Does Class Size Reduce the Gender Gap? A Natural Experiment in Law[·]

43 J. LEGAL STUD. (forthcoming)

Daniel E. Ho[†]
Stanford Law School

Mark G. Kelman[°]
Stanford Law School

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 2001-2011. We provide evidence (i) that small sections closed a slight (but substantively and highly statistically significant) gender gap existing in large sections from 2001-08; (ii) that reforms in 2008, which modified the grading system and instituted small, graded, writing and simulation-intensive courses, eliminated the gap entirely; and (iii) 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.

[·] Thanks to 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 Center for Law and Social Sciences Workshop at the University of Southern California School of Law, the Faculty Colloquium at the University of Colorado Boulder School of Law, the faculty workshop at Duke Law School, the Faculty Colloquium at Emory University School of Law, and the 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, Mila Fernandez-Ronquillo for help in collecting law school data, and the many faculty members responding to our inquiries about pedagogy.

[†] Professor of Law, Stanford Law School; Address: 559 Nathan Abbott Way, Stanford, CA 94305; Tel: 650-723-9560; Fax: 650-725-0253; Email: dho@law.stanford.edu, URL: http://dho.stanford.edu.

[°] James C. Gaither Professor of Law & Vice Dean, Stanford Law School, Address: 559 Nathan Abbott Way, Stanford, CA 94305; Tel: 650-723-4069; Fax: 650-725-0253.

1. INTRODUCTION

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

One promising intervention to reduce achievement gaps is to reduce class size. Smaller classes may, for instance, enable teachers to better understand and teach to students at different levels. Jencks and Phillips (1998) 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" (p. 44). The best evidence comes from the Tennessee STAR[1] experiment, which randomly assigned students in kindergarten through third grade to large and small classrooms. Results suggest that 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% of students leaving the experiment[2]), noncompliance (with 10% switching from large to small classrooms), and nonresponse (with 3 to 12% 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 demographic gaps.[3]

A separate literature, focusing on gender gaps, particularly in math and science, examines the role of competition and 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," where 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; Niederle and Vesterlund 2010). Ors, Palomino, and Peyrache (2013) find that men outperform women on entrance exams to a top-ranked French business school, which is reversed in less competitive high school finishing exams. Carrell, Page, and West (2010), in a study that is closest to ours in research design, study a natural experiment in the U.S. Air Force Academy, where students are randomly assigned to professors for mandatory courses. Female professors greatly improve women's performance in math and science courses (see also Dee).[4]

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 (e.g., Banks 1988; Guinier et al. 1994; Rhode 1993; Rhode 2001; Weiss and Melling 1988). Voluminous research confirms that women participate less frequently in the classroom, although some document relative parity in (or greater comfort by women with) small courses (Yale Law Women 2012; Banks 1988; Weiss and Melling 1988, p. 1334-35). Because grades matter considerably in the legal profession, numerous scholars have examined

[1] STAR stands for "Student-Teacher Achievement Ratio."

[2] See Krueger (1999) (Table 1, documenting attrition rates from 47 to 53% for students entering the experiment in kindergarten or first grade).

[3] See, e.g., Fredriksson, Öckert, and Oosterbeek (2012); Hoxby (2000); Angrist and Lavy (1999); and Fryer and Levitt (2004).

[4] In 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 gender of the instructor.

the gender gap in law school grades, with heterogeneous findings across schools.[5] 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 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 2001-2011. We collect rich individual-level covariate and grade information for every student in every mandatory first-year course to study whether small sections reduce the gender gap in law. We find they do.

Our study has several virtues. First, unlike observational studies, where class size is often confounded (e.g., 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.[6] 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 if section assignments were not randomized (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 importantly, ability) to identify the effect of small classes by gender.

Third, our study has advantages even relative to other experimental approaches. In Tennessee STAR, for instance, some 60% of students leave or switch away from their assigned classrooms.[7] 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 was 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,[8] 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 post-graduate professional school setting.

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

[5] See Kay and Gorman (2008, p. 302) ("Studies have offered conflicting evidence as to whether there is a gender difference in law school grades."); Clydesdale (2004) (finding no gender difference in first-year GPAs); Wightman (1996) (finding a slight gender gap in first-year GPAs); Guinier et al. (1994) (finding a gender gap in first-year GPAs at the University of Pennsylvania); Bowers (2000) (finding gender gap in first-year GPAs at University of Texas); Homer and Schwartz (1989) (finding a gender gap in contracts and property at UC Berkeley); Taber et al. (1988) (finding no gender gap in membership in the Order of the Coif at Stanford Law School).

[6] The closest are De Paola, Ponzo, and Scoppa (2013), Krueger (1999), and Fredriksson, Öckert, and Oosterbeek (2012).

[7] See Krueger (1999) (Table 1, attrition rates) and Hanushek (1999) (discussing attrition and failure to sit for exams).

[8] But see Monks and Schmidt (2010), who note,"[o]nly a handful of studies ha[s examined] class size . . . in tertiary education."

Section 5 examines the effects of class size on the gender gap from 2001-08, when the school employed numerical GPA grades. Applying a difference-in-differences approach, we show that small sections eliminate a small, but highly statistically significant, gender gap that exists in large sections. Section 6 examines the evidence after educational reforms of 2008, which changed the grading system to an Honors/Pass 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 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 other educational settings, students had no choice of which course to enroll in. Student enrollment choices (e.g., 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 first-year, mandatory curriculum, and then detail the precise mechanisms of (1) grouping students into sections and (2) 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, or 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. LRW courses were graded on a mandatory credit / restricted credit / no-credit basis. In other courses, students could elect to be graded on a credit / no-credit basis (the so-called "3K option"[9]), and the 3.4 mean requirement applied regardless of the grading option.

Beginning in Fall 2008, the law school instituted a series of pedagogical reforms. First, courses would be graded on an Honors / Pass basis (the HP system). The required range was 30-40% Honors for doctrinal courses. The rationales for grade reform were to reduce "grade curve shopping" and to eliminate what was perceived as a falsely precise, and to many students an intimidating, numerical GPA system.[10] 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 fall quarter courses continued to meet for the same duration as previously. Winter courses were adjusted to the quarter system.

[9] "3K" refers to the fact that there are three grades under that option: credit, restricted credit, and no credit. In practice, restricted credit and no credit were rarely used.

[10] *See* Andy Guess, *Stanford Drops Letter Grades*, INSIDE HIGHER ED, June 2, 2008; Orin Kerr, VOLOKH CONSPIRACY, Sep. 27, 2008.

Two modifications were that: (i) LRW would be graded and shortened to the fall term, and (ii) 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, with a wide range of writing and simulation exercises (initially, drafting a complaint, three briefs, and a bench memo; delivering and judging oral arguments; taking and defending a deposition). The required range in LRW and Federal Litigation was 35-50% Honors.

Throughout the entire 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 would be taught exclusively to the small section. The exact substantive field (e.g., contracts or criminal law) would vary both *within* and *across* entering classes, based largely on faculty availability. Other doctrinal courses were typically taught in a large class, combining two small sections (i.e., roughly 60 students). When Federal Litigation was introduced, small sections were split into groups of roughly 18 students (10 sections per incoming class), further divided into legal teams of four to five students each. Depending on the instructor, Federal Litigation class meetings were often held exclusively between the instructor and the legal team. At all times, exams in doctrinal courses, on which final grades are overwhelmingly based, were graded blindly, ruling out the possibility of sheer instructor grading bias.[11]

2.2. Grouping and Assignment Mechanisms

To understand the mechanism by which students ended up in particular small sections, we detail two decisions: (1) *grouping* students into small sections, and (2) *assigning* small sections to specific classes. These decisions were made to ensure fairness in and representativeness (i.e., balance) across section assignments, not to study class size effects.

Grouping students into small sections worked 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 LSAT and undergraduate GPA), assigning numbers 1 to 6 to each student. To balance the academic index, but to retain the simplicity of assignment, the Dean systematically cycled through the numbers 1-6 (first in order then in reverse order) going down the list of sorted names: e.g., 1, 2, 3, 4, 5, 6, 6, 5, 4, 3, 2, 1, and so on. The academic index amongst Stanford students is coarse due to range compression: for instance, the class of 2005 had only 7 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 specific 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 based on instructor "fit" or predicted ability to succeed

[11] Blind grading may not rule out the possibility that instructor's may devalue "female voice" (Gilligan 1982) on exams.

in a particular small or large section. Student characteristics were not consulted in assigning sections to courses, except for very rare circumstances.[12]

Grouping students into sections, as Appendix A shows, is best characterized as approximating a form of stratified block randomization (see 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 precise order of students in the list is stochastic, as matriculation decisions for specific students can hinge on chance factors (e.g., deferrals of admission). It is very unlikely that the student list thereby has a (periodic) relationship (e.g., every twelfth student has a low income background), which would confound the section grouping. 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 are balanced along all covariates. Appendix A demonstrates that section grouping was effectively a form of block randomization, stratifying on gender and ethnicity.[13]

3. DATA

We compile data from the Office of the Admissions on first-year students and match these to data from the Office of Registrar on grades awarded to each student in a course. Our primary data consists of 15,689 grades assigned in mandatory first-year courses by 91 instructors to 1,897 students from 2001-2012. Table 1 provides a breakdown of the raw data for the two observation periods under the GPA system (2001-08) and the HP system (2008-11). Prior to 2008, the overall mean grade was 3.46, which is higher than the mandatory mean of 3.4 due to students electing the 3K option. (Instructors graded all exams collectively, without knowledge of the grading option.) The overall proportion of Honors was 0.42, which exceeds 40% because LRW and Federal Litigation are subject to a 50% cap on Honors.

			Grades		
Period	Students	Instructors	All	GPA / HP	Mean
2001-8	1,193	62	9,539	5,600	3.46
2008-11	704	58	6,150	6,141	0.42

Table 1: Summary statistics for sample. During the 2001-08 period, the law school employed a numerical GPA grading system. Beginning in Fall 2008, the law school switched to an Honors / Pass (HP) system. "All" grades include courses graded on the 3K or mandatory credit basis, while "GPA / HP" grades include only those evaluated by numerical GPA or H/P grades.

Table 2 reports summary statistics of incoming credentials 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 represent minority groups (e.g, 15% of women are Asian-American, compared to 8% of men). These differences along observables are important in understanding the gender gap and

[12] These involved instances where section assignment was adjusted to avoid conflicts of interest (e.g., when faculty members were related to the student). These were exceedingly rare and grouping of sections remained intact.

[13] Grouping and assignments into Federal Litigation sections worked comparably.

class size effects --- all model-based estimates we present below control for ethnicity or student fixed-effects.

	Men	Women	SD
Academic background			
Law School Admissions Test score (LSAT)	169.0	168.9	4.2
Undergraduate degree GPA	3.81	3.82	0.19
Academic index (LSAC)	3.42	3.41	0.15
Master's degree	0.18	0.12	0.36
Ph.D.	0.05	0.03	0.19
Age	24.6	23.8	2.8
Ethnicity			
White	0.59	0.51	0.50
Latino	0.12	0.12	0.32
Asian-American	0.08	0.15	0.32
African-American	0.08	0.11	0.29
Undergraduate institution			
Stanford	0.10	0.11	0.30
Harvard	0.06	0.07	0.25
Yale	0.07	0.07	0.25
Berkeley	0.03	0.04	0.18

Table 2: Covariates at time of matriculation. The first two columns present the means by gender, and the third column presents the pooled standard deviation (SD).

Figure 1 plots the raw distribution of grades assigned by gender. The grey 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 0.05 GPA points lower than those for men (p-value < 0.0001). 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, doctorate, professional degree, fixed effects for undergraduate institution, instructors, and courses).[14] 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* gender --- despite the gap, individual women and men perform along the entire range of GPAs.

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

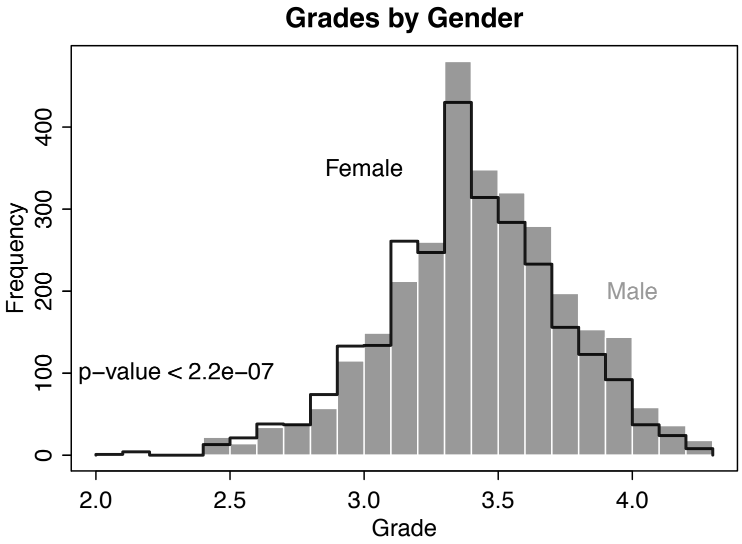

Figure 1: Raw gender gap. The histograms present the grades assigned in individual courses by gender. The grey histogram represents male students and the black lines represent women.

Although the gender gap is small in absolute magnitude, the gap represents roughly 15% of the pooled GPA standard deviation --- in a profession that prizes law school performance (see Henderson 2003). To illustrate the gap's substantive importance, we examine data on 487 clerkship applications by Stanford students from 2003-2008 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: a 0.05 GPA increase from 3.6 to 3.65 is associated with a 7% (statistically significant) increase in the probability of securing a federal appellate clerkship.[15] Similarly, we use data from the fall on-campus recruitment, which is the primary method by which students secure private sector jobs (the modal job for students upon graduation). We use data on 2,949 on-campus interviews in fall of 2008 and calculate the rate at which students are offered callback interviews relative to the number of on-campus interviews. (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.) Again, we confirm that grades have a strong positive correlation with the rate at which students are offered callback interviews: a 0.05 GPA increase from 3.25 to 3.3 is associated by a nearly 5% increase in the callback rate.[16] It is worth noting that law firms appear to have become even more grade-sensitive since 2008.[17] The private callback rate from 2008 may thereby understate the effect of grades on the current labor market. In short, while small in absolute magnitude, the 0.05 GPA gender gap matters.

4. RANDOMIZATION CHECKS

[15] This is estimated with a 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 to an appellate clerkship.

[16] This is estimated 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.

[17] See, e.g., Jacqueline Bell, *Law School Grads Face Tight Job Market*, LAW360, Aug. 6, 2008.

Although there are strong reasons to believe that the assignment mechanism of sections to specific 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 on the *x*-axis against 12 covariates on the *y*-axis. Each red dot represents the mean (or proportion) for one small section. The white line represents the mean (or proportion) for the incoming class. The grey intervals represent the 95% confidence interval 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.

The figure 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 line up are closer to the class mean than under pure randomization. Other covariates approximate the randomization distribution. Although some sections fall outside of the 95% interval, the rate is much lower than Type I error rates: under randomization, we would expect roughly 40 such deviations [= 0.05 α-level × 6 sections / 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 section assignments.

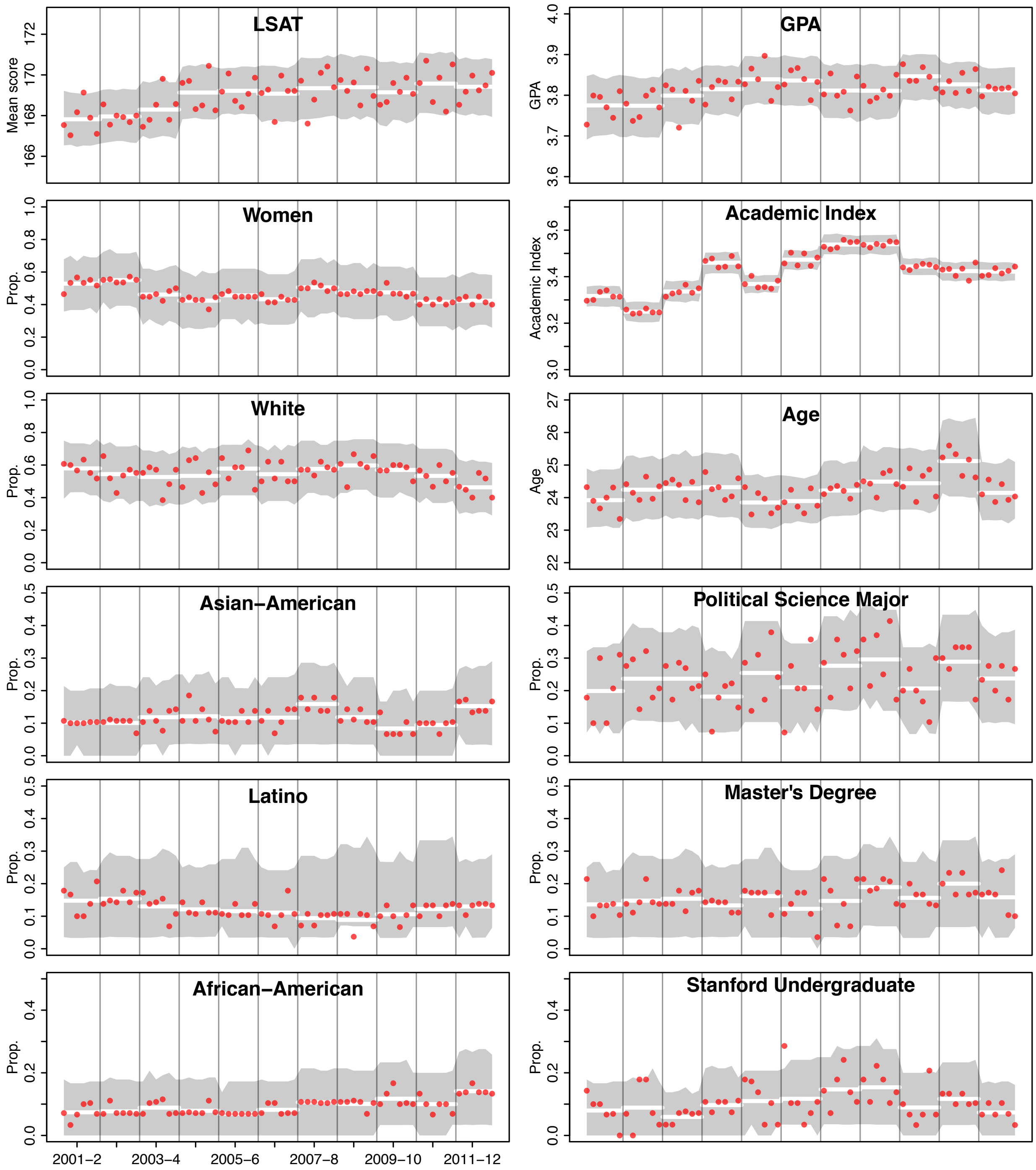

Figure 2: Randomization checks for small sections. Each red dot represents the mean for one of six small sections in an entering class, sorted chronologically by entering class on the *x*-axis. Vertical lines separate unique entering classes. White horizontal lines plot the mean for the entire entering class. Grey intervals plot the (simulated) 95% confidence intervals of means under the null of randomization.

5. CLASS SIZE EFFECTS, 2001-2008

We now focus on assessing the causal effect of class size during the time of the GPA system (2001-08). Due to the number of changes --- particularly in grading --- Section 6 examines the post-2008 period separately.

Figure 3 plots quantile-quantile plots comparing the raw grade distributions for men (on *x*-axes) and women (on *y*-axes) conditional on section size. 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 0.05 points higher than women in large sections (*p*-value < 0.01).

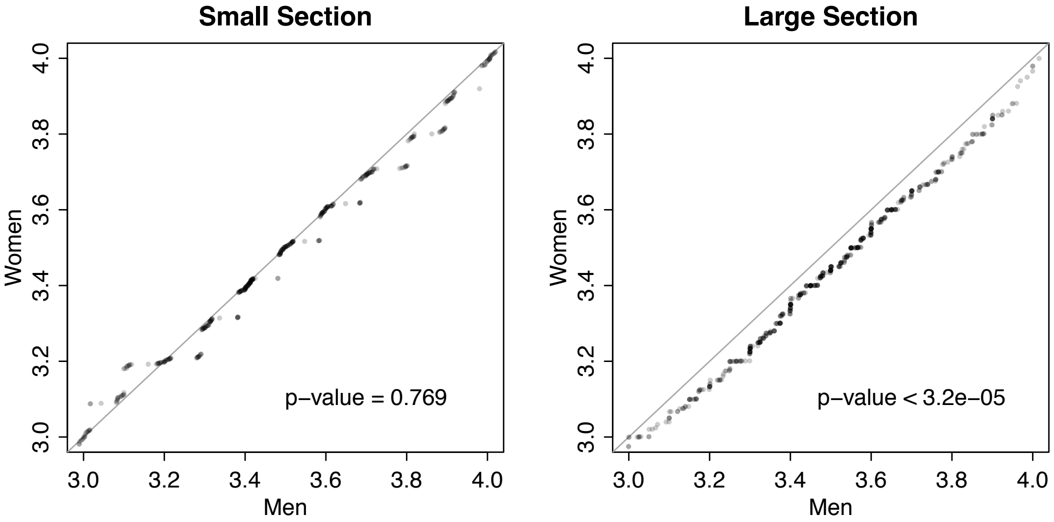

Figure 3: Quantile-Quantile plots comparing performance on men and women in small and large sections, with dots randomly jittered for visibility. The left panel shows that there is no statistically distinguishable difference between men and women in small sections. The right panel shows that men on average earn 0.05 GPA points more than women.

Table 3 provides summary statistics on the differences in means between men and women across large and small classes. The bottom right cell calculates the raw difference-in-differences (*p*-value < 0.05): women tend to outperform men by 0.05 GPA points in small sections relative to large sections.

	Large	Small	Difference (small-large)
Men	3.488	3.461	-0.026
Women	3.433	3.454	0.021
Difference (men-women)	0.054**	0.007	**0.047****

Table 3: Raw grade averages by men and women in large and small sections. The bottom row presents the gender difference conditional on class size, subtracting female from male performance. The right column presents the class size difference conditional on gender, subtracting performance in large from small sections. The bottom right cell presents the difference-in-differences. ** indicate statistical significance at $\alpha = 0.05$.

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

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

where (a) s indexes students, i indexes instructors, and c indexes course subjects; (b) $Y_{s,i,c}$ represents the numerical grade earned by student s in course c taught by instructor i; (c) $T_{s,i,c}$ equals 1 if the student was enrolled in the "treatment" of a small section and 0 if not, and (d) G_s equals 1 if the gender of student s is female and 0 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 (i) any student-specific, course-invariant effects (chiefly, ability), (ii) instructor-specific, course-invariant effects, and (iii) course-specific effects.

By construction, student fixed effects (α) control for gender, age, LSAT score, undergraduate GPA, and any other student-specific entering characteristics. 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 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 3K elections can cause courses to deviate from the 3.4 mean.

Table 4 presents results. Column A reports the simplest estimates, with fixed effects for ethnicity. The gender gap decreases slightly to 0.039 GPA points, but is reversed entirely in small sections. Columns B-D add student, instructor, and course fixed-effects sequentially. Model estimates remain stable: while small sections cause women to improve performance by 0.04 GPA points, they diminish men's performance by 0.03 GPA points. These results provide considerable evidence that small classes diminish the gender gap existing in large sections.

	A	B	C	D
Small section × female (τ)	0.045** (0.023)	0.044** (0.018)	0.042** (0.019)	0.041** (0.019)
Small section (λ)	-0.027* (0.015)	-0.024* (0.012)	-0.029** (0.014)	-0.032** (0.014)
Female	-0.038*** (0.008)			
Ethnicity FE	Yes	No	No	No
Student FE (α)	No	Yes	Yes	Yes
Instructor FE (η)	No	No	Yes	Yes
Course FE (κ)	No	No	No	Yes
Parameters	10	1,184	1,222	1,227
R^2	0.10	0.52	0.53	0.53
N	5,600	5,600	5,600	5,600

Table 4: Difference-in-differences linear regression estimates. Standard errors, clustered by course section, are presented in parentheses. "FE" indicates fixed effects. "Parameters" indicates the number of parameters estimated in the regression. "*N*" indicates the sample size. $^{*}/^{**}/^{***}$ denote statistical significance at α-levels of 0.1, 0.05, and 0.01, respectively.

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

6. THE VANISHING GAP, 2008-2011

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 left column shows that the gender gap vanishes under the HP system. Women earn Honors in roughly 42% of courses, compared to 41% of courses by men. The second and third columns suggest that women continue to perform slightly better than men in small sections. The effect, however, appears to be entirely driven by Federal Litigation.

	All	Large	Small	Fed. Lit.
Men	0.414	0.386	0.452	0.471
Women	0.417	0.367	0.483	0.545
Difference (women-men)	0.002	-0.018	0.030	0.074**

Table 5: Gender differences under HP grading system. The first two rows present the proportion of Honors earned by men and women across all classes (first column), large sections (second column), small sections (third column) (including Federal Litigation and LRW), and Federal Litigation (fourth column). Federal Litigation and LRW are subject to a grading guideline of 35-50% Honors. The bottom row presents the gender difference conditional on course. ** denote statistical significance at α-level of 0.05, using Fisher's exact test.

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 has disappeared. Women systematically earn more Honors in Federal Litigation, a result robust to the full set of fixed effects. Relative to a large section, Federal Litigation increases women's probability of earning Honors by 0.18, compared to only 0.08 for men. The differential grading guideline raises the probability of Honors, but does so disproportionately for women.

	A	B	C	D
Fed. lit. × female	0.48**	0.62***	0.62***	0.62***
	(0.19)	(0.24)	(0.24)	(0.24)
Fed. Lit.	0.39***	0.58**	-0.26	-0.26
	(0.08)	(0.13)	(0.16)	(0.16)
Small section × female	-0.03	-0.05	-0.05	-0.05
	(0.15)	(0.16)	(0.16)	(0.16)
Female	-0.00			
	(0.07)			
Ethnicity FE	Yes	No	No	No
Student FE	No	Yes	Yes	Yes
Instructor FE	No	No	Yes	Yes
Course FE	No	No	No	Yes
Parameters	14	710	767	772
Res. Dev.	7834	5375	5340	5339
N	6,141	6,141	6,141	6,141

Table 6: Difference-in-differences estimates from logistic regression model of the probability of an "Honors" grade in a course. Standard errors, clustered by course section, are presented in parentheses. "FE" indicates fixed effects. "Parameters" indicates the number of parameters estimated in the regression. "*N*" indicates the sample size. **/*** denote statistical significance at α-levels of 0.05 and 0.01, respectively. Models A-B include LRW and LRW × female fixed effects, as LRW is subject to a different grading guideline; as LRW instructors are unique to the course, these are not estimated in Models C-D, which include instructor fixed effects.

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 the gender gap 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 last four years of the GPA system, employing comparable grading guidelines of no more than 40% Honors. Modeling these shadow Honors strongly rejects the null hypothesis of no gender differences under the GPA system (p-value<0.001). 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 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. Lastly, 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 HP system may have removed, at least subjectively, a degree of competitiveness from first-year exams. Recall that one of the predominant assumptions about the HP system is that it takes the pressure off; 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 found that when the University of Virginia medical school changed from letter to pass-fail grading, women disproportionately exhibited gains in psychological well-being.[18]

Our finding about Federal Litigation provides more insight into specific pedagogical techniques potentially affecting the gender gap. What distinguishes Federal Litigation from other doctrinal courses (and LRW to a lesser extent) is that it (a) is based entirely on simulation, assigning students into an affirmative litigation position in a real case, (b) has no final exam under timed conditions, (c) provides substantial feedback throughout the class, and (d) is the smallest mandatory first-year class, with effectively only 4-5 students for many class meetings.[19] Each of these pedagogical features may affect the gender gap (see Rhode 1993 (simulation and feedback); Miller and Mitchell 1994 (timed exams)). The scope of simulation-intensive exercises would simply not be possible in large classes. Importantly, there is no evidence of the gender gap reversal in LRW courses, which share the same set of core instructors.[20] 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 mechanism by which class size and pedagogy differentially affect students. To develop a sense of the mechanism, we surveyed each of the instructors who taught first-year courses at the law school from 2001-2008 (with all but one instructor responding), and consulted final exams, syllabi, and course evaluations whenever available. We collected information on exam type (e.g., open vs. closed book, duration), class participation (e.g., 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% of small sections administered practice exams with model answers and/or class discussion, compared to 14% of large sections (*p*-value = 0.001); and 29% of small sections administered practice exams with grades and/or individualized feedback, compared to 7% of large sections (*p*-value = 0.001). The primary reason for this difference is practical: unlike other parts of the university, law faculty perform

[18] 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. See also Robins et al. (1995) (finding that pass/fail grading at the University of Michigan medical school reduced anxiety without a reduction in performance).

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

[20] The finding that women outperform men is, if anything, *more* pronounced when excluding Federal Litigation sections taught by four instructors that do not teach the LRW course. This rules out the possibility that the Federal Litigation effect is driven by instructor gender bias with non-anonymous grading.

nearly all grading and less than one-fifth of courses employ teaching assistants, making 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 through the two quarters: students are assigned real advocacy roles with discrete issues in an actual case involving vivid issues. The extensive interactive exercises (e.g., 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 (see, e.g., Guinier et al. 1994, pp. 39-40, Weiss and Melling 1988, pp. 1317, 1348), the simulation basis of Federal Litigation may be the mechanism reversing the gender gap. This evidence is consistent studies suggesting that interactive engagement techniques can reduce the science gender gap (Lorenzo, Crouch, and Mazur 2006, Rosser 1995).[21]

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 instructor 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.[22]

Second, because Stanford employs "norm-referenced" grading (i.e., 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. Put differently, we cannot distinguish whether small sections closed an achievement or test score gap. Criterion-referenced grading would be the obvious, but likely infeasible, way forward. A less ideal approach would be to re-grade exams in different sections of the same subject matter 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 forward. 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 (a) deploy a form of randomization, and (b) to 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 (see, e.g., Areeda 1996).

Lastly, 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 ABA-accredited law schools in 2013 was 66 students (SD=17). The vast majority of law schools enroll sections that are far larger than Stanford's. Moving toward the small section in our experiment may hence demand considerable resources. On the other

[21] 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).

[22] We cannot reject the hypotheses that gender effects are the same for (a) instructors teaching both small and large sections, and (b) instructors teaching exclusively small or large sections.

hand, Stanford managed to create Federal Litigation without substantial additional cost. The school shifted existing instructors to create Federal Litigation sections, suggesting that not all reductions in class sizes need be a drain on resources.

In sum, our study demonstrates that 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 (and possibly induced by competitive pressure of) the grading system. This finding might explain the cacophony of findings about the existence of gender gap across law schools. The key now is how to address it when it does. 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 yields balance along gender and ethnicity beyond what would be expected under pure randomization. Here, we show that grouping of 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 twelve covariates under pure randomization, from 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 groups (Asian-American, Latino, and African-American). Within each stratum, we apply block randomization, assigning section numbers 1-6 randomly without replacement within 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 4 plots results. The dark dashes along the *x*-axes indicate observed means across six sections. The black outlined histogram plots the pure randomization distribution and the grey histogram plots the block randomization distribution. Dark dashes track the latter extraordinarily well, showing that the Associate Dean's grouping is comparable to block randomization.

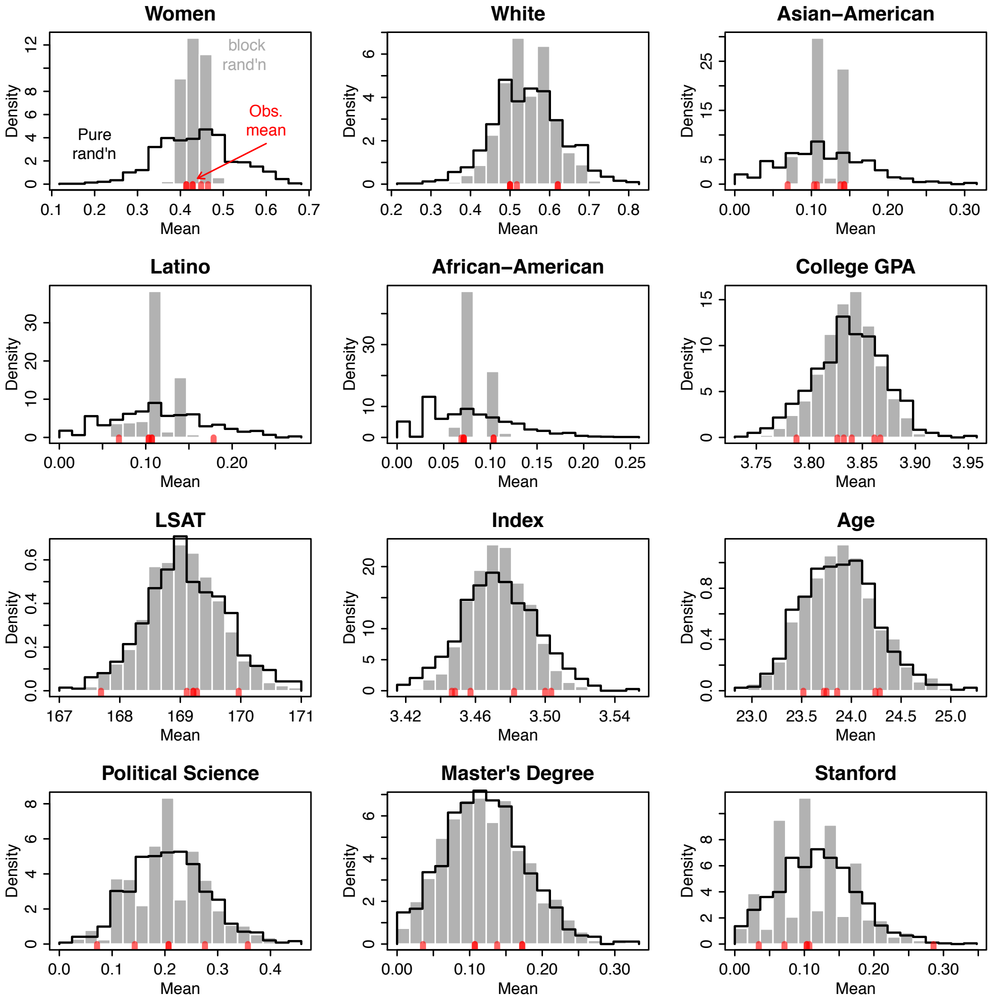

Figure 4: Stratified block randomization. The red dashes along the *x*-axis represent the observed means of the covariate for six small sections for the incoming class in 2006, and are slightly transparent for visibility. The black outlined histogram represents the randomization distribution of the mean under pure randomization. The grey histogram represents the randomization distribution of the mean under stratified block randomization, stratifying on gender, Asian-American, Latino, and African-American indicators. Randomization distributions are approximated via Monte Carlo simulation. These plots show that the Associate Dean's balancing is tantamount to a form of stratified block randomization: for gender and ethnicity, balance is greater than would be expected under pure randomization; observed means follow the stratified block randomization distribution.

APPENDIX B: SECTION ASSIGNMENT VIOLATING IDENTIFICATION ASSUMPTIONS

Even were assignment non-random, confounded assignment mechanisms that would artificially generate the gender effects are difficult to conjure up. Applying differences-in-differences --- when large sections are simply composites of small sections (i.e., including *the same set of students*) --- rules out many simple manipulations. Two assignment mechanisms that would violate our identification assumptions are as follows. One possibility is that the Associate Dean observes information about students that is course and / or instructor specific, and disproportionately assigns female students who are predicted to perform well to the cognate small section. Because the Associate Dean does not actually take into account any information on how the numbers 1-6 map to specific 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 were only two sections and that students are either high achieving or low achieving. If one section combined a small number of women with large number of men while another combined a large number of women with a small number of men --- all while keeping the relative proportion of high achievement students constant within a section --- that could artificially generate our findings. Consider the hypothetical section assignments in Table 7.

	Small Section A		Small Section B		Large Section	
	H/N	Prop.	H/N	Prop.	H/N	Prop.
High achieving women / all women	3/4	0.75	8/20	0.40	11/24	0.46
High achieving men / all men	15/20	0.75	6/15	0.40	21/35	0.60

Table 7: Hypothetical section assignment that would violate a difference-in-differences identification strategy. Each row presents statistics conditional on gender. The "H/N" columns represents the number of high achieving individuals divided by the number of all individuals in that class (conditional on gender). The "Prop." Columns indicate the proportion of individuals that are high-achieving in that class (conditional on gender). In this scenario, there would be no gender gap in small sections (and norm-grading would lead to an equivalent grade distribution across small sections A and B), but a considerable gender gap would exist in large sections.

The first cell indicates that small section A has three high achieving women out of four women and 15 high achieving men out of 20 men. 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 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 the grade for students who elected to take the course on a credit / no-credit basis. (As the HP system eliminates grading elections, nonresponse poses no problem for the 2008-2011 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. Amongst respondents (i.e., students taking the course for a grade), treated individuals may actually 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. Generally, students opted to take one course on a credit/no-credit basis during the first term. Roughly 78% of all courses were taken on a graded basis, with 79% and 78% of small and large sections taken on a graded basis, respectively (*p*-value = 0.48). By comparison, the fraction of students remaining in the Tennessee STAR experiment is under 50% (Krueger 1999, p. 503) and even amongst students remaining in the experiment some 10% may not sit for the examination in a 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 2001-08, 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 the 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.

1. Missingness at Random

Under "missingness at random" (MAR), the 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, enabling us to draw an inference about the small section effect for the population of matriculated students. We do so via Gibbs sampling, iterating between (a) imputing missing potential outcomes given the model parameters, and (b) 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 0.01. Under MAR, results are (unsurprisingly) comparable to those in Table 4.

2. Balance Conditional on Response

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

	Men		Diff.	*Women*		Diff.	
	Small	Large	(small-large)	Small	Large	(small-large)	Diff.-in-diff.
Graded (rate)	0.82	0.78	0.037**	0.76	0.78	-0.021	0.057**
LSAT	168.9	168.9	>-0.001	168.7	168.7	-0.007	-0.006
College GPA	3.80	3.81	-0.006	3.82	3.82	0.003	0.009
Other grades	3.49	3.48	0.005	3.45	3.44	0.009	0.004
Age	24.5	24.4	0.085	23.7	23.7	0.017	-0.068
African-American	0.06	0.06	>-0.001	0.09	0.09	<0.001	<0.001
Asian-American	0.07	0.08	-0.003	0.16	0.17	-0.007	-0.004
Latino	0.11	0.11	-0.004	0.12	0.13	-0.008	-0.002
Grade	3.46	3.49	-0.026	3.45	3.43	0.021	0.047**

Table 8: Patterns of taking course on a graded basis. The first row presents the rate at which men and women take small and large sections on a graded basis. There is a slight difference, with roughly one male being more likely to take a small section on a graded basis. The next rows investigate whether this marginal male student --- whose grading basis may be affected by treatment assignment --- varies along covariates, to potentially explain the disappearance of the gender gap in small sections. The right column presents difference-in-differences. None of the covariates plausibly accounts for the difference in grades earned (presented in the last row). ** denote statistical significance at α-level of 0.05.

To investigate this, Table 8 reports patterns of missingness for men and women by section size. The first row provides some evidence that more men appear to be taking small section on a graded basis. Roughly one to two students per small section may be changing their grading option due to 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 the large section (recall that large sections are composites of small sections).

The eight middle rows calculate means of covariates amongst students taking the course on a graded basis. There is no evidence that the marginal student differs sharply: covariates remain balanced. The last column reports differences-in-differences in covariates showing that none appear to plausibly account for the difference-in-differences in the grade received.

3. Principal Stratification

A powerful approach to address nonresponse is by focusing 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.[23] (Potential grades are otherwise ill-defined.) Generalizing the framework to our setting, where students choose both how many and which courses to 3K, 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 off of students taking both a small and large section on a graded basis, which might be considered the subpopulation of students whose grading basis is not affected by class size.

4. Sensitivity Analysis

As Table 8 showed, there is evidence that small sections may induce one male student (and at most one male and one female student) to change their 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 (affecting 42 male students [= 7 years × 6 sections/year]) and lower-performing female students in large sections (affecting 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% (i.e., focusing only on male or female students below the median male or female student in the course), removing marginal male student grades from every small section and marginal female student grades from every large section.

[23] In the parlance of principal stratification, these students are "always-takers." See Rubin 2006.

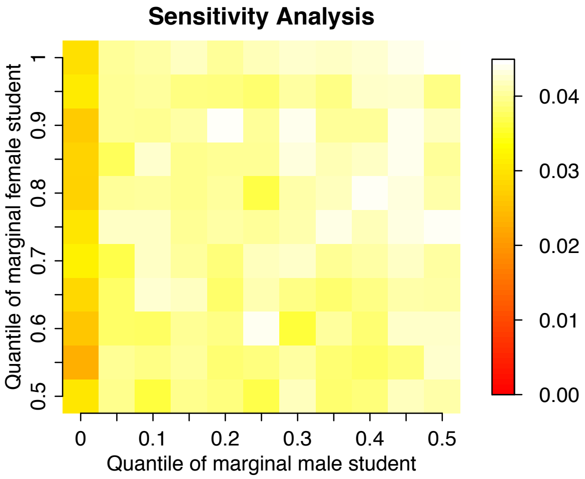

Figure 5: Sensitivity analysis of difference-in-differences estimates. Each cell represents the estimate of τ, under different assumptions about the marginal male and female student, represented by the x and y-axes. The right legend denotes the size of the point estimate of τ. The x-axis represents the identity of the marginal male student: for instance, at the 0% quantile, grades of all male students receiving the lowest grade in their respective small sections are removed. Conversely, at the 100% quantile on the y-axis, grades for all female students receiving the highest grade in their respective small sections are removed.

Figure 5 presents results. The x-axis represents the percentile of the marginal male student and the y-axis represents the percentile of the marginal female student. Across scenarios, the effect estimates remain comparable to our main results. The one exception is the left column, 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." *The Quarterly Journal of Economics* 114 (2) (May 1): 533–575. doi:10.2307/2587016.

Areeda, Phillip E. 1996. "The Socratic Method (SM) (Lecture at Puget Sound, 1/31/90)." *Harvard Law Review* 109 (5) (March 1): 911–922. doi:10.2307/1342257.

Banks, Taunya Lovell. 1988. "Gender Bias in the Classroom." *J. Legal Educ.* 38: 137.

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

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 (5): 655–662.

Bowers, Allison L. 2000. "Women at the University of Texas School of Law: A Call for Action." *Tex. J. Women & L.* 9: 117.

Box, George EP, J. Stuart Hunter, and William Gordon Hunter. 2005. *Statistics for Experimenters: Design, Innovation, and Discovery*. Vol. 2. Wiley Online Library.

Carrell, Scott E., Marianne E. Page, and James E. West. 2010. "Sex and Science: How Professor Gender Perpetuates the Gender Gap." *The Quarterly Journal of Economics* 125 (3): 1101–1144.

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 & Social Inquiry* 29 (4): 711–769.

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

Ferguson, Ronald F. 1998. "Can Schools Narrow the Black-White Test Score Gap?" In *The Black-White Test Score Gap*, edited by Christopher Jencks and Meredith Phillips, 318–374. Washington, DC: Brookings Institution.

Frangakis, Constantine E., and Donald B. Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58 (1): 21–29.

Fredriksson, Peter, Björn Öckert, and Hessel Oosterbeek. 2012. "Long-Term Effects of Class Size." *The Quarterly Journal of Economics* (November 18). doi:10.1093/qje/qjs048. http://qje.oxfordjournals.org/content/early/2013/01/21/qje.qjs048.abstract.

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 (2) (May 1): 447–464. doi:10.1162/003465304323031049.

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

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini. 2003. "Performance in Competitive Environments: Gender Differences." *The Quarterly Journal of Economics* 118 (3) (August 1): 1049–1074. doi:10.1162/00335530360698496.

Greiner, D. James, and Donald B. Rubin. 2010. "Causal Effects of Perceived Immutable Characteristics." *Review of Economics and Statistics* 93 (3) (August 13): 775–785. doi:10.1162/REST_a_00110.

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): 1–110.

Hancock, Terence. 1999. “The Gender Difference: Validity of Standardized Admission Tests in Predicting MBA Performance.” *Journal of Education for Business* 75 (2): 91–93.

Hanushek, Eric A. 1999. “The Evidence on Class Size.” Edited by Susan E. Mayer and Paul Peterson. *Earning and Learning: How Schools Matter*: 131–68.

Henderson, William D. 2003. “LSAT, Law School Exams, and Meritocracy: The Surprising and Undertheorized Role of Test-Taking Speed, The.” *Tex. L. Rev.* 82: 975.

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

Ho, Daniel E., and Kosuke Imai. 2006. “Randomization Inference With Natural Experiments.” *Journal of the American Statistical Association* 101 (475).

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 (3): 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 (396): 945–960.

Homer, Suzanne, and Lois Schwartz. 1989. “Admitted but Not Accepted: Outsiders Take an Inside Look at Law School.” *Berkeley Women’s LJ* 5: 1.

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 (3): 669–687.

Hoxby, Caroline M. 2000. “The Effects of Class Size on Student Achievement: New Evidence from Population Variation.” *The Quarterly Journal of Economics* 115 (4): 1239–1285.

Jacobs, Jerry A. 1996. “Gender Inequality and Higher Education.” *Annual Review of Sociology* 22 (January 1): 153–185. doi:10.2307/2083428.

Jencks, Christopher, and Meredith Phillips. 1998. *The Black-white Test Score Gap*. 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 (1): 19–26.

Krueger, Alan B. 1999. “Experimental Estimates of Education Production Functions.” *The Quarterly Journal of Economics* 114 (2): 497–532.

Little, Roderick JA, and Donald B. Rubin. 2002. “Statistical Analysis with Missing Data.”

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

Miller, Diane L., and Charles E. Mitchell. 1994. “Evaluation Achievement in Mathematics: Exploring the Gender Biases of Timed Testing.” *Education* 114 (3): 436–438.

Monks, James, and Robert Schmidt. 2010. “The Impact of Class Size and Number of Students on Outcomes in Higher Education.”

Mosteller, Frederick. 1995. “The Tennessee Study of Class Size in the Early School Grades.” *The Future of Children* 5 (2) (July 1): 113–127. doi:10.2307/1602360.

———. 1999. “How Does Class Size Relate to Achievement in Schools?” *Earning and Learning: How Schools Matter*: 117–130.

Niederle, Muriel, and Lise Vesterlund. 2007. “Do Women Shy Away From Competition? Do Men Compete Too Much?” *The Quarterly Journal of Economics* 122 (3) (August 1): 1067–1101. doi:10.1162/qjec.122.3.1067.

———. 2010. “Explaining the Gender Gap in Math Test Scores: The Role of Competition.” *The Journal of Economic Perspectives* 24 (2): 129–144.

Ors, Evren, Frédéric Palomino, and Eloïc Peyrache. 2013. “Performance Gender Gap: Does Competition Matter?” *Journal of Labor Economics* 31 (3): 443–499.

Pocock, Stuart J., and Richard Simon. 1975. “Sequential Treatment Assignment with Balancing for Prognostic Factors in the Controlled Clinical Trial.” *Biometrics* 31 (1) (March 1): 103–115. doi:10.2307/2529712.

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 (1): 010107.

Rhode, Deborah L. 1993. “Missing Questions: Feminist Perspectives on Legal Education.” *Stanford Law Review* 45 (6) (July 1): 1547–1566. doi:10.2307/1229112.

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

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 (4): 327–9.

Rosser, Sue V. 1995. *Teaching the Majority: Breaking the Gender Barrier in Science, Mathematics, and Engineering.* ERIC.

Rubin, Donald B. 2006. “Causal Inference through Potential Outcomes and Principal Stratification: Application to Studies with ‘censoring’ Due to Death.” *Statistical Science* 21 (3): 299–309.

———. 2008. “For Objective Causal Inference, Design Trumps Analysis.” *The Annals of Applied Statistics* 2 (3) (September 1): 808–840. doi:10.2307/30245110.

Samida, Dexter. 2004. “The Value of Law Review Membership.” *The University of Chicago Law Review* 71 (4) (October 1): 1721–1748. doi:10.2307/1600537.

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.” *Stan. L. Rev.* 40: 1209.

Weiss, Catherine, and Louise Melling. 1988. “The Legal Education of Twenty Women.” *Stanford Law Review* 40 (5) (May 1): 1299–1369. doi:10.2307/1228867.

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

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

Yale Law Women. 2012. “Speak Up About Gender: Ten Years Later.”