make, relying on administrative data regularly gathered by every institution of higher education, and we encourage faculty, staff, and administrators involved in postsecondary STEM education to examine their own data. We hope that these results will provide the impetus for widespread equity analyses of this kind. When significant GPDs are found, steps should be taken to investigate and address them.

Social psychological interventions designed to improve student performance provide a potential solution, which is being widely explored. Because they do not require changing the structure or mode of instruction of courses, these relatively simple interventions (e.g., values affirmation or sense-of-belonging writing exercises) are attractive approaches to reducing GPDs. Although they have been found effective in some contexts (Miyake et al., 2010; Unkovic, Sen, & Quinn, 2016; Walton, Logel, Peach, Spencer, & Zanna, 2015), replication has not always been

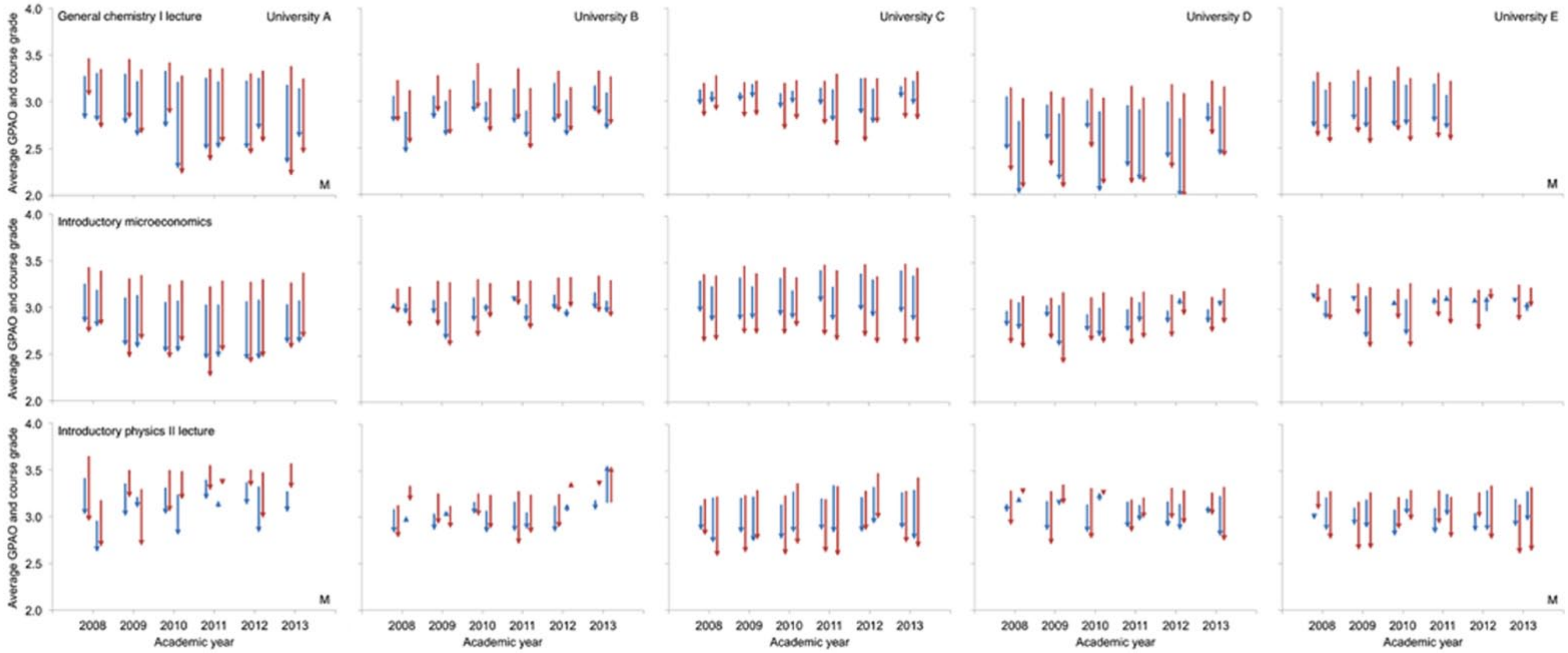

FIGURE 4. *Variation in grade and grade point average in other courses (GPAO) in similar courses across institutions. Differences between average GPAO (arrowtail) and average grade (arrowhead) for men (blue) and women (red) by academic year (e.g., 2008 indicates fall 2008 and spring 2009) in selected introductory chemistry, economics, and physics lecture and mixed (M) lecture and laboratory courses. This figure shows that gendered performance differences (GPDs) emerge from a combination of differences in GPAO, average grade, or both. Despite this, the GPDs are consistent from term to term, across changes in both the groups of students and instructors.*

possible (Madsen et al., 2013), and there is much to learn about how to apply these interventions at scale (Paunesku et al., 2015; Yeager & Walton, 2011). Large-scale randomized trials of the impact of values affirmation on GPDs are now in progress at one of our institutions.

These results suggest connections between GPDs and the structure of evaluation in courses. It is possible that modest changes to evaluative schemes might reduce GPDs, for example, reducing the time pressure in exams. It is also important to recognize differences among individual STEM disciplines (Cheryan, Ziegler, Montoya, & Jiang, 2017). The solution, then, is not to broadly prescribe a formulaic ratio of multiple-choice to constructed-response assessment items, suggest strict changes in class size, or ask faculty to send encouraging e-mails to particular groups of male or female students. Indeed, we agree with the assessment of Halpern et al. (2007) that "there are no single or simple answers to the complex questions about sex differences in science and mathematics" (p. 1).

This work should compel those at institutions of higher education to ask, as many are already doing (Elliott, 2016), how we can learn from this information to change practices in whatever ways are appropriate in our local contexts. Understanding student performance in context is an important step in pursuing equity (M. Wright, McKay, Hershock, Miller, & Tritz, 2014). Systems capable of personalizing at scale and responding to differences among students rather than prescribing a single solution for all hold some promise. Huberth et al. (2015), for example, describe a digital mentoring tool that is now being tested for its ability to reduce stereotype threat for women in high-enrollment undergraduate science courses.

Grades are consequential performance measures and clearly impact persistence (King, 2015). It is unclear whether men or women are more sensitive to their STEM grades in persistence decisions (Ost, 2010; Rask, 2010), and these differences may be field dependent. Regardless, grade penalties that are worse for female students than for male students create yet another headwind impeding gender equity in STEM. There is widespread evidence that faculty, especially male STEM faculty, are reluctant to accept evidence of gendered biases in STEM (Handley, Brown, Moss-Racusin, & Smith, 2015; Moss-Racusin, Molenda, & Cramer, 2015). In this light, continued investigation of GPDs, coupled with efforts to understand their correlates and causes, is imperative.

Unexplained GPDs of the kind reported here cannot be ignored or simply allowed to persist.

Acknowledgments

The Sloan Foundation (G-2014-14496) and the provosts of our universities supported this Big Ten Academic Alliance Learning Analytics project. Numerous faculty and staff at each university helped us determine which courses would be comparable. Peter Radcliffe and Sarah Ruhland at University of Minnesota kept this project afloat in a time of transition. Special support for this project was provided at Indiana University by vice provost Dennis Groth; Dawit Gelan and Yanan Feng at the Bloomington Assessment and Research office helped as well. Finally, we thank the anonymous reviewers for their substantive suggestions, which greatly improved this work.

References

Achen, A. C., & Courant, P. N. (2009). What are grades made of? *Journal of Economic Perspectives*, *23*(3), 77–92.

Blickenstaff, J. C. (2005). Women and science careers: Leaky pipeline or gender filter? *Gender and Education*, *17*(4), 369–386.

Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014). Women in academic science: A changing landscape. *Psychological Science in the Public Interest*, *15*(3), 75–141.

Cheryan, S., Ziegler, S. A., Montoya, A. K., & Jiang, L. (2017). Why are some STEM fields more gender balanced than others? *Psychological Bulletin*, *143*(1), 1–35. https://dx.doi.org/10.1037/bul0000052

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

DiPrete, T. A., & Buchmann, C. (2013). *The rise of women*. New York, NY: Russell Sage Foundation.

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., Brownell, S. E., Thummaphan, P., Lan, M.-C., & Wenderoth, M. P. (2015). Caution, student experience may vary: Social identities impact a student's experience in peer discussions. *CBE-Life Sciences Education*, *14*(4), ar45. https://doi.org/10.1187/cbe.15-05-0108

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.

Ehrlinger, J., & Dunning, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology*, *84*(1), 5–17. https://doi.org/10.1037/0022-3514.84.1.5

Elliott, S. L. (2016). From the editor-in-chief: Questions of gender equity in the undergraduate biology classroom. *Journal of Microbiology & Biology Education*, *17*(2), 186–188. https://doi.org/10.1128/jmbe.v17i2.1136

Ellis, J., Fosdick, B. K., & Rasmussen, C. (2016). Women 1.5 times more likely to leave STEM pipeline after calculus compared to men: Lack of mathematical confidence a potential culprit. *PLOS ONE*, *11*(7), e0157447. https://doi.org/10.1371/journal.pone.0157447

Federer, M. R., Nehm, R. H., & Pearl, D. K. (2016). Examining gender differences in written assessment tasks in biology: A case study of evolutionary explanations. *CBE-Life Sciences Education*, *15*(1), ar2. https://doi.org/10.1187/cbe.14-01-0018

Freeman, D. G. (1999). Grade divergence as a market outcome. *Journal of Economic Education*, *30*(4), 344–51.

Freeman, S., O'Connor, E., Parks, J. W., Cunningham, M., Hurley, D., Haak, D., . . . Wenderoth, M. P. (2007). Prescribed active learning increases performance in introductory biology. *CBE-Life Sciences Education*, *6*(2), 132–139. https://doi.org/10.1187/cbe.06-09-0194

Garner, M., & Engelhard, G. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education*, *12*(1), 29–51.

Gayles, J. G., & Ampaw, F. (2014). The impact of college experiences on degree completion in STEM fields at four-year institutions: Does gender matter? *Journal of Higher Education*, *85*(4), 439–468. https://doi.org/10.1080/00221546.2014.11777336

Gershenfeld, S., Hood, D. W., & Zhan, M. (2016). The role of first-semester GPA in predicting graduation rates of underrepresented students. *Journal of College Student Retention: Research, Theory & Practice*, *17*(4), 469–488.

Goldman, R. D., Schmidt, D. E., Hewitt, B. N., & Fisher, R. (1974). Grading practices in different major fields. *American Educational Research Journal*, *11*(4), 343–357. https://doi.org/10.3102/00028312011004343

Good, C., Rattan, A., & Dweck, C. S. (2012). Why do women opt out? Sense of belonging and women's representation in mathematics. *Journal of Personality and Social Psychology*, *102*(4), 700–717. https://doi.org/10.1037/a0026659

Grossman, J. W., & Porche, M. V. (2014). Perceived gender and racial/ethnic barriers to STEM success. *Urban Education*, *49*(6), 698–727.

Grunspan, D. Z., Eddy, S. L., Brownell, S. E., Wiggins, B. L., Crowe, A. J., & Goodreau, S. M. (2016). Males under-estimate academic performance of their female peers in undergraduate biology classrooms. *PLOS ONE*, *11*(2), e0148405. https://doi.org/10.1371/journal.pone.0148405

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest*, *8*(1), 1–51. https://doi.org/10.1111/j.1529-1006.2007.00032.x

Handley, I. M., Brown, E. R., Moss-Racusin, C. A., & Smith, J. L. (2015). Quality of evidence revealing subtle gender biases in science is in the eye of the beholder. *Proceedings of the National Academy of Sciences*, *112*(43), 13201–13206.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, *99*(467), 609–618. https://doi.org/10.1198/016214504000000647

Hansen, B. B. (2007). Optmatch: Flexible, optimal matching for observational studies. *R News*, *7*(2), 18–24.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.

Hazari, Z., Sonnert, G., Sadler, P. M., & Shanahan, M.-C. (2010). Connecting high school physics experiences, outcome expectations, physics identity, and physics career choice: A gender study. *Journal of Research in Science Teaching*, *47*(8), 978–1003. https://doi.org/10.1002/tea.20363

Hill, C., Corbett, C., & St. Rose, A. (2010). *Why so few? Women in science, technology, engineering, and mathematics*. Washington, DC: American Association of University Women.

Huberth, M., Chen, P., Tritz, J., & McKay, T. A. (2015). Computer-tailored student support in introductory physics. *PLOS ONE*, *10*(9), e0137001.

Jewell, R. T., & McPherson, M. A. (2012). Instructor-specific grade inflation: Incentives, gender, and ethnicity. *Social Science Quarterly*, *93*(1), 95–109. https://doi.org/10.1111/j.1540-6237.2011.00827.x

Johnson, A. C. (2007). Unintended consequences: How science professors discourage women of color. *Science Education*, *91*(5), 805–821. https://doi.org/10.1002/sce.20208

Johnson, H. J., Barnard-Brak, L., Saxon, T. F., & Johnson, M. K. (2012). An experimental study of the effects of stereotype threat and stereotype lift on men and women's performance in mathematics. *Journal of Experimental Education*, *80*(2), 137–149. https://doi.org/10.1080/00220973.2011.567312

Keiser, H. N., Sackett, P. R., Kuncel, N. R., & Brothen, T. (2016). Why women perform better in college than admission scores would predict: Exploring the roles of conscientiousness and course-taking patterns. *Journal of Applied Psychology*, *101*(4), 569–581.

King, B. (2015). Changing college majors: Does it happen more in STEM and do grades matter? *Journal of College Science Teaching*, *44*(3), 44–51.

Koester, B. P., Grom, G., & McKay, T. A. (2016). *Patterns of gendered performance difference in introductory STEM courses*. Retrieved from https://arxiv.org/abs/1608.07565

Koester, B. P., Fogel, J., Murdock III, W., Grom, G., & McKay, T. A. (2017). Building a transcript of the future. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 299–308). New York, NY: Association for Computing Machinery.

Kokkelenberg, E. C., Dillon, M., & Christy, S. M. (2008). The effects of class size on student grades at a public university. *Economics of Education Review*, *27*(2), 221–233. https://doi.org/10.1016/j.econedurev.2006.09.011

Kost, L. E., Pollock, S. J., & Finkelstein, N. D. (2009). Characterizing the gender gap in introductory physics. *Physical Review Special Topics-Physics Education Research*, *5*(1), 010101.

Kostal, J. W., Kuncel, N. R., & Sackett, P. R. (2016). Grade inflation marches on: Grade increases from the 1990s to 2000s. *Educational Measurement: Issues and Practice*, *35*(1), 11–20.

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

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

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.

Maltese, A. V., Melki, C. S., & Wiebke, H. L. (2014). The nature of experiences responsible for the generation and maintenance of interest in STEM. *Science Education*, *98*(6), 937–962. https://doi.org/10.1002/sce.21132

Maltese, A. V., & Tai, R. H. (2011). Pipeline persistence: Examining the association of educational experiences with earned degrees in STEM among U.S. students. *Science Education*, *95*(5), 877–907. https://doi.org/10.1002/sce.20441

Mann, A., & DiPrete, T. A. (2013). Trends in gender segregation in the choice of science and engineering majors. *Social Science Research*, *42*(6), 1519–1541.

Meyer, M. (1908). The grading of students. *Science*, *28*(712), 243–250.

Micari, M., Pazor, P., & Hartmann, M. J. Z. (2007). A matter of confidence: Gender differences in attitudes toward engaging in lab and course work in undergraduate engineering. *Journal of Women and Minorities in Science and Engineering*, *13*(3), 279–293.

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.

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

Moss-Racusin, C. A., Molenda, A. K., & Cramer, C. R. (2015). Can evidence impact attitudes? Public reactions to evidence of gender bias in STEM fields. *Psychology of Women Quarterly*, *39*(2), 194–209. https://doi.org/10.1177/0361684314565777

Murphy, M. C., Steele, C. M., & Gross, J. J. (2007). Signaling threat: How situational cues affect women in math, science, and engineering settings. *Psychological Science*, *18*(10), 879–885. https://doi.org/10.1111/j.1467-9280.2007.01995.x

Ost, B. (2010). The role of peers and grades in determining major persistence in the sciences. *Economics of Education Review*, *29*(6), 923–934. https://doi.org/10.1016/j.econedurev.2010.06.011

Paunesku, D., Walton, G. M., Romero, C., Smith, E. N., Yeager, D. S., & Dweck, C. S. (2015). Mind-set interventions are a scalable treatment for academic underachievement. *Psychological Science*, *26*(6), 784–793. https://doi.org/10.1177/0956797615571017

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.

President's Council of Advisors on Science and Technology. (2012). *Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics*. Retrieved from http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-engage-to-excel-final_2-25-12.pdf

Pulfrey, C., Buchs, C., & Butera, F. (2011). Why grades engender performance-avoidance goals: The mediating role of autonomous motivation. *Journal of Educational Psychology*, *103*(3), 683–700. https://doi.org/10.1037/a0023911

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.

Rask, K. (2010). Attrition in STEM fields at a liberal arts college: The importance of grades and pre-collegiate preferences. *Economics of Education Review*, *29*(6), 892–900. https://doi.org/10.1016/j.econedurev.2010.06.013

Rauschenberger, M. M., & Sweeder, R. D. (2010). Gender performance differences in biochemistry. *Biochemistry and Molecular Biology Education*, *38*(6), 380–384.

Rodger, S., Murray, H. G., & Cummings, A. L. (2007). Gender differences in cooperative learning with university students. *Alberta Journal of Educational Research*, *53*(2), 157–173.

Rojstaczer, S., & Healy, C. (2012). Where A is ordinary: The evolution of American college and university grading, 1940–2009. *Teachers College Record*, *114*(7).

Snyder, T. D., & Dillow, S. A. (2015). *Digest of education statistics 2013* (NCES 2015-011). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

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

Stump, G. S., Hilpert, J. C., Husman, J., Chung, W. T., & Kim, W. (2011). Collaborative learning in engineering students. *Journal of Engineering Education*, *100*(3).

Sue, D. W. (2010). *Microaggressions in everyday life: Race, gender, and sexual orientation*. Hoboken, NJ: Wiley.

Unkovic, C., Sen, M., & Quinn, K. M. (2016). Does encouragement matter in improving gender imbalances in technical fields? Evidence from a randomized controlled trial. *PLOS ONE*, *11*(4), e0151714.

Urry, M. (2015). Science and gender: Scientists must work harder on equality. *Nature*, *528*(7583), 471–473. https://doi.org/10.1038/528471a

Vogt, C. M., Hocevar, D., & Hagedorn, L. S. (2007). A social cognitive construct validation: Determining women's and men's success in engineering programs. *Journal of Higher Education*, *78*(3), 337–364. https://doi.org/10.1353/jhe.2007.0019

Walton, G. M., Logel, C., Peach, J. M., Spencer, S. J., & Zanna, M. P. (2015). Two brief interventions to mitigate a "chilly climate" transform women's experience, relationships, and achievement in engineering. *Journal of Educational Psychology*, *107*(2), 468–485. https://doi.org/10.1037/a0037461

Weaver, A. J., & Raptis, H. (2001). Gender differences in introductory atmospheric and oceanic science exams: Multiple choice versus constructed response questions. *Journal of Science Education and Technology*, *10*(2), 115–126. https://doi.org/10.1023/A:1009412929239

Wright, C. D., Eddy, S. L., Wenderoth, M. P., Abshire, E., Blankenbiller, M., & Brownell, S. E. (2016). Cognitive difficulty and format of exams predicts gender and socioeconomic gaps in exam performance of students in introductory biology courses. *CBE-Life Sciences Education*, *15*(2), ar23. https://doi.org/10.1187/cbe.15-12-0246

Wright, M. C., McKay, T., Hershock, C., Miller, K., & Tritz, J. (2014). Better than expected: Using learning analytics to promote student success in gateway science. *Change: The Magazine of Higher Learning*, *46*(1), 28–34. https://doi.org/10.1080/00091383.2014.867209

Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: They're not magic. *Review of Educational Research*, *81*(2), 267–301. https://doi.org/10.3102/0034654311405999

Authors

REBECCA L. MATZ is an academic specialist in the Hub for Innovation in Learning and Technology at Michigan State University; email: matz@msu.edu. Her research focuses on assessment and learning analytics in undergraduate science, technology, engineering, and mathematics education.

BENJAMIN P. KOESTER is a senior research associate in the Department of Physics at the University of Michigan; email: bkoester@umich.edu. His research focuses on adapting and applying algorithmic, quantitative, statistical, and mathematical methods of astrophysics, bioinformatics, and learning analytics.

STEFANO FIORINI is a lead research management analyst in the Office of Assessment and Research at Indiana University; email: sfiorini@indiana.edu. His research focuses on developing and implementing strategies to respond to campus information needs as well as collaborating with campus partners on new analytical initiatives.

GALINA GROM is a graduate student in the Department of Physics at the University of Michigan; email: grom@umich.edu. Her research focuses on using institutional information to assess dropout rates, student performance, and the efficacy of interventions designed to close performance gaps.

LINDA SHEPARD is the senior assistant vice provost for undergraduate education at Indiana University; email: lshepard@indiana.edu. Her research focuses on developing institutional resources to be used to inform campus policy, strategic initiatives, and program assessment.

CHARLES G. STANGOR is a professor of psychology at the University of Maryland; email: stangor@umd.edu. His research focuses on the enhancement and assessment of academic achievement in higher education, with a particular focus on reducing educational achievement gaps between men and women and among ethnic groups.

BRAD WEINER is a former analyst in the Office of the Vice Provost for Undergraduate Education at the University of Minnesota; email: bweiner@capturehighered.com. He is currently the director of data science at Capture Higher Ed, 2303 River Road, Suite 201, Louisville, KY 40206.

TIMOTHY A. MCKAY is an Arthur F. Thurnau Professor of Physics, Astronomy, and Education at the University of Michigan; email: tamckay@umich.edu. His research focuses on finding new ways to use data to assess teaching methods, support students, and improve learning.