Does Class Size Affect the Gender Gap? A Natural Experiment in Law

Daniel E. Ho and Mark G. Kelman

ABSTRACT

We study a unique natural experiment in which Stanford Law School randomly assigned first-year students to small or large sections of mandatory courses from 2001 to 2011. We provide evidence that assignment to small sections closed a slight (but substantively and highly statistically significant) gender gap existing in large sections from 2001 to 2008; that reforms in 2008 that modified the grading system and instituted small graded writing and simulation-intensive courses eliminated the gap entirely; and that women, if anything, outperformed men in small simulation-based courses. Our evidence suggests that pedagogical policy—particularly small class sizes—can reduce, and even reverse, achievement gaps in post-graduate education.

1. INTRODUCTION

Demographic achievement and test score gaps pose severe challenges to educational policy. Such gaps have been widely documented, from the

DANIEL E. HO is Professor of Law and Robert E. Paradise Faculty Fellow for Excellence in Teaching and Research, Stanford Law School. MARK G. KELMAN is James C. Gaither Professor of Law and Vice Dean, Stanford Law School. We thank Kristen Altenburger, Rebecca Morris, and Michael Morse for invaluable research assistance; Ian Ayres, Stuart Benjamin, Beth Colgan, Faye Deal, Stephen Galoob, Paul Goldstein, Gillian Hadfield, Amalia Kessler, Jon Klick, Jeff Kobrick, Larry Kramer, Liz Magill, Jenny Martinez, Jeannie Merino, Bernie Meyler, Alison Morantz, Deborah Rhode, Shirin Sinnar, George Triantis, Laura Trice, and participants at the 2013 Center for Law and Social Sciences Workshop at the University of Southern California School of Law, the 2014 faculty colloquium at the University of Colorado at Boulder Law School, the 2014 faculty workshop at Duke University School of Law, the 2014 faculty colloquium at Emory University School of Law, the 2014 faculty workshop at Stanford Law School, and the 2013 Conference on Empirical Legal Studies at the University of Pennsylvania Law School for helpful comments and conversations; Elizabeth Di Giovanni, Kim Borg, Nicole Cagampan, Kyung Chong, Faye Deal, Carol Ida, Anna O'Neill, Susan Robinson, and Mila Fernandez-Ronquillo for help

[*Journal of Legal Studies,* vol. 43 (June 2014)]

black-white test score gap (Jencks and Phillips 1998) to gender gaps in science, collegiate outcomes, and law and business schools (Xie and Shauman 2003; Jacobs 1996; Hancock 1999; Epstein 1993). Less understood is whether policies and pedagogical choices can reduce achievement gaps and, if so, how.

One promising intervention to reduce achievement gaps is to reduce class size. Having smaller classes may, for instance, enable teachers to better understand and teach students at different levels. Jencks and Phillips conclude that to narrow the gap, "[t]he two policies that . . . combine effectiveness with ease of implementation are cutting class size and screening out teachers with weak academic skills" (Jencks and Phillips 1998, p. 44). The best evidence comes from the Tennessee Student-Teacher Achievement Ratio (STAR) experiment, in which students in kindergarten through third grade were randomly assigned to large or small classrooms. Results suggest that assignment to smaller classrooms improved performance overall and reduced racial test score gaps (Ferguson 1998; Krueger 1999; Mosteller 1995). But these estimates are disputed. Hanushek (1999) argues that high attrition rates (with up to 50 percent of students leaving the experiment),[1] noncompliance (with 10 percent switching from large to small classrooms), and nonresponse (with 3–12 percent not taking exams) provide reasons to doubt the class size effects. Quasi-experimental and observational studies are less certain about the effect of smaller classes on achievement generally and on demographic gaps (see, for example, Fredriksson, Öckert, and Oosterbeek 2012; Hoxby 2000; Angrist and Lavy 1999; Fryer and Levitt 2004).

A separate literature, focusing on gender gaps, particularly in math and science, examines the role of competition and the gender of the instructor. Gneezy, Niederle, and Rustichini (2003) show that competition exacerbates gender differences in a maze-solving task. They randomly assign experimental subjects to compensation based on a tournament incentive, in which only the highest performer receives payment, or payment per task. The gender gap increases threefold in the competitive tournament condition (see also Niederle and Vesterlund 2007, 2010). Ors, Palomino, and Peyrache (2013) find that men outperform women on entrance exams to a top-ranked French business school, a finding that is reversed for less competitive finishing exams at the end

in collecting law school data; and the many faculty members responding to our inquiries about pedagogy.

[1] See Krueger (1999, table 1), which documents attrition rates from 47 to 53 percent for students entering the experiment in kindergarten or first grade.

of high school. Carrell, Page, and West (2010), in a study that is closest to ours in research design, study a natural experiment at the U.S. Air Force Academy, where students are randomly assigned to professors for mandatory courses. Having female professors greatly improves women's performance in math and science courses (see also Dee 2007).[2]

The gender gap in legal education has attracted a great deal of academic attention. Scholars argue that Socratic and adversarial teaching styles common in large law school classes disadvantage women (see, for example, Banks 1988; Guinier et al. 1994; Rhode 1993, 2001; Weiss and Melling 1988). Voluminous research confirms that women participate less frequently in the classroom, although some studies document relative parity in (or greater comfort by women with) small courses (Yale Law Women 2012; Banks 1988; Weiss and Melling 1988, pp. 1334–35). Because law school grades matter considerably in the legal profession, numerous scholars examine the gender gap in law school grades, with heterogeneous findings across schools.[3] Guinier et al. (1994, p. 96) advocate comprehensive reform to address gender disparities, emphasizing that "small class size may be a necessary condition," a common refrain in calls for reform. But while much ink has been spilled describing gender differences, few studies—and none applying experimental methods—systematically assess what pedagogical policies might mitigate the gender gap in law school performance.

Our article marries these literatures by examining whether having smaller classes reduce gender gaps in performance. We study a unique setting in which Stanford Law School randomly assigned students to small or large sections of mandatory first-year courses from 2001 to 2011. We collect rich individual-level covariate and grade information for every student in every mandatory first-year course to study whether assignment to small sections reduces the gender gap in law school performance. We find that they do.

Our study has several virtues. First, unlike observational studies, in

[2] In examining our data, we do not find that gender of the instructor has an effect on the gender gap or that the class size effect is explained by the gender of the instructor.

[3] Kay and Gorman (2008, p. 302) observe, "Studies have offered conflicting evidence as to whether there is a gender difference in law school grades." Clydesdale (2004) finds no gender difference in first-year grade point averages (GPAs); Wightman (1996) finds a slight gender gap in first-year GPAs; Guinier et al. (1994) find a gender gap in first-year GPAs at the University of Pennsylvania; Bowers (2000) finds a gender gap in first-year GPAs at the University of Texas; Homer and Schwartz (1989) find a gender gap in Contracts and Property courses at the University of California, Berkeley; and Taber et al. (1988) find no gender gap in membership in the Order of the Coif at Stanford Law School.

which class size is often confounded (for example, by type of student), we leverage Stanford's randomization of mandatory first-year courses. To our knowledge, virtually no studies capitalize on random assignment to focus specifically on the effect of class size on gender gaps in academic achievement.[4] In addition, because we observe all information that the Office of Admissions takes into account when assigning students to sections, treatment assignment would be unconfounded even without randomization (Barnow, Cain, and Goldberger 1980; Ho and Rubin 2011; Rubin 2008). Second, because large sections are composites of small sections, we observe how the same students perform in small versus large sections across gender lines. Applying a difference-in-differences design to our data allows us to control for all student-fixed attributes (most important, ability) to identify the effect of small classes by gender.

Third, our study has advantages even relative to other experimental approaches. In the Tennessee STAR experiment, for instance, some 60 percent of students leave or transfer out of their assigned classrooms.[5] In contrast, in our study, all students remain in the class as assigned; no students drop out, course section assignments are mandatory, and all students sit for the final exam. Fourth, Stanford's assignment and grouping were conducted to maximize representativeness across sections, not with any evaluation of class size in mind. Hawthorne effects, whereby instructors modify teaching because of the experiment, are thereby impossible. Last, while many have conjectured that class size effects vary at different levels of education, prior work focuses overwhelmingly on early education,[6] despite mounting evidence of achievement gaps in higher education. Our study contributes to the literature by providing one of the first examinations of class size effects in a postgraduate professional school setting.

This article proceeds as follows. Section 2 discusses the unique natural experiment that Stanford inadvertently conducted from 2001 to 2012. Section 3 describes the fine-grained student and course data we collected with the help of the law school's admissions and registrar offices. Section 4 verifies random assignment to sections by assessing balance along a host of covariates. Section 5 examines the effects of class size on the

[4] The studies that come closest to doing so are De Paola, Ponzo, and Scoppa (2011), Krueger (1999), and Fredriksson, Öckert, and Oosterbeek (2012).

[5] See Krueger (1999, table 1) on attrition rates and Hanushek (1999) for a discussion of attrition and failure to sit for exams.

[6] But see Monks and Schmidt (2010, p. 1), who note that "[o]nly a handful of studies have [examined] class size . . . in tertiary education."

gender gap from 2001 to 2008, when the school employed numerical grade point averages (GPAs). Applying a difference-in-differences approach, we show that assignment to small sections eliminates a small but highly statistically significant gender gap that exists in large sections. Section 6 examines the evidence after the educational reforms of 2008, which changed the grading system to an honors/pass (H/P) basis and instituted small graded writing and simulation-intensive courses. We show that the gender gap vanishes under this new system and rule out the possibility that this is solely due to the coarseness of the grading system. If anything, women systematically outperform men in simulation-based courses, which have even fewer students than small sections. Section 7 concludes.

2. THE STANFORD EXPERIMENT

Stanford's first-year curriculum provides a compelling natural experiment because the school randomly assigned small sections of students to specific courses. In addition to randomly matching sections to courses, the school sought to make each small section representative of the entering class as a whole, adopting what is best characterized as a form of (stratified) block randomization to group students into sections. Unlike in other educational settings, students had no choice of course enrollment. Students' enrollment choices (for example, in elective courses beyond the first year) would otherwise confound estimates of the effect of class size. We first discuss the role of small sections in Stanford's mandatory first-year curriculum and then detail the precise mechanisms of grouping students into sections and assigning sections to courses.

2.1. The First-Year Curriculum

From fall 2001 to spring 2008, Stanford's mandatory first-year curriculum consisted of six core doctrinal courses (Civil Procedure, Constitutional Law, Contracts, Criminal Law, Property, and Torts) and one writing course (Legal Research and Writing [LRW]). Doctrinal courses were graded on a numerical 4.0 GPA scale ranging from 2.1 to 4.3, with a mean requirement of 3.4 in a course. Legal Research and Writing courses were graded on a mandatory-credit/restricted-credit/no-credit basis. In other courses, students could elect to be graded under the so-called 3K grading system (a system where credit, restricted credit, and no credit were the grading options), with the 3.4 mean GPA requirement applied regardless of the grading option chosen. The 3K system was effectively a pass/fail system.

Beginning in fall 2008, the law school instituted a series of pedagogical reforms. First, courses would be graded using the H/P system. The required range was 30–40 percent honors for doctrinal courses. The rationales for grade reform were to reduce grade curve shopping and to eliminate what was perceived to be a falsely precise and, to many students, intimidating, numerical GPA system (Guess 2008; Kerr 2008). As part of grade reform, students would no longer be able to elect the 3K option.

Second, the law school transitioned from a semester to a quarter system in fall 2009, keeping the first-year curriculum largely unchanged. Mandatory courses in the fall term continued to meet for the same duration as previously. Winter courses were adjusted to the quarter system. Two modifications were that LRW was graded and shortened to the fall term, and the school introduced an even smaller, two-quarter, simulation-based Federal Litigation course in lieu of LRW in the winter and spring terms. The case used in Federal Litigation involved First Amendment, personal jurisdiction, and class certification issues. Students were assigned to specific sides and sets of issues and were given a wide range of writing and simulation exercises (initially, drafting a complaint, three briefs, and a bench memo; delivering and judging oral arguments; and taking and defending a deposition). The required range in LRW and Federal Litigation was 35–50 percent honors.

Throughout the observation period, the entering class, ranging from 166 to 180 students, was split into six small sections of up to 30 students. In addition to LRW, one fall doctrinal course was taught exclusively to the small section. The substantive field (for example, Contracts or Criminal Law) varied both within and across entering classes, largely on the basis of faculty availability. Other doctrinal courses were typically taught in a large class combining two small sections (roughly 60 students). When Federal Litigation was introduced, small sections were split into groups of roughly 18 students (10 sections per incoming class), which were further divided into legal teams of four or five students each. Depending on the instructor, Federal Litigation class meetings were often held exclusively between the instructor and a legal team. At all times, exams in doctrinal courses, on which final grades are overwhelmingly based, were graded blindly, ruling out the possibility of patent instructor grading bias.[7]

[7] Blind grading may not rule out the possibility that instructors may devalue the female voice (Gilligan 1982) on exams.

2.2. Grouping and Assignment Mechanisms

To understand the mechanism by which students were assigned to small sections, we detail two decisions: grouping students into small sections and assigning small sections to classes. These decisions were made to ensure fairness in and representativeness (or balance) across section assignments, not to study class size effects.

Students were grouped into small sections as follows. First, after finalizing most of the entering class, the associate dean of admissions sorted the list of entering students by academic index (a function of the Law School Admission Test [LSAT] score and undergraduate GPA), assigning numbers 1–6 to each student. To balance the academic index but retain the simplicity of assignment, the dean systematically cycled through the numbers 1–6 (first in order and then in reverse order), going down the list of sorted names: for example, 1, 2, 3, 4, 5, 6, 6, 5, 4, 3, 2, 1, and so on. The academic index among Stanford students is coarse because of range compression: for instance, the class of 2005 had only seven unique values of the academic index, and the order within a stratum of an index value was random. Second, the associate dean made a series of adjustments to balance gender and ethnicity across sections while retaining parity in terms of LSAT scores, advanced degrees, and undergraduate institutions.

Assigning the six sections to instructors and courses was random. Because the associate dean was unaware of how the six numbers mapped onto specific courses and instructors, she could not match students on the basis of instructor fit or predicted ability to succeed in a particular small or large section. Students' characteristics were not considered in assigning sections to courses, except in very rare circumstances.[8]

Grouping students into sections, as Appendix A shows, is best characterized as approximating a form of stratified block randomization (Box, Hunter, and Hunter 2005). The emphasis on balancing gender and ethnicity is akin to stratifying on these variables, increasing, if anything, the efficiency of analysis. The order of students in the list is stochastic, as matriculation decisions for specific students can hinge on factors of chance (for example, deferrals of admission). It is very unlikely that the student list thereby has a (periodic) relationship (for example, every 12th student has a low-income background) that would confound the group-

[8] These involved conflicts of interest (for example, when a faculty member was related to a student), which were exceedingly rare, and grouping of sections remained intact.

Table 1. Summary Statistics for the Sample

			Grades			
Period	Students	Instructors	All	GPA	H/P	Mean
2001–8	1,193	62	9,539	5,600		3.46
2008–11	704	58	6,150		6,141	.42

Note. Numerical grade point averages (GPAs) were used 2001–8; the honors/pass (H/P) system was instituted in fall 2008. "All" grades include courses graded either under the 3K system (an option with three grades: credit, restricted credit, and no credit) or on a mandatory-credit basis.

ing into sections. Gender, ethnicity, the academic index, and other covariates are, by construction, balanced across sections.

While there are strong reasons, based on institutional knowledge of the assignment mechanism, to believe that the school randomized students into small sections, Section 4 verifies empirically that small sections were balanced along all covariates.

3. DATA

We compile data from the Office of Admissions on first-year students and match these to data from the Office of the Registrar on grades awarded to each student in a course. Our primary data consist of 15,689 grades assigned by 91 instructors to 1,897 students in mandatory first-year courses from 2001 to 2012. Table 1 provides a breakdown of the raw data for the two observation periods under the GPA system (2001–8) and the H/P system (2008–11). Prior to 2008, the overall mean GPA was 3.46, which is higher than the mandatory mean of 3.4 because of students electing the 3K option. (Instructors graded all exams collectively, without knowledge of the students' choice of grading option.) The overall proportion of honors was .42, which exceeds 40 percent because LRW and Federal Litigation are subject to a 50 percent cap on honors.

Table 2 reports summary statistics of incoming credentials and demographics by gender. The two most crucial covariates are LSAT score and undergraduate degree, which are comparable for men and women. Women differ in other respects, however: they are nearly a year younger and more likely to self-identify as members of minority groups (for example, 15 percent of women are Asian American, compared with 8 percent of men). These differences along observables are important in understanding the gender gap and class size effects; all the model-based estimates we present control for ethnicity or student fixed effects. The

Table 2. Demographic Covariates at Time of Matriculation

	Mean		Pooled
	Men	Women	SD
Academic background:			
LSAT score	169	168.9	4.2
Undergraduate GPA	3.81	3.82	.19
LSAC index	3.42	3.41	.15
Master's degree	.18	.12	.36
Ph.D. degree	.05	.03	.19
Demographic background:			
Age	24.6	23.8	2.8
White	.59	.51	.50
Latino	.12	.12	.32
Asian American	.08	.15	.32
African American	.08	.11	.29
Undergraduate institution:			
Stanford	.10	.11	.30
Harvard	.06	.07	.25
Yale	.07	.07	.25
Berkeley	.03	.04	.18

Note. LSAT = law school admission test; GPA = grade point average; LSAC = Law School Admission Council.

histograms in Figure 1 plot the raw distribution of grades assigned in individual courses by gender. The gray histogram plots the grade distribution for men, and the black outline plots the grade distribution for women. The figure shows that there is a small but persistent gender gap. On average, women earn grades that are .05 GPA points lower than those for men ($p < .0001$). The gap persists and remains highly statistically significant when controlling for the full set of covariates (LSAT score, undergraduate GPA, academic index, age, ethnicity, master's degree, doctoral degree, professional degree, and fixed effects for undergraduate institution, instructors, and courses).[9] Slight demographic differences therefore do not account for the gender gap. Although obvious, it is worth noting that the variation within gender far exceeds that across genders—despite the gap, individual women and men perform along the entire range of GPAs.

Although the gender gap is small in absolute magnitude, the gap represents roughly 15 percent of the pooled GPA standard deviation—in a profession that prizes law school performance (Henderson 2003). To

[9] Because of substantial overlap between the characteristics of men and women at entrance, the gender gap persists when preprocessing via matching to reduce the degree of extrapolation (Ho et al. 2007).

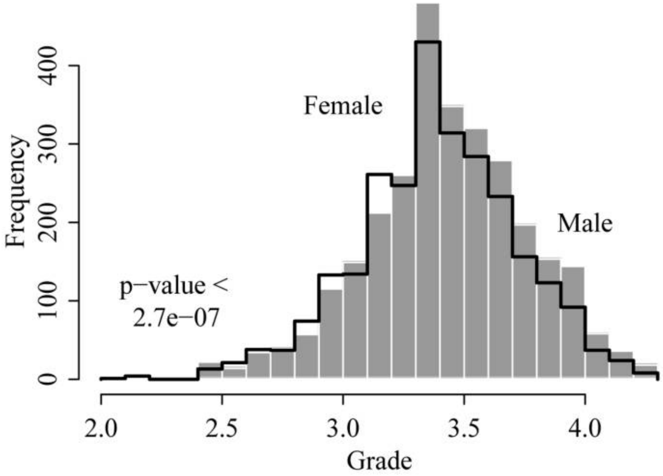

Figure 1. Raw gender gap in grades

illustrate the gap's substantive importance, we examine data on 487 clerkship applications by Stanford students from 2003 to 2008 and data from 2,949 on-campus interviews in fall 2008, the predominant process for securing private-sector jobs. Grades and clerkship placements are highly correlated: an increase in GPA from 3.6 to 3.65 is associated with a 7 percent (statistically significant) increase in the probability of securing a federal appellate clerkship.[10] Similarly, grades have a strong positive correlation with the rate at which students are offered callback interviews: an increase in GPA from 3.25 to 3.3 is associated with a nearly 5 percent increase in the callback rate.[11] It is worth noting that law firms appear to have become even more grade sensitive since 2008 (Bell 2008). The callback rate from 2008 may thereby understate the effect of grades on the current labor market. In short, while small in absolute magnitude, the .05 GPA gender gap matters.

[10] We estimate this correlation using logistic regression with placement in a federal appellate clerkship as the outcome and GPA at the time of application as the explanatory variable, conditional on applying for an appellate clerkship.

[11] We estimate this correlation using a local polynomial (loess) model. There is no evidence that the association between first-year GPA and callback rates differs between men and women. On-campus interviews are scheduled via a lottery preventing employers from observing law school transcripts, so grades manifest themselves primarily in the rate of callback interviews.

4. RANDOMIZATION CHECKS

Although there are strong reasons to believe that the assignment of sections to courses (and section grouping) was random, we perform a series of randomization checks to test for violations. As large sections are composites of small sections, we check for whether the six small sections in any year of admission exhibit imbalance on key covariates. Figure 2 plots the year of admission against 12 covariates. Each black dot represents the mean (or proportion) for one of six small sections in an entering class; entering classes are separated by vertical lines. The gray shading represents the (simulated) 95 percent confidence intervals assuming randomization, calculated by 1,000 Monte Carlo simulations. Under randomization, the observed mean (or proportion) should generally fall within the intervals. Nearly all do.

Figure 2 also reveals that the associate dean's additional demographic shuffling balances gender and ethnicity beyond what would be expected by chance. The observed proportions of women and minorities are closer to the class mean than would be the case under pure randomization. Other covariates approximate the randomization distribution. Although some sections fall outside of the 95 percent interval, the rate is much lower than type I error rates: under randomization, we would expect roughly 40 such deviations [= .05 α level × 6 sections per entering class × 11 entering classes × 12 covariates]. In short, the results strongly confirm that small sections were effectively randomized. In Appendix A, we show that the process is essentially a form of (stratified) block randomization, thereby improving balance on gender and ethnicity beyond pure randomization. Indeed, the associate dean was gladly willing to substitute a formal stratified block randomization algorithm that essentially replicated her manual assignment to sections.

5. CLASS SIZE EFFECTS, 2001–8

We now focus on assessing the causal effect of class size during the time of the GPA system (2001–8). Because of the number of changes—particularly in grading—Section 6 examines the post-2008 period separately.

Figure 3 presents quantile-quantile plots comparing the raw grade distributions for men and women conditional on section size, with dots randomly jittered for visibility. In the absence of a gender gap, the dots should line up along the 45 degree line. The left panel shows that men and women perform similarly in small sections, while the right panel exhibits the gender gap. On average, men earn GPAs that are .05 point

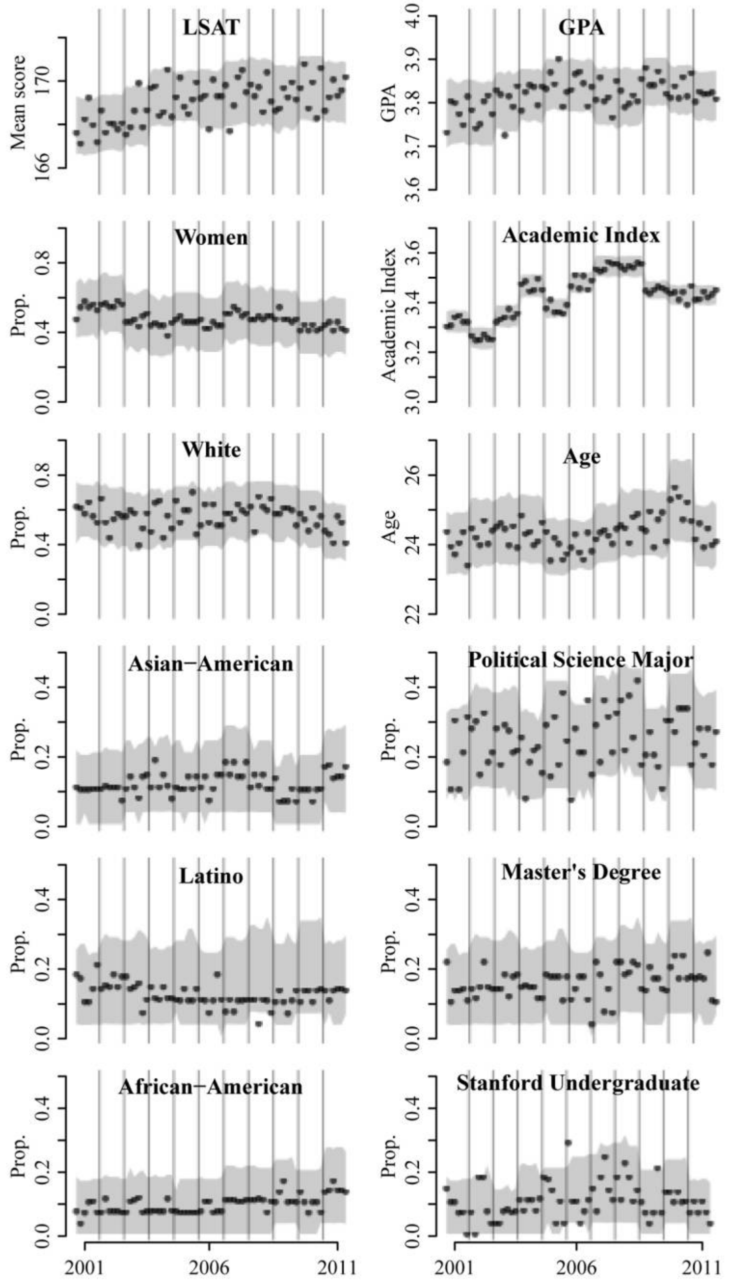

Figure 2. Randomization checks for small sections

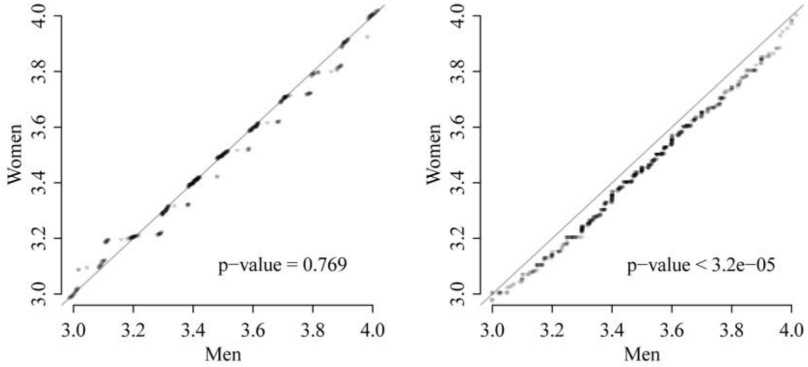

Figure 3. Quantile-quantile plots comparing performance of men and women in small (*left*) and large (*right*) sections.

higher than those of women in large sections ($p < .01$). Table 3 provides summary statistics for the differences in means between men and women across large and small classes. Calculation of the raw difference in differences ($p < .05$) shows that women tend to outperform men by .05 GPA point in small sections relative to large sections.

To more rigorously assess the class size and gender effect, we use a difference-in-differences identification strategy. We estimate the following equation:

$$E(Y_{s,i,c}) = \tau T_{s,i,c} G_s + \lambda T_{s,i,c} + \alpha_s + \eta_i + \kappa_c,$$

where s indexes students, i indexes instructors, and c indexes course subjects, $Y_{s,i,c}$ represents the numerical grade earned by student s in course c taught by instructor i, $T_{s,i,c}$ equals one if the student was enrolled in the treatment of a small section and zero if not, and G_s equals one if the gender of student s is female and zero if male. Standard errors are clustered by course section. The parameters α_s, η_i, and κ_c are student, instructor, and course fixed effects capturing, respectively, any student-specific, course-invariant effects (chiefly, ability); instructor-specific, course-invariant effects; and course-specific effects.

By construction, student fixed effects (α) control for gender, age, LSAT score, undergraduate GPA, and any other student-specific characteristics present upon entering law school. The parameter of interest (τ) is identified by changes in the performance of female students across small and large sections relative to male students across small and large

Table 3. Raw Grade Averages in Large and Small Sections, by Gender

	Large	Small	Small − Large
Men	3.488	3.461	−.026
Women	3.433	3.454	.021
Men − women	.054*	.007	.047*,[a]

Note. Gender difference is conditional on class size. Class size difference is conditional on gender.

[a] Difference-in-differences value.

*$p < .05$.

sections. This formalizes the hypothesis that smaller class sizes may have differential effects on performance by gender. Because of random assignment, the identification assumption is credibly met: it is very unlikely that there are exogenous factors that are unique to female students specific to small sections. Appendix B discusses two highly implausible mechanisms that would confound treatment assignment. Absent grading elections, we would not expect instructor- and course-specific deviations from the mandatory mean. We nonetheless include instructor and course fixed effects in the saturated model because students who choose the 3K grading option can cause courses to deviate from the 3.4 mean.

Table 4 presents results. The simplest estimates are reported with fixed effects for ethnicity. The gender gap decreases slightly to .039 GPA point but is reversed entirely in small sections. Student, instructor, and course fixed effects are added sequentially. Model estimates remain stable: while assignment to small sections causes women to improve performance by .04 GPA point, it diminishes men's performance by .03 GPA point. These results provide considerable evidence that being in small classes diminishes the gender gap existing for women in large sections.

Appendix C investigates the possibility that grading elections bias our estimates. If more women relative to men, for example, exercise the 3K grading option in small sections, observed grades achieved by women may be inflated in small sections solely because lower-performing women remain ungraded on the GPA scale. Because class size does not appear to have a substantial effect on students' grading elections (affecting at most one or two students per small section), graded students remain statistically indistinguishable in covariates across small and large sections, and the difference-in-differences approach identifies the effect solely on the basis of students electing to be graded in both small and large sections, grading elections do not appear to threaten our findings.

Table 4. Effect of Class Size: Difference-in-Differences Estimates

	A	B	C	D
Small Section × Female (τ)	.045*	.044*	.042*	.041*
	(.023)	(.018)	(.019)	(.019)
Small Section (λ)	−.027$^+$	−.024$^+$	−.029*	−.032*
	(.015)	(.012)	(.014)	(.014)
Female	−.038**			
	(.008)			
Ethnicity fixed effects	Yes	No	No	No
Student fixed effects (α)	No	Yes	Yes	Yes
Instructor fixed effects (η)	No	No	Yes	Yes
Course fixed effects (κ)	No	No	No	Yes
Parameters	10	1,184	1,222	1,227
R^2	.10	.52	.53	.53

Note. Standard errors, clustered by course section, are in parentheses. $N = 5{,}600$.

$^+p < .1$.

$^*p < .05$.

$^{**}p < .01$.

6. THE VANISHING GAP, 2008–11

We now examine the gender gap and class size effects after the pedagogical reforms instituted in 2008. Table 5 reports the proportion of honors earned by men and women. The gender gap vanishes under the H/P system. Women earn honors in roughly 42 percent of courses, compared with 41 percent of courses for men. Women also continue to perform slightly better than men in small sections. The effect, however, appears to be entirely driven by Federal Litigation.

To investigate this, we apply a similar difference-in-differences strategy to test for small-section effects and Federal Litigation. Table 6 reports logistic regression estimates comparable to those in Table 4. Model A confirms that the gender gap disappears. Women systematically earn more honors in Federal Litigation, a result robust to the full set of fixed effects. Relative to being in a large section, being in Federal Litigation increases women's probability of earning honors by .18, compared with only .08 for men. The differential grading guideline increases the probability of honors but does so disproportionately for women.

Why did the gender gap disappear? As the gender gap disappeared only over time (and is not induced by a randomized intervention), it is difficult to assess precisely what caused it to vanish. We can, however, rule out several explanations. First, it is not the case that grade reform, by dichotomizing grades into honors and pass, masked an underlying gender difference. To show this, we calculate shadow honors under the

Table 5. Gender Differences under the Honors/Pass Grading System: Proportion of Honors

	All Sections	Large Sections	Small Sections	Federal Litigation
Men	.414	.386	.452	.471
Women	.417	.367	.483	.545
Women − men	.002	−.018	.030	.074*

Note. Small sections include Federal Litigation and Legal Research and Writing, which are subject to a grading guideline of 35–50 percent honors.
*$p < .05$.

last 4 years of the GPA system, employing comparable grading guidelines of no more than 40 percent honors. Our model of these shadow honors strongly rejects the null hypothesis of no gender differences under the GPA system ($p < .001$). Intuitively, this can be seen from Figure 1, which shows that the small gap manifests itself along the entire range of the distribution.

Second, the relative qualifications of entering women and men did not change in any material way around 2008. Academic qualifications for men and women, such as LSAT scores and undergraduate GPAs, were comparable over the entire observation period and were smooth before and after 2008. Third, closing the gender gap was also not likely due to a spillover effect from Federal Litigation. Federal Litigation began only in the winter quarter, and our evidence suggests that the gender gap diminished even during the fall quarter. Last, because the transition from the semester to the quarter system left the first-term mandatory first-year courses largely intact, it is also unlikely that the change in the academic calendar eliminated the gender gap.

One explanation for the vanishing gender gap appears more plausible. The H/P system may have removed, at least subjectively, a degree of competitiveness from first-year exams. Recall that one of the predominant assumptions about the H/P system is that it reduces the pressure, and critiques of legal education often focus on gender dimensions of competition in the first year. Our findings are thereby consistent with laboratory experiments demonstrating that increasing the degree of competitiveness can greatly exacerbate gender gaps (Gneezy, Niederle, and Rustichini 2003; Niederle and Vesterlund 2010). Bloodgood et al. (2009) similarly find that when the University of Virginia medical school changed from letter to pass/fail grading, women disproportionately ex-

Table 6. Effects of Small Sections and Federal Litigation: Difference-in-Differences Estimates from Logistic Regression

	A	B	C	D
Federal Litigation × Female	.48*	.62**	.62**	.62**
	(.19)	(.24)	(.24)	(.24)
Federal Litigation	.39**	.58*	−.26	−.26
	(.08)	(.13)	(.16)	(.16)
Small Section × Female	−.03	−.05	−.05	−.05
	(.15)	(.16)	(.16)	(.16)
Female	−.00			
	(.07)			
Ethnicity fixed effects	Yes	No	No	No
Student fixed effects	No	Yes	Yes	Yes
Instructor fixed effects	No	No	Yes	Yes
Course fixed effects	No	No	No	Yes
Parameters	14	710	767	772
Residual deviance	7,834	5,375	5,340	5,339

Note. Models A and B include Legal Research and Writing (LRW) and LRW × female fixed effects because of the different grading guideline; as LRW instructors are unique, these are not estimated in models C and D (with instructor fixed effects). Standard errors, clustered by course section, are in parentheses. $N = 6{,}141$.

$^{*}p < .05$.

$^{**}p < .01$.

hibited gains in psychological well-being.[12] Robins et al. (1995) find that pass/fail grading at the University of Michigan Medical School reduced anxiety without a reduction in performance.

Our finding for Federal Litigation provides more insight into specific pedagogical techniques that potentially affect the gender gap. What distinguishes Federal Litigation from other doctrinal courses (and, to a lesser extent, LRW) is that it is based entirely on simulation, assigning students an affirmative litigation position in a real case; has no final exam under timed conditions; provides substantial feedback throughout the class; and is the smallest mandatory first-year class, with effectively only four or five students for many class meetings.[13] Each of these pedagogical features may affect the gender gap (see Rhode [1993] on simulation and feedback and Miller and Mitchell [1994] on timed exams). The scope of simulation-intensive exercises is simply not possible in large

[12] In their setting, grade reform did not appear to affect performance. One methodological challenge to the study is that there is some evidence that grade reform affected enrollment decisions along gender lines.

[13] The Federal Litigation effect does not appear to stem from the gender of the instructor. We cannot reject the null hypothesis that the effect is the same across male and female instructors.

classes. Importantly, there is no evidence of the gender gap reversal in LRW courses, which share the same set of core instructors.[14] This strongly suggests that distinct pedagogical techniques available in small classes matter.

7. CONCLUSION

Our findings suggest that class size and pedagogical policy have a considerable role to play in addressing gender gaps in professional school. Much work remains to be done in understanding the precise mechanisms by which class size and pedagogy differentially affect students. To develop a sense of the mechanisms, we surveyed each of the instructors who taught first-year courses at the law school from 2001 to 2008 (with all but one instructor responding) and consulted final exams, syllabi, and course evaluations whenever available. We collected information on exam type (for example, open versus closed book, duration), class participation (for example, pure cold call, panel system), assignments, use of formal simulation techniques, practice exams, and teaching assistants. The one pronounced difference was in formally administering practice exams: 45 percent of small sections had practice exams with model answers and/or class discussion, compared with 14 percent of large sections ($p = .001$), and 29 percent of small sections had practice examinations with grades and/or individualized feedback, compared with 7 percent of large sections ($p = .001$). The primary reason for this difference is practical: unlike in other divisions of the university, nearly all grading is done by law faculty, and fewer than one-fifth of courses employ teaching assistants, which makes feedback and grading of practice exams more difficult in large sections.

As mentioned, Federal Litigation provides more suggestive evidence on the mechanism. The course is heavily simulation based, with extensive feedback throughout the two quarters: students are assigned real advocacy roles with discrete issues in an actual case involving compelling issues. The extensive interactive exercises (for example, multiple oral arguments) are infeasible in a larger class. Consistent with the evidence that women express and exhibit preferences for direct representation and clinical education (Guinier et al. 1994, pp. 39–40; Weiss and Melling 1988, pp. 1317–48), the simulation structure of Federal Litigation may

[14] The finding that women outperform men is statistically indistinguishable across instructors teaching Legal Research and Writing and Federal Litigation and instructors teaching only Federal Litigation. This rules out the possibility that the Federal Litigation effect is driven by instructors' gender bias due to familiarity with the students.

be the mechanism reversing the gender gap. This evidence is consistent with studies suggesting that interactive engagement techniques can reduce the science gender gap (Lorenzo, Crouch, and Mazur 2006; Rosser 1995).[15]

We conclude with caveats on interpretation. First, randomization of students to small sections is a principal strength of our design; instructors are not necessarily randomly assigned. Instead, some deference is paid to instructors' preferences about section size; instructors with small-section preferences may simply teach differently. That does not invalidate estimates of the effect of these small sections on the gender gap, but it may mean that shifting large-section instructors to small sections may not automatically close the gender gap.[16]

Second, because Stanford employs norm-referenced grading (that is, the GPA ranks students only relative to one another, not based on an external criterion), our study does not permit us to directly assess the effects of class size on absolute degrees of learning. Criterion-referenced grading would be the obvious, but likely infeasible, way forward. A less ideal approach would be to regrade exams covering the same subject matter from different sections based on an absolute standard, but such test equating is challenging when exams may test for instructor- and section-specific knowledge.

Third, while our study provides well-identified quantities for matriculated Stanford students, the effects may not readily generalize to other schools. Our study nonetheless paves a path for additional research. As Mosteller (1999, p. 125) concludes, because of the dearth of randomized controlled trials of pedagogy, "in the last 100 years, education has not made much progress in evaluating processes of education." Yet many other schools have comparable concerns of fairness in assigning students to teachers, providing plausible settings by which to deploy a form of randomization and assess effects of pedagogy and class size. Fourth, our study cannot address whether small classes ultimately benefit a student's legal career beyond law school. Some may argue, for instance, that the Socratic method better prepares students for legal practice (Areeda 1996).

[15] But compare Pollock, Finkelstein, and Kost (2007), who are unable to replicate the interactive engagement findings in a setting with classroom sizes three times those of Lorenzo, Crouch, and Mazur (2006).

[16] We cannot reject the hypotheses that gender effects are the same for instructors teaching both small and large sections and instructors teaching exclusively small or large sections.

Last, the ultimate policy choice involves a more complex tradeoff between the benefits of reduced class size and the costs of staffing and classroom resources. The typical size of a first-year section across 201 American Bar Association–accredited law schools in 2013 was 66 students (SD = 17). The vast majority of law schools enroll sections that are far larger than those at Stanford. Instituting small sections as in our experiment may hence demand considerable resources. On the other hand, Stanford managed to create Federal Litigation without substantial additional cost by shifting existing instructors to create Federal Litigation sections, which suggests that not all reductions in class size need be a drain on resources.

In sum, our study demonstrates that reducing class size can eliminate, and even reverse, the gender gap in professional schools. Our findings also suggest that the gender gap may be highly contextual, depending on (and possibly induced by competitive pressure of) the grading system. This might explain the cacophony of findings about the existence of gender gap across law schools. The key now is how to address a gap when it does exist. And pedagogy may have a crucial role to play.

APPENDIX A: STRATIFIED BLOCK RANDOMIZATION

The law school's consideration of demographic factors in section assignments results in a balance of gender and ethnicity beyond what would be expected under pure randomization. Here we show that grouping students into six sections approximates a form of (stratified) block randomization (Box, Hunter, and Hunter 2005; Kernan et al. 1999). For simplicity of exposition, we focus on the entering class in 2006. For reference, we again calculate the distribution of means of 12 covariates under pure randomization, using 1,000 Monte Carlo simulations. We similarly simulate the distribution of covariates under block randomization. Within each simulation, we form six strata of unique combinations of gender and minority group (Asian American, Latino, and African American). Within each stratum, we apply block randomization, assigning section numbers 1–6 randomly without replacement in the stratum. This guarantees that sections will have equal numbers of women and minorities, with the only small imbalance stemming from strata with fewer students than sections. Figure A1 plots the results. The dark dashes along the x-axes indicate observed means of the covariate across six small sections. The black outlined histogram plots the pure randomization distribution, and the gray histogram plots the block randomi-

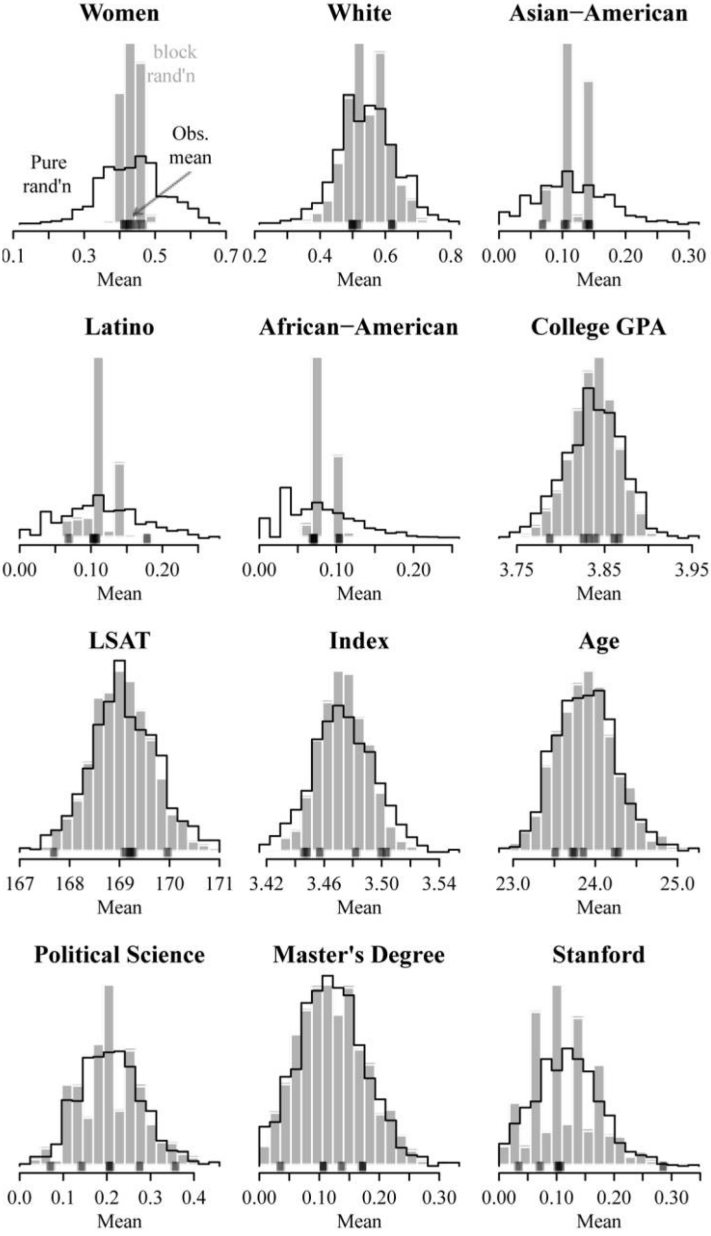

Figure A1. Pure and stratified block randomization for the incoming class of 2006

Table B1. Hypothetical Section-Assignment-Violating Identification Strategy

	Small Section				Large Section	
	A		B			
	H/N	%	H/N	%	H/N	%
High-achieving women/all women	3/4	.75	8/20	.40	11/24	.46
High-achieving men/all men	15/20	.75	6/15	.40	21/35	.60

Note. Statistics are conditional on gender. Percentages are the proportion of high-achieving individuals in the class. H/N = the number of high-achieving individuals divided by the number of all individuals in that class.

zation distribution. The dark dashes track the latter extraordinarily well, showing that the associate dean's grouping is comparable to block randomization.

APPENDIX B: SECTION-ASSIGNMENT-VIOLATING IDENTIFICATION ASSUMPTIONS

Even were assignment nonrandom, confounded assignment mechanisms that would artificially generate the gender effects would be difficult to conjure. Applying differences-in-differences estimation when large sections are simply composites of small sections (including the same set of students) rules out many simple manipulations. Two assignment mechanisms would violate our identification assumptions. One possibility is that the associate dean observes information about students that is course and/or instructor specific and then disproportionately assigns female students who are predicted to perform well to the cognate small section. Because the associate dean does not take into account any information on how the numbers 1–6 map to particular courses, this assignment mechanism can be easily ruled out on substantive grounds.

Another possibility is that sections are reverse stratified on ability by gender. For simplicity, imagine that there are only two sections and that students are either high achieving or low achieving. If one section combines a small number of women with a large number of men while another combines a large number of women with a small number of men, all while keeping the relative proportion of high-achieving students constant within a section, that could artificially generate our findings. Consider the hypothetical section assignments in Table B1.

The first column indicates that, in small section A, three of four women and 15 of 20 men are high achieving. These data reveal no gender gaps in small sections (with equivalent grade distributions across sections under norm-referenced grading) but a large gender gap in the consolidated large section. This form of section assignment, however, is emphatically not what the law school practices. To the contrary, small sections are designed to be as representative of the incoming class as possible, including by gender and ability.

APPENDIX C: SENSITIVITY TO NONRESPONSE

The grading election by students can be viewed as a kind of nonresponse: we are unable to observe grades for students who elect to take a course on a credit/no-credit basis. (As the H/P system eliminates grading elections, nonresponse poses no problem for the 2008–11 results.) Even when treatment is randomized, nonrandom nonresponse threatens the validity of estimates for two reasons (see Horiuchi, Imai, and Taniguchi 2007). First, nonresponse can invalidate the randomization. Among respondents (that is, students taking the course for a grade), treated individuals may be quite different from individuals in the control group. Second, nonresponse affects the target population, as we may no longer be able to estimate the average treatment effect of class size on the population of matriculated students.

At the outset, there are substantive reasons to doubt strong nonresponse bias. Students generally opted to take one course on a credit/no-credit basis during the first term. Roughly 78 percent of all courses were taken on a graded basis, with 79 and 78 percent of small and large sections taken on a graded basis, respectively ($p = .48$). By comparison, the fraction of students remaining in the Tennessee STAR experiment is under 50 percent (Krueger 1999, p. 503), and even among students remaining in the experiment, some 10 percent may not sit for the examination in a given year (Hanushek 1999). Students possess relatively little knowledge about how they might fare relative to the rest of class during the first year in law school. Other than LRW, which was ungraded from 2001 to 2008, grades are nearly exclusively based on one final exam at the end of the term. Moreover, for purposes of employment, the critical statistic is the observed cumulative GPA. From that perspective, the descriptive fact of a gender gap in large courses, and none in small, is relevant regardless. There is some evidence, however, that grading election may differ for small and large sections conditional on gender. On average, one more male student chooses to take a small section on a graded basis, compared to a large section. We pursue several approaches to assess the sensitivity of our inferences to this nonresponse.

C1. Missingness at Random

Under a missingness-at-random (MAR) model, grading election is assumed to be independent of potential grades earned, conditional on covariates and the observed treatment (Little and Rubin 2002; Horiuchi, Imai, and Taniguchi 2007; Hill, Reiter, and Zanutto 2006). The credibility of MAR depends critically on the range of covariates employed, which militates in favor of the more saturated outcome model. Under the MAR assumption, we can impute missing grades, which enables us to draw an inference about the small-section effect for the population of matriculated students. We do so via Gibbs sampling, iterating between imputing missing potential outcomes given the model parameters and drawing model parameters given the potential observed. Under MAR, the one-tailed p-value of τ, based on 1,000 draws from the posterior, is .01. Under MAR, results are (unsurprisingly) comparable to those in Table 4.

C2. Balance Conditional on Response

One of the critical questions with nonresponse is whether it destroys balance. In our setting, the question is whether the marginal student (that is, the student whose grading option is affected by the section size) differs in underlying ability, thereby confounding the gender class size estimate.

To investigate this, Table C1 reports patterns of missingness for men and women by section size. The table provides some evidence that more men appear to be taking small sections on a graded basis. Roughly one or two students per small section may be changing their grading option because of class size. If the marginal male student performs poorly on the exam, that may contaminate estimates. The mechanism by which section size should differentially affect men and women, however, is not obvious, especially because any information about relative standing in a small section should also affect a student's inferences about standing in a large section (recall that large sections are composites of small sections).

Table C1 also presents means of demographic covariates for students taking the course on a graded basis. There is no evidence that the marginal student differs sharply: the covariates remain balanced. The column reporting difference in differences in covariates shows that none appear to plausibly account for the difference in differences in the grade received.

Table C1. Patterns of Grade Elections and Missingness

	Men			Women			Difference in Differences
	Small	Large	Small − Large	Small	Large	Small − Large	
Graded (rate)	.82	.78	.037*	.76	.78	−.021	.057*
LSAT	168.9	168.9	> −.001	168.7	168.7	−.007	−.006
College GPA	3.80	3.81	−.006	3.82	3.82	.003	.009
Other grades	3.49	3.48	.005	3.45	3.44	.009	.004
Age	24.5	24.4	.085	23.7	23.7	.017	−.068
African American	.06	.06	> −.001	.09	.09	< .001	< .001
Asian American	.07	.08	−.003	.16	.17	-.007	−.004
Latino	.11	.11	−.004	.12	.13	−.008	−.002
Grade	3.46	3.49	−.026	3.45	3.43	.021	.047*

Note. The graded rate is the rate at which men and women enroll in small and large sections on a graded basis. Values for the demographic variables indicate whether the marginal male student varies along covariates. None of the covariates plausibly account for the difference in grades earned. LSAT = Law School Admission Test score; GPA = grade point average.

*$p < .05$.

C3. Principal Stratification

A powerful approach to addressing nonresponse is to focus on effects within principal strata (Frangakis and Rubin 2002; Rubin 2006). In the case of nonresponse (without compliance problems), the relevant principal stratum can be conceived of as students who always take a course on a graded basis, regardless of class size.[17] (Potential grades are otherwise ill defined.) Generalizing the framework to our setting, in which students choose both how many and which courses to take on the 3K grading system basis, poses somewhat of a challenge. Unlike typical principal stratification settings, however, our estimates are not identified by raw differences between treated and control units conditional on response. The difference-in-differences estimates are identified based on students taking both a small and large section on a graded basis, which might be considered the subpopulation of students whose grading choice is not affected by class size.

C4. Sensitivity Analysis

As Table C1 shows, there is evidence that being in a small section may induce one male student (and at most one male and one female student) to change his grading option. We therefore investigate the sensitivity of our results to assumptions about marginal students. Our approach is to remove the grade for lower-performing male students in small sections (which affects 42 male students [= 7 years × 6 sections per year]) and lower-performing female students in large sections (which affects 104 female students) and to examine the sensitivity of τ. In the worst-case scenario, the marginal male student taking the small section for a grade (only because of section size) is the worst male student, and the marginal female student taking a large section for a grade (only because of section size) is also the worst female student. As that seems unrealistic, we vary the percentile of the marginal student from 0 to 50 percent (that is, we focus only on male or female students below the median male or female student in the course), removing the grades of marginal male students from every small section and the grades of marginal female students from every large section.

Figure C1 presents the results. Across scenarios, the estimates of the effects remain comparable to our main results in Table 4. The one ex-

[17] In the parlance of principal stratification, these students are “always-takers.” See Rubin (2006).

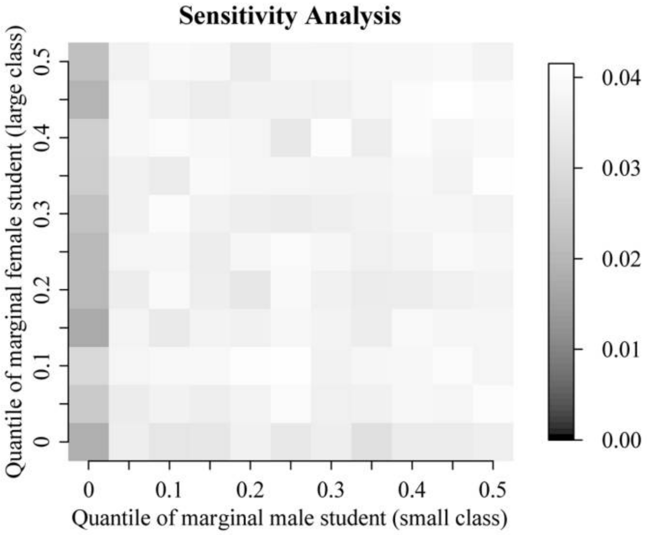

Figure C1. Sensitivity analysis of difference-in-differences estimates

ception is when the marginal male student is the worst-performing student in the section: point estimates of τ remain positive but become statistically indistinguishable from zero. Substantively, it seems unlikely that the worst-performing students are the ones taking small sections on a graded basis solely because of class size. In short, these sensitivity analyses suggest that nonresponse does not invalidate our findings in Table 4.

REFERENCES

Angrist, Joshua D., and Victor Lavy. 1999. Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. *Quarterly Journal of Economics* 114:533–75.

Areeda, Phillip E. 1996. The Socratic Method (SM) (Lecture at Puget Sound, 1/31/90). *Harvard Law Review* 109:911–22.

Banks, Taunya Lovell. 1988. Gender Bias in the Classroom. *Journal of Legal Education* 38:137–46.

Barnow, B., G. Cain, and Arthur Goldberger. 1980. Issues in the Analysis of Selectivity Bias. *Evaluation Studies Review Annual* 5:43–59.

Bell, Jacqueline. 2008. Law School Grads Face Tight Job Market. Law360. http://www.law360.com/articles/65265/law-school-grad-face-tight-job-market.

Bloodgood, Robert A., Jerry G. Short, John M. Jackson, and James R. Martindale. 2009. A Change to Pass/Fail Grading in the First Two Years at One Medical School Results in Improved Psychological Well-Being. *Academic Medicine* 84:655–62.

Bowers, Allison L. 2000. Women at the University of Texas School of Law: A Call for Action. *Texas Journal of Women and the Law* 9:117.

Box, George E. P., J. Stuart Hunter, and William Gordon Hunter. 2005. *Statistics for Experimenters: Design, Innovation, and Discovery*. Vol. 2. Hoboken, N.J.: John Wiley & Sons.

Carrell, Scott E., Marianne E. Page, and James E. West. 2010. Sex and Science: How Professor Gender Perpetuates the Gender Gap. *Quarterly Journal of Economics* 125:1101–44.

Clydesdale, Timothy T. 2004. A Forked River Runs through Law School: Toward Understanding Race, Gender, Age, and Related Gaps in Law School Performance and Bar Passage. *Law and Social Inquiry* 29:711–69.

Epstein, Cynthia Fuchs. 1993. *Women in Law.* Champaign: University of Illinois Press.

Dee, Thomas S. 2007. Teachers and the Gender Gaps in Student Achievement. *Journal of Human Resources* 42:528–54.

De Paola, Maria, Michela Ponzo, and Vincenzo Scoppa. 2011. Class Size Effects on Student Achievement: Heterogeneity across Abilities and Fields. *Education Economics* 21:35–53.

Ferguson, Ronald F. 1998. Can Schools Narrow the Black-White Test Score Gap? Pp. 318–74 in *The Black-White Test Score Gap,* edited by Christopher Jencks and Meredith Phillips. Washington, D.C.: Brookings Institution Press.

Frangakis, Constantine E., and Donald B. Rubin. 2002. Principal Stratification in Causal Inference. *Biometrics* 58:21–9.

Fredriksson, Peter, Björn Öckert, and Hessel Oosterbeek. 2013. Long-Term Effects of Class Size. *Quarterly Journal of Economics* 128:249–85.

Fryer, Roland G., and Steven D. Levitt. 2004. Understanding the Black-White Test Score Gap in the First Two Years of School. *Review of Economics and Statistics* 86:447–64.

Gilligan, Carol. 1982. *In a Different Voice: Psychological Theory and Women's Development.* Cambridge, Mass.: Harvard University Press.

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini. 2003. Performance in Competitive Environments: Gender Differences. *Quarterly Journal of Economics* 118:1049–74.

Greiner, D. James, and Donald B. Rubin. 2010. Causal Effects of Perceived Immutable Characteristics. *Review of Economics and Statistics* 93:775–85.

Guess, Andy. 2008. Stanford Law Drops Letter Grades. Inside Higher Ed. http://www.insidehighered.com/news/2008/06/02/stanford#sthash.3tmf01mY.dpbs.

Guinier, Lani, Michelle Fine, Jane Balin, Ann Bartow, and Deborah Lee Stachel. 1994. Becoming Gentlemen: Women's Experiences at One Ivy League Law School. *University of Pennsylvania Law Review* 143:1–110.

Hancock, Terence. 1999. The Gender Difference: Validity of Standardized Admission Tests in Predicting MBA Performance. *Journal of Education for Business* 75:91–93.

Hanushek, Eric A. 1999. The Evidence on Class Size. Pp. 131–68 in *Earning and Learning: How Schools Matter,* edited by Susan E. Mayer and Paul Peterson. Washington, D.C.: Brookings Institution Press.

Henderson, William D. 2003. The LSAT, Law School Exams, and Meritocracy: The Surprising and Undertheorized Role of Test-Taking Speed. *Texas Law Review* 82:975–1051.

Hill, Jennifer L., Jerome P. Reiter, and Elaine L. Zanutto. 2006. A Comparison of Experimental and Observational Data Analyses. Pp. 49–60 in *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family.* New York: Wiley.

Ho, Daniel E., and Kosuke Imai. 2006. Randomization Inference with Natural Experiments. *Journal of the American Statistical Association* 101:888–900.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis* 15:199–236.

Ho, Daniel E., and Donald B. Rubin. 2011. Credible Causal Inference for Empirical Legal Studies. *Annual Review of Law and Social Science* 7:17–40.

Holland, Paul W. 1986. Statistics and Causal Inference. *Journal of the American Statistical Association* 81:945–60.

Homer, Suzanne, and Lois Schwartz. 1989. Admitted but Not Accepted: Outsiders Take an Inside Look at Law School. *Berkeley Women's Law Journal* 5:1–74.

Horiuchi, Yusaku, Kosuke Imai, and Naoko Taniguchi. 2007. Designing and Analyzing Randomized Experiments: Application to a Japanese Election Survey Experiment. *American Journal of Political Science* 51:669–87.

Hoxby, Caroline M. 2000. The Effects of Class Size on Student Achievement: New Evidence from Population Variation. *Quarterly Journal of Economics* 115: 1239–85.

Jacobs, Jerry A. 1996. Gender Inequality and Higher Education. *Annual Review of Sociology* 22:153–85.

Jencks, Christopher, and Meredith Phillips. 1998. *The Black-White Test Score Gap*. Washington, D.C.: Brookings Institution Press.

Kay, Fiona, and Elizabeth Gorman. 2008. Women in the Legal Profession. *Annual Review of Law and Social Science* 4:299–332.

Kernan, Walter N., Catherine M. Viscoli, Robert W. Makuch, Lawrence M. Brass, and Ralph I. Horwitz. 1999. Stratified Randomization for Clinical Trials. *Journal of Clinical Epidemiology* 52:19–26.

Kerr, Orin. 2008. The Psychology of Grading. The Volokh Conspiracy. http://www.volokh.com/the-psychology-of-grading/.

Krueger, Alan B. 1999. Experimental Estimates of Education Production Functions. *Quarterly Journal of Economics* 114:497–532.

Little, Roderick J. A., and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*. New York: Wiley.

Lorenzo, Mercedes, Catherine H. Crouch, and Eric Mazur. 2006. Reducing the Gender Gap in the Physics Classroom. *American Journal of Physics* 74:118–22.

Miller, Diane L., and Charles E. Mitchell. 1994. Evaluation Achievement in Mathematics: Exploring the Gender Biases of Timed Testing. *Education* 114:436–38.

Monks, James, and Robert Schmidt. 2011. The Impact of Class Size and Number of Students on Outcomes in Higher Education. *B.E. Journal of Economic Analysis and Policy* 11, art. 62, pp. 1–17.

Mosteller, Frederick. 1995. The Tennessee Study of Class Size in the Early School Grades. *Future of Children* 5:113–27.

———. 1999. How Does Class Size Relate to Achievement in Schools? Pp. 117–30 in *Earning and Learning: How Schools Matter*. Washington, D.C.: Brookings Institution Press.

Niederle, Muriel, and Lise Vesterlund. 2007. Do Women Shy Away from Competition? Do Men Compete Too Much? *Quarterly Journal of Economics* 122: 1067–1101.

———. 2010. Explaining the Gender Gap in Math Test Scores: The Role of Competition. *Journal of Economic Perspectives* 24:129–44.

Ors, Evren, Frédéric Palomino, and Eloïc Peyrache. 2013. Performance Gender Gap: Does Competition Matter? *Journal of Labor Economics* 31:443–99.

Pocock, Stuart J., and Richard Simon. 1975. Sequential Treatment Assignment with Balancing for Prognostic Factors in the Controlled Clinical Trial. *Biometrics* 31:103–15.

Pollock, Steven J., Noah D. Finkelstein, and Lauren E. Kost. 2007. Reducing the Gender Gap in the Physics Classroom: How Sufficient Is Interactive Engagement? *Physical Review Special Topics—Physics Education Research* 3:010107-1–4.

Rhode, Deborah L. 1993. Missing Questions: Feminist Perspectives on Legal Education. *Stanford Law Review* 45:1547–66.

———. 2001. *The Unfinished Agenda: Women and the Legal Profession*. Report of the American Bar Association Commission on Women in the Profession. Chicago: American Bar Association.

Robins, Lynne S., Joseph C. Fantone, Mary S. Oh, Gwen L. Alexander, Marshal Shlafer, and Wayne K. Davis. 1995. The Effect of Pass/Fail Grading and Weekly Quizzes on First-Year Students' Performances and Satisfaction. *Academic Medicine* 70:327–29.

Rosser, Sue V. 1995. *Teaching the Majority: Breaking the Gender Barrier in Science, Mathematics, and Engineering*. New York: Teachers College Press.

Rubin, Donald B. 2006. Causal Inference through Potential Outcomes and Principal Stratification: Application to Studies with “Censoring” Due to Death. *Statistical Science* 21:299–309.

———. 2008. For Objective Causal Inference, Design Trumps Analysis. *Annals of Applied Statistics* 2:808–40.

Samida, Dexter. 2004. The Value of Law Review Membership. *University of Chicago Law Review* 71:1721–48.

Taber, Janet, Marguerite T. Grant, Mary T. Huser, Rise B. Norman, James R. Sutton, Clarence C. Wong, Louise E. Parker, and Claire Picard. 1988. Gender, Legal Education, and the Legal Profession: An Empirical Study of Stanford Law Students and Graduates. *Stanford Law Review* 40:1209–97.

Weiss, Catherine, and Louise Melling. 1988. The Legal Education of Twenty Women. *Stanford Law Review* 40:1299–1369.

Wightman, Linda F. 1996. *Women in Legal Education: A Comparison of the Law School Performance and Law School Experiences of Women and Men.* Newton, Penn.: Law School Admission Council.

Xie, Yue, and Kimberlee A. Shauman. 2003. *Women in Science: Career Processes and Outcomes.* Vol. 26. Cambridge, Mass.: Harvard University Press.

Yale Law Women. 2012. Yale Law School Faculty and Students Speak up about Gender: Ten Years Later. Report prepared by Yale Law Women. Yale Law School, New Haven, Conn.