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 Galloob, Paul Goldstein, Gillian Hadfield, Amalia Kessler, Jon Klick, Larry Kramer, Liz Magill, Jenny Martinez, Bernie Meyler, Alison Morantz, Deborah Rhode, 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, 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.
I. **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 that are at different levels in the same classroom. 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\(^2\) 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 even Tennessee’s experimental estimates are disputed. Hanushek (1999) argues that high attrition rates (with up to 50% of students leaving the experiment\(^3\)), noncompliance (with 10% switching from large to small classrooms), and nonresponse (with 3 to 12% not taking exams) provide reasons to doubt Tennessee’s class size effects. Quasi-experimental and observational studies are less certain whether smaller classes improve achievement and whether they can narrow the gap.\(^4\)

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. In a laboratory experiment, they randomly assign participants to compensation based on a “tournament incentive,” where only the highest performer receives payment, or payment per task completed. 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).\(^5\)

---

\(^1\) As we discuss in Section VII, we do not distinguish between achievement gaps and test score gaps for our purposes here, although the difference is conceptually important.

\(^2\) STAR stands for “Student-Teacher Achievement Ratio.”

\(^3\) See Krueger (1999) (Table 1, documenting attrition rates from 47 to 53% for students entering the experiment in kindergarten or first grade).

\(^4\) See, e.g., Fredriksson, Öckert, and Oosterbeek (2012) (finding that smaller class sizes in Sweden in the last three years of primary school have short-term cognitive benefits, as well as improving long-term labor market outcomes); Hoxby (2000) (finding no evidence of class size effects on achievement); Angrist and Lavy (1999) (finding gains to smaller classes for fourth and fifth graders, but not third graders); and Fryer and Levitt (2004) (finding only weak evidence that school inputs, such as class size, account for the black-white test score gap, as black and white students attended schools generally similar in class size).

\(^5\) 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 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 are seen to matter considerably in the legal profession, numerous scholars have examined the gender gap in law school grades, with heterogeneous findings across schools. Guinier et al. (1994, p. 96) advocate comprehensive reform of legal education to address gender disparities, emphasizing, common to calls for reform, that “small class size may be a necessary condition.” 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 family background, type of student), we leverage Stanford’s randomization into mandatory first-year courses to study the causal effect of class size. 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. 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 covariates (most importantly, ability) to identify the effect of small classes by gender.

---

6 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).

7 The closest are De Paola, Ponzo, and Scoppa (2013), who find that large class sizes in an Italian university have a larger negative effect on low-ability student in math, Krueger (1999), who briefly examines heterogeneous treatment effects in the context of Tennessee STAR, finding that “smaller classes have a larger initial effect, but smaller cumulative effect, for boys as compared to girls,” and Fredriksson, Öckert, and Oosterbeek (2012), applying a fuzzy regression discontinuity to examine class size effects in Swedish primary schools and briefly examine heterogeneity effects across gender, finding no evidence of differences.
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.\(^8\) 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 change 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.\(^9\) This is despite questions about the endurance of gains to early interventions (see Mosteller 1999, Hanushek 1999) and 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 --- unconfounded by student selection because of random assignment --- in a post-graduate professional school setting.

This Article proceeds as follows. Section II discusses the unique natural experiment that Stanford inadvertently conducted from 2001-2012. Section III describes fine-grained student and course data we collected with the help of the law school’s admissions and registrar offices. Section IV verifies random section assignment by assessing balance along a host of covariates. Section V 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 VI 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, smaller even than small sections. Section VII concludes.

II. 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.

---

\(^8\) See Krueger (1999) (Table 1, attrition rates) and Hanushek (1999) (discussing attrition and failure to sit for exams).

\(^9\) One rare study of higher education is Monks and Schmidt (2010), who note that “[o]nly a handful of studies have focused on the role that class size may play in outcomes in tertiary education.”
A. 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”\(^\text{10}\)), and the 3.4 mean requirement applied regardless of the grading option.\(^{11}\)

Beginning in Fall 2008, the law school instituted a series of pedagogical reforms. First, courses would be graded on an Honors, Pass, or no credit basis (the HP system).\(^{12}\) The required range was 30-40% Honors in a doctrinal course.\(^{13}\) The rationales for grade reform were to reduce “grade curve shopping” upper-division writing courses (not subject to a grading guideline) and to eliminate what was perceived as a falsely precise, and to many students an intimidating, numerical GPA system.\(^{14}\) 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 period of time as they had previously. Winter courses were adjusted to the quarter system.\(^{15}\) Two modifications of concern were that (i) LRW would be graded and shortened to a single term in the fall beginning Fall 2010, and (ii) in lieu of LRW for the winter and spring quarters, the school introduced an even smaller, two-quarter, simulation-based “Federal Litigation” course about the pretrial stages of a federal lawsuit. The case used in the course involved First Amendment, personal jurisdiction, and class certification issues in the context of a suit to prevent screening a documentary about abuses at a mental health facility. In Federal Litigation, students were assigned to specific sides and sets of issues, with a wide range of writing and simulation exercises (initially, they drafted one complaint, three briefs, and one

---

10 “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.
11 To be precise, students had an unlimited number of 3K options during the first-term, and two 3K options for the remainder of law school.
12 Technically, instructors could also assign a grade of “restricted credit” or “fail,” but these grade were scarcely used. Some upper-division courses, but no mandatory first-year courses that are the subject of our study, were also graded on a mandatory credit / no-credit basis.
13 The range applied to any course in which student grades were based on responses to a common prompt. For the first year, this applied to all doctrinal courses.
14 See Andy Guess, Stanford Drops Letter Grades, INSIDE HIGHER ED, June 2, 2008 (noting the argument that “shifting from the precision of letter grades to broader categories will reduce . . . pressure [on students]”); Orin Kerr, VOLOKH CONSPIRACY, Sep. 27, 2008 (“Students perceive that [an Honors/Pass system] takes pressure off them.”); see also Bloodgood et al. (2009) (finding that changing from letter to pass-fail grading at the University of Virginia medical school improved psychological well-being of students); Robins et al. (1995) (finding that pass/fail grading at the University of Michigan medical school reduced anxiety without a reduction in performance).
15 Specifically, Property, which met three times a week under the 13-week semester, would meet four times a week for the nine-week quarter. Constitutional Law was divided into a mandatory quarter course and an upper-division elective.
bench memo; delivering three, and judging one, oral arguments; taking and defending a deposition\textsuperscript{16}). 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 comprised of two small sections (i.e., roughly 60 students). When Federal Litigation was introduced, small sections were split even further into groups of roughly 18 students (10 sections per incoming class), which were further divided into legal teams generally four to five students each. The majority of Federal Litigation class meetings were held exclusively between the instructor and the legal team of four to five students. At all times, exams in doctrinal courses, on which final grades are overwhelmingly (if not exclusively) based, were graded blindly, ruling out the possibility of sheer instructor grading bias.\textsuperscript{17}

B. 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, students made matriculation decisions. Second, upon 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 across sections, 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 names: e.g., 1, 2, 3, 4, 5, 6, 6, 5, 4, 3, 2, 1, and so on (see Ho and Imai 2006). The academic index amongst Stanford students is very coarse due to range compression: for instance, the entering class of 2005 had only 7 unique values of the academic index, and the order within a stratum of an index value was random. Third, 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 (i.e., mapping the numbers 1-6 to courses) was random. Because the Associate Dean was at all times unaware of how the six numbers mapped onto specific courses and instructors, she could not match students in any way based on instructor “fit” or predicted ability to succeed in a particular small or large section.

\textsuperscript{16} Class requirements were modified over time.
\textsuperscript{17} Blind grading of course does not rule out the possibility that instructor’s may devalue “female voice” (Gilligan 1982) on exams.
Specific students characteristics were not consulted in assigning sections to courses, except for very rare circumstances.\footnote{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, involving a handful of students from 2001-11, and grouping of sections remained intact.}

Grouping students into sections, as we show in Appendix A, 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 matriculations decisions for specific students can hinge on chance factors (e.g., deferrals of admission, significant others). 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 IV 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.\footnote{Grouping and assignments into Federal Litigation sections worked comparably.}

III. 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.

<table>
<thead>
<tr>
<th>Period</th>
<th>Students</th>
<th>Instructors</th>
<th>GPA / HP Grades</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001-8</td>
<td>1,193</td>
<td>62</td>
<td>9,539</td>
<td>3.46</td>
</tr>
<tr>
<td>2008-11</td>
<td>704</td>
<td>58</td>
<td>6,150</td>
<td>0.42</td>
</tr>
</tbody>
</table>

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.
Figure 1: Histogram of class sizes for mandatory first-year courses. The left bin represents Federal Litigation, introduced in 2008, which had a median enrollment of 18 students. The modal bin represents small sections for a doctrinal course, with a median enrollment of 29 students. The right bins represent large sections, with a median enrollment of 58 students.

Figure 1 presents the distribution of class sizes for mandatory first-year courses. As expected, given the small section assignments, distribution is distinctly bimodal pre-2008 (before Federal Litigation is introduced). Courses taught exclusively to small sections ranged from 26 to 31 enrolled students, with a median enrollment of 29. With the exception of one year (two large sections), courses taught to large sections (i.e., consolidated small sections) range from 54 to 62 students, with a median of 58 students.

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

<table>
<thead>
<tr>
<th>Academic background</th>
<th>Men</th>
<th>Women</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Law School Admissions</td>
<td>169.0</td>
<td>168.9</td>
<td>4.2</td>
</tr>
<tr>
<td>Test score (LSAT)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Undergraduate degree GPA</td>
<td>3.81</td>
<td>3.82</td>
<td>0.19</td>
</tr>
<tr>
<td>Academic index (LSAC)</td>
<td>3.42</td>
<td>3.41</td>
<td>0.15</td>
</tr>
<tr>
<td>Master’s degree</td>
<td>0.18</td>
<td>0.12</td>
<td>0.36</td>
</tr>
<tr>
<td>Ph.D.</td>
<td>0.05</td>
<td>0.03</td>
<td>0.19</td>
</tr>
<tr>
<td>Age</td>
<td>24.6</td>
<td>23.8</td>
<td>2.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>0.59</td>
<td>0.51</td>
<td>0.50</td>
</tr>
<tr>
<td>Latino</td>
<td>0.12</td>
<td>0.12</td>
<td>0.32</td>
</tr>
<tr>
<td>Asian-American</td>
<td>0.08</td>
<td>0.15</td>
<td>0.32</td>
</tr>
<tr>
<td>African-American</td>
<td>0.08</td>
<td>0.11</td>
<td>0.29</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Undergraduate institution</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford</td>
<td>0.10</td>
<td>0.11</td>
<td>0.30</td>
</tr>
<tr>
<td>Harvard</td>
<td>0.06</td>
<td>0.07</td>
<td>0.25</td>
</tr>
<tr>
<td>Yale</td>
<td>0.07</td>
<td>0.07</td>
<td>0.25</td>
</tr>
<tr>
<td>Berkeley</td>
<td>0.03</td>
<td>0.04</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Figure 2 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).²⁰ The slight demographic differences from Table 2 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.

²⁰Because of substantial overlap between entering characteristics of men and women, the gender gap also persists when preprocessing via matching to reduce the degree of extrapolation (see Ho et al. 2007). Matching women and men on LSAT scores, undergraduate GPA, academic index, and age (via the propensity score estimated by logistic regression) and imposing exact matches on ethnicity, Master’s degrees, doctoral degrees, professional degrees, law school instructor, course, and undergraduate institution results in a matched sample of 1,352 student enrollments. Fitting a linear model with all these explanatory variables on this matched sample still results in a highly statistically significant difference in law school performance between men and women. Note that because gender is a (largely) immutable characteristic, this should not be interpreted as a causal effect of gender per se --- gender plausibly affects many dimensions of life and it is not clear what variables are “pretreatment” (see Greiner and Rubin 2010, Holland 1986). For the treatment of class size, all these variables (including gender) are pretreatment covariates, as they temporally occur before assignment to course sections.
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, e.g., Henderson 2003 (“[A] wide variety of academic and career opportunities . . . often hinge on relatively small variations in law school grades’’)). To illustrate the gap’s substantive importance, we examine clerkship and private sector employment data. We use data on 525 clerkship applications by Stanford students from 2003-2008 and examine the correlation between successful clerkship placements and GPA at the time of application. Grades and clerkship placements are highly correlated: a 0.05 GPA increase from 3.6 to 3.65, for instance, is associated with a 6% (statistically significant) increase in the probability of securing a federal appellate clerkship.\footnote{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 (358 students).} 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: a 0.05 GPA increase from 3.25 to 3.3 is associated by a nearly 5% increase in the callback rate.\footnote{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.}

It is worth noting that law firms appear to have become even more grade-sensitive since 2008.\footnote{See, e.g., Jacqueline Bell, Law School Grads Face Tight Job Market, Law360, Aug. 6, 2008.} Indeed, the degree of scrutiny by employers to small numerical differences in GPAs was precisely one of the reasons for grade reform. 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.

IV. RANDOMIZATION CHECKS

Although there are strong reasons to believe that the assignment mechanism of sections to specific courses (and the grouping of sections) was random, we perform a series of randomization checks to verify the crucial assumption. Because large sections are comprised of consolidated small sections, we check here for whether the six small sections in any given year of admission exhibit any imbalance on key covariates.

Figure 3 plots the year of admission on the x-axis against 12 key 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. If small sections are randomly assigned, the observed mean (or proportion) should generally fall within the grey interval. Nearly all do.

As the bottom five panels in the left column show, the Associate Dean’s additional demographic shuffling balances gender and ethnicity more than would be expected by chance alone. The observed proportion of women and minorities line up more closely along the class mean than would be expected under pure randomization. Other covariates, however, approximate the randomization distribution. Although there are a few isolated sections that fall outside of the 95% randomization interval, such deviations are far fewer than we would expect by Type I error alone: under randomization, we would expect roughly 40 such deviations \([= 0.05 \alpha\text{-level} \times 6 \text{ sections/entering class} \times 11 \text{ entering classes} \times 12 \text{ covariates}]\).

In short, the results in Figure 3 strongly confirm that small sections were effectively randomized. In Appendix A, we show that the Associate Dean’s grouping is essentially a form of (stratified) block randomization, thereby improving balance on gender and ethnicity beyond what would be expected under 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 3: 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.
V. 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 VI examines the post-2008 period separately.

Figure 4 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).

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.

<table>
<thead>
<tr>
<th></th>
<th>Large</th>
<th>Small</th>
<th>Difference (small-large)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>3.488</td>
<td>3.461</td>
<td>-0.026</td>
</tr>
<tr>
<td>Women</td>
<td>3.433</td>
<td>3.454</td>
<td>0.021</td>
</tr>
<tr>
<td>Difference (men-women)</td>
<td>0.054**</td>
<td>0.007</td>
<td><strong>0.047</strong></td>
</tr>
</tbody>
</table>

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 α = 0.05.
To more rigorously assess the class size gender effect, we pursue 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 (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 \( \alpha_s, \eta_i, \) and \( \kappa_c \) 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 (\( \alpha \)) control for gender, age, LSAT score, undergraduate GPA, and any other student-specific entering characteristics. The parameter of interest (\( \tau \)) 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 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 difference-in-differences estimation, adding only fixed effects for ethnicity compared to Table 3. The gender gap narrows somewhat 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 cause men to diminish performance by 0.03 GPA points. These results provide considerable evidence that small classes diminish the gender gap existing in large sections.
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 \( \alpha \)-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.\(^{24}\) 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.

**VI. 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 under the HP system, the gender gap vanishes entirely. 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, the simulation-based writing course.

---

\(^{24}\) Even if that were the case, it would suggest that women are able to better discern their relative abilities in small sections than in large sections.
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 all but disappeared. We cannot reject the null that small sections have no differential effect for women. Federal Litigation, however, appears to result in systematically more women earning Honors, robust to the full set of fixed effects presented by Models B-D. Holding constant student and instructor attributes, Federal Litigation (compared to a large section) increases women’s probability of earning Honors by 0.18, compared to 0.08 for men. Even though the differential grading guideline should be expected to raise the probability of Honors, it 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 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 2, 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. Figure 5 plots the time series by gender of the academic index, LSAT scores, and undergraduate GPAs. The dashed lines represent the means for women and the grey lines represent the means for men, with bands plotting the 95% probability interval. The vertical lines indicate the year of grade reform, and there were no sharp changes. Qualifications between women and men were substantively comparable over the entire observation period.

Third, closing the gender gap was also not likely due to a spillover effect from Federal Litigation. In Federal Litigation, students were given feedback throughout the course, thereby potentially generating confidence that may have benefitted women even in other courses.\textsuperscript{25} The problem with this spillover explanation is that 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.\textsuperscript{26}

\textsuperscript{25}Psychological research documents sharp differences along gender lines in overconfidence.

\textsuperscript{26}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.
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 (Rhode 1993, p. 1563 (drawing on feminist theory to call for less reliance on “large lectures or quasi-Socratic discussion [and] greater emphasis on [inter alia] simulations [with] interactive, experiential learning”), (b) has no final exam under timed conditions (see Miller and Mitchell 1994 (finding timed exams increase gender gaps)), (c) provides substantial feedback throughout the class (Rhode 1993, p. 1563 (concluding that “feedback on a more regular basis” would address the gender gap in law school), and (d) is the smallest mandatory first-year class, with effectively only 4-5 students for most class meetings. Indeed, Federal Litigation’s 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. This strongly suggests that distinct pedagogical techniques available in small classes matter.

VII. 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 provide 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 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 other 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., three oral arguments) are infeasible in a larger class. Consistent with the evidence that women express and exhibit preferences for direct

---

27 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.

28 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.
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 would also be consistent with evidence outside of the law school context, finding that the science gender gap can be reduced by incorporating interactive engagement techniques (Lorenzo, Crouch, and Mazur 2006, Rosser 1995).29

We conclude by discussing some caveats on interpretation. First, while a principal strength of our study is randomization of students to small sections, instructors are not necessarily randomly assigned. Instead, some deference is paid to instructor preferences of section size; instructors with small section preferences may simply teach differently. That of course 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.30

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, with high internal validity, 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. To provide a sense of the magnitude of resource costs, Figure 6 plots the distribution of typical first-year sections across 201 law schools. The vast majority of law schools enroll sections that are far larger than Stanford’s (denoted by the red vertical line). The resource cost to move toward anything close to the small section in our experiment may hence be considerable. On the other hand, Stanford

29 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).

30 We also cannot reject the hypothesis that the gender effect is the same for (a) instructors teaching both small and large sections, and (b) instructors teaching exclusively small or large sections.
managed to create Federal Litigation without substantial additional resource cost, as all the instructors were already teaching at the school. The school simply shifted existing instructors to create Federal Litigation sections\textsuperscript{31} --- and instructors in turn divided most class meetings to meet only with the 4-5 student legal team (in lieu of plenary class meetings) --- suggesting that not all reductions in class sizes need be a drain on resources.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure6}
\caption{Typical first-year section size for full-time students in 201 law schools accredited by the American Bar Association and Law School Admissions Council (ABA-LSAC). The vertical line plots Stanford’s small section size of roughly 30 students. Source: ABA-LSAC (2013).}
\end{figure}

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 may explain why findings vary considerably as to whether a gender gap exists in law school. The key now is how to address it when it does. And we show that pedagogical policy may have a crucial role to play.

\textsuperscript{31} Six of the Federal Litigation instructors were shifted from teaching LRW (which was shortened to the fall term in 2009) to teaching Federal Litigation in the winter and spring terms; four of the instructors were already teaching Federal Litigation for second and third year students.
APPENDIX

A. Stratified Block Randomization

As our randomization checks reveal, 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. If the Associate Dean is grouping students in a way that is comparable to block randomization, the six observed means should follow the mode of the latter distribution.

Figure 7 plots results. The red 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. Red dashes track the latter extraordinarily well. Based on this evidence, the Associate Dean has expressed her willingness to simplify the grouping procedure using stratified block randomization. A further refinement would be a form of randomization minimization (Pocock and Simon 1975), which we also implement. Our results suggest, however, that minimization yields few improvements in approximating the Associate Dean’s observed grouping relative to stratified block randomization.

---

32 The only manual adjustment to stratified block randomization is that the total section size can deviate slightly due to strata that are indivisible by six.
Figure 7: 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.

B. Section Assignment Violating Identification Assumptions

Even were assignment non-random, confounded assignment mechanisms that would artificially generate the gender effects we detect 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 below:

<table>
<thead>
<tr>
<th></th>
<th>Small Section A</th>
<th>Small Section B</th>
<th>Large Section</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H/N</td>
<td>Prop.</td>
<td>H/N</td>
</tr>
<tr>
<td>High achieving women / all women</td>
<td>3/4</td>
<td>0.75</td>
<td>8/20</td>
</tr>
<tr>
<td>High achieving men / all men</td>
<td>15/20</td>
<td>0.75</td>
<td>6/15</td>
</tr>
</tbody>
</table>

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-referenced 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 a total of four women and 15 high achieving men out of a total of 20 men. These data reveal no gender gaps in small sections (and would yield 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, the aim is to keep all small sections as representative of the incoming class as possible, including gender and ability.

C. Sensitivity to Nonresponse

We now investigate whether nonresponse affects our estimates for the effect of small sections on women from 2001-08. 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. (Because the grading election was eliminated under the HP system, 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),

---

33 The table is of course just an example of Simpson’s paradox.
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 therefore take several different 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 $\tau$, 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.

---

34 Even when practice exams are available, the typical feedback is in the form of a model answer, providing only a sense of the upper bound in a class.

35 Of course, a large number of courses taken on a credit/no-credit basis may also be viewed skeptically by potential employers. But elections for that grading option rarely occurred.
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.\(^{36}\) (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 [\(= 7 \text{ years } \times 6 \text{ sections/year}\)]) and to remove the grade for higher-performing female students in large sections (affecting 104 female students), and to examine how sensitive our estimates of \(\tau\) are. In the worst-case scenario, the marginal male

\(^{36}\) In the parlance of principal stratification, these students are “always-takers.” See Rubin 2006.
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 the best female student. As that seems unrealistic, we vary the percentile of the marginal student from 0 to 50% for men (i.e., focusing only on male students below the median male student in the course), and from 50 to 100% for women, removing marginal male student grades from every small section and marginal female student grades from every large section.

Figure 8 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 in Table 4. The one exception is the left column, when the marginal male student is the worst-performing student in the section: point estimates of $\tau$ 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


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


