The Effect of Teacher Gender on Students of Differing Ability: Evidence from a Randomized Experiment

Niklaus Julius

October 15, 2019

Abstract

Using data from a well-executed field experiment, I study heterogeneity in the impact of teacher gender on students of differing ability by implementing the Conditional Average Treatment Effect estimator of Abrevaya et al. (2015). I find that female students see a limited degree of heterogeneity in the effect of teacher gender, while male students see almost none. My results suggest that estimates of the average effect of teacher gender do not mask significant heterogeneity, and can be used to motivate policy without exacerbating inequality between students of different abilities.

Keywords: teacher gender, student achievement, conditional average treatment effect

JEL Codes: I21, I24, I26, J24

1 Introduction

Over the past half-century, the U.S. has seen a remarkable evolution of the gender gap in educational outcomes. For the birth cohort of 1950, the U.S. high school graduation rate was 85% for men and women, and the proportion with at least some college education by age 35 was 55% (50%) for men (women) (Autor and Wasserman, 2013). For the birth cohort of 1975, the male high school graduation rate was 88%, while for women it was 91%. College attendance rates display the same pattern - for the birth cohort of 1975, women were approximately 17% more likely to attend college than males, and almost 23% more likely to complete a four-year degree (Autor and Wasserman, 2013). This trend is not unique to the United States - the majority of OECD countries have experienced qualitatively similar reversals of the traditional gender gaps in education (Vincent-Lancrin, 2008; Fortin et al., 2015).

In contrast with the marked change in the relative performance of women in education overall, the shifts in math performance have been muted. Men continue to outnumber women in most science and engineering fields, both in education and the labor force (Hill et al., 2010). While the gap between men and women at the high end of the distribution in math has shrunk considerably, the ratio of men to women in the top decile is approximately 2:1, and women are slightly overrepresented in the bottom decile (Hedges and Nowell, 1995). Recent research suggests that the underlying distribution of math ability is equal (particularly in the top decile) (Joensen and Nielsen, 2016; Vesterlund and Niederle, 2010), raising the question of why the gender gap in math performance has not closed.

One potential explanation for the 'sticky' gender gap in math performance is that teacher gender may have different effects on the outcomes of different students. There is some empirical support for this hypothesis - in a recent working paper, Cappelen et al. (2019) find evidence of a gender bias that affects low-performing males, while other scholars have found evidence that female teachers negatively impact the math outcomes of female students (Antecol et al., 2015; Beilock et al., 2010).

To date, the study of how different students are impacted by different teachers has mostly focused on demographic characteristics. This amounts to estimating average treatment effects for different population subgroups, which can potentially mask significant within-group heterogeneity (Bitler et al., 2006). I address this gap in the literature using data from a field experiment conducted by Mathematica Policy Research to evaluate the Teach for America program. Exploiting the random assignment of students to teachers, I estimate the Conditional Average Treatment Effect (henceforth CATE) of being assigned to a female teacher on the math and reading test scores of male and female students in primary school, conditioning on pre-treatment test scores as a proxy for ability. My estimates shed light on how the effect of being assigned a female teacher changes with both gender and ability - changes that may be significant, particularly if a bias such as that found by Cappelen et al. (2019) is present. Understanding how teacher gender impacts students of differing ability and gender also has important implications for how teachers should be assigned to students.

The fact that my data comes from a well-executed randomized experiment and includes a rich set of covariates allows me to deploy powerful non-parametric techniques that require the relatively strong assumption of unconfoundedness, rather than imposing functional form restrictions. While my sample is not representative of the U.S. student population, it is representative of the most disadvantaged students and schools - a subset of particular importance to policymakers, as students in these schools are least likely to continue on to higher education, and therefore most likely to face difficulties that challenge individuals without a college education in modern society[1].

I find that there is limited heterogeneity in the effect of teacher gender on students of different ability. Female teachers appear to have a small positive impact on most male students in math, while having a statistically insignificant negative effect on the lowest-ability

[1] Men with less than a four-year college education have seen a dramatic reduction in real income over the last decade (Autor and Wasserman, 2013), are less likely to enter the labor force (Krueger, 2017), and face increased risk of poverty, physical, and mental health problems. The prospects for women with less than a four-year college education are significantly worse than for women with more education, but are less grim than those for men.

female students in math. For the majority of students in my sample, assignment to a female teacher has no statistically significant effect on math or reading outcomes. Outside the very bottom of the ability distribution, the effect of teacher gender does not significant differ based on the gender of the student. While my results echo much of the previous economics literature on teacher gender effects by finding no significant *average* effect of teacher gender on students, they also suggest that those average effect estimates do not mask heterogeneity (at least for primary school students) and can thus be used to motivate policy without potentially exacerbating inequality.

The remainder of the paper is organized as follows. Section 2 reviews related literature. Section 3 provides a brief overview of the institutional background of the experiment, the experiment itsef, and the resulting data. Section 4 briefly introduces the theoretical framework for the CATE estimator and sets out my estimation strategy. Section 5 presents my main results. Section 6 discusses my results, possible mechanisms, and policy implications. Finally, Section 7 concludes.

2 Literature Review

This paper contributes directly to the literature that studies dynamics based on demographic features in educational settings. This literature has two distinct strands. One focuses on estimating the effect of demographic matching, but leaves the question of the mechanism unanswered. The other strand considers potential mechanisms, and has to date focused primarily on biases in teacher grading behavior.

2.1 Reduced-Form Effects of Demographic Matching

The starting point for this literature in economics was Ehrenberg et al. (1995). Ehrenberg et al. used data from the National Educational Longitudinal Study of 1988 (henceforth the NELS:88) and found that demographic matching had little association with how much students learned, but a significant impact on teacher perceptions of the student. Dee (2004), using data from Tennessee's Project STAR experiment to investigate the effect of assignment to a teacher of the same race on student outcomes, found a significant positive effect on math and reading achievement for both black and white students. Dee (2005) used NELS:88 data, like Ehrenberg et al., but exploited a unique feature of the data - namely that for each student, two teachers were surveyed - to control for student-specific fixed effects, and found that dynamics between student and teachers based on gender, race, and ethnic background had consistently large effects on teacher perceptions of student performance. Interestingly, the effects associated with race and ethnicity appeared to differ significantly as a function of student socioeconomic status and geographic location. Dee (2007) goes further, using the same NELS:88 data, studying objective outcomes of gender matching (in particular, test scores), and finds that assignment to a same-gender teacher significantly improves achievement for students of both genders, as well as improving teacher perceptions of that student and student engagement in that teacher's subject.

The NELS:88 data consists of a representative study of 8th grade students in 1988. Other scholars have extended the investigation to consider different parts of the educational system. Bettinger and Long (2005), and Hoffmann and Oreopoulos (2009), used administrative data from the university students[2] to study the effects of female instructors on female students in university education. Hoffmann and Oreopoulos found that assignment to a same-sex instructor boosted student's relative performance and their likelihood of course completion, but found that effects on upper-year course selection were insignificant, while Bettinger and Long (2005) found very mixed results - their primary conclusion is that the effect of gender matching changes dramatically based on the subject being taught - for instance, the effects are positive and strong in mathematics and statistics, while being weak for economics.

[2] Bettinger and Long uses administrative data on full-time undergraduate students in Ohio during 1998 and 1999. Hoffmann and Oreopoulos use administrative data from the University of Toronto's Arts and Science Faculty.

They also add to the growing number of studies that find largely negligble effects of gender matching on male students. Using data from the U.S. Air Force Academy, Carrell et al. (2010) found limited impacts of gender matching on male students, but found significant positive effects on female student performance in math and science classes. In addition, Carrell et al. found significant effects on the likelihood of female students taking future math or science courses, and on their likelihood of graduating with a STEM degree. Fairlie et al. (2014), using data from a large and diverse community college, find that assignment to an underrepresented minority instructor reduces the performance gap between white and underrepresented minority students significantly, while also confirming that white students perform better when matched with instructors of the same race/ethnicity.

Moving into postgraduate education, Neumark and Gardecki (1998) studies the influence of role models and the effectiveness of mentoring by female faculty on the success of female graduate students, specifically in economics. Their data was obtained by directly surveying all Ph.D.-granting institutions in the U.S. for information on female Ph.D. graduates from 1973 to 1996. Their primary focus is on quality of first job placement, and they find no evidence that outcomes for female graduate students are improved by adding female faculty members, or by having a female dissertation chair. There is limited evidence of positive effects on other outcomes, such as time to graduate and probability of graduating, but this evidence does not fully survive robustness checks. Hilmer and Hilmer (2007) undertakes a somewhat more focused analysis, looking at the effect of same-gender mentoring (where the mentor is identified as the dissertation advisor) on initial job placements and early career publishing outcomes, for a sample of individuals that graduated from a top-30 program between 1990 and 1994. Hilmer and Hilmer find female students with male advisors are significantly more likely to accept research-oriented first jobs than male students with male advisors, while finding little evidence of an effect on early career publication success.

In recent years, scholars have begun to exploit newer data to shed further light on the effects of demographic matching. Egalite et al. (2015) uses a large administrative dataset that studies a huge number of students over time in the Florida public school system, and find small but significant effects for racial/ethnic matching between students and their teachers, with an exception for Hispanic students. Winters et al. (2013), also using Florida data, studies the impact of a female instructor (not a same-sex instructor) on student achievement in math and reading, finding positive effects for both male and female students (although the effect is stronger for female students in math), with the effects occurring primarily between the 6th and 10th grade levels.

One implication of Winters et al. (2013) is that gender matching might not have an effect on student outcomes in primary school. There remains, however, some uncertainty regarding when children begin to understand or internalize gender stereotypes. For instance, Ambady et al. (2001) suggests that this begins to occur around 10 years of age (5th grade), while Steele (2003) finds effects beginning at 7 years of age (2nd grade). The only study of which I am aware that directly studies the effect of gender matching prior to 6th grade is Antecol et al. (2015), who use data from the National Evaluation of Teach for America and find that female teachers have a negative impact on female student achievement in math, while having no effect on male student achievement, or on female student achievement in reading. Furthermore, Antecol et al. suggest that the negative effect on female math achievement is consistent with the math anxiety hypothesis.

Examination of the effect of teacher gender on non-cognitive outcomes is still a nascent strand of the literature. Gong et al. (2018) uses a representative survey of middle-school teachers and students in China. By focusing on those schools where teacher assignment is random, they find that female teachers raise female students test scores as well as their mental status and social acclimation, relative to male students.

2.2 Potential Mechanisms

The most commonly cited theory as to the mechanism underlying demographic matching effects is that demographically similar teachers serve as role models for their students (Hess and L. Leal, 1997). Exposure to demographically similar teachers may increase student motivation and ambition (Maria Villegas et al., 2012), or reduce the effect of stereotype threat (Steele, 1997; Beilock et al., 2010). A second theory proposes that demographic matching influences the expectations teachers form for their students, and that these expectations have material influence on relevant student outcomes. Prior research has found that teacher expectations are affected by demographic matching (Gershenson et al., 2016; Ouazad, 2014; Ouazad and Page, 2012). The latter claim (that teacher expectations influence student outcomes) is largely uncontroversial, but Mechtenberg (2009) develops a model of cheap-talk grading which can generate the stylized achievement gaps observed empirically. Finally, it could be that a teacher is less likely to exhibit unconscious biases against demographically similar students, either directly through biased grading (as in Terrier (2016)) or through reduced likelihood of extreme responses to student misbehavior, which has been found to be affected by demographically based biases (Downey and Pribesh, 2004; Holt and Gershenson, 2017).

Determining which of these theories, if any, are correct poses a number of empirical difficulties. The data necessary to distinguish between potential mechanisms is difficult to acquire, although in recent years some progress has been made along these lines. Lavy (2008) is one of the first studies to consider a particular mechanism - in particular, that of biased grading behavior. Using data on high-school students in Israel that included blind and non-blind test scores, Lavy found that male students face discrimination (i.e. teachers are biased against male students in their grading behavior) in each of the subjects considered (four in the humanities, four in the sciences, and mathematics). In a followup, Lavy and Sand (2018) confirm that exposure to a teacher who exhibits biased grading behavior benefits those students who are favored by the teacher in terms of objective achievement outcomes. Terrier (2016), using data on French middle-school students, also finds a bias against male students in math, concluding that "gender-biased grading accounts for 21 percent of boys falling behind girls in math during middle school."

Ouazad and Page (2012) report results from an interesting experiment in which British students in grade 8 were given the opportunity to bet an endowment on their future test scores. Some students would be graded by an anonymous external examiner, and others by their teacher. Ouazad and Page found that student choices do in fact reflect beliefs about teacher biases, but did not find a gender-matching effect. In fact, they found that students systematically invested more when taught by male teachers[3].

Other potential mechanisms have also been considered. Carlana (2019), for instance, uses the Gender-Science Implicit Association Test to measure teacher biases, and then checks whether those biases affect student achievement using data from Italian schools. She finds that the gender gap in math performance increases substantially when students are assigned to teachers with strong gender stereotypes, but finds no similar effect in literature performance. Bassi et al. (2018) uses videotapes of teachers in schools in Chile, and identifies specific teacher behaviors that serve as proxies for the level of attention teachers pay to a specific student. They find that teachers pay more attention to - and interact more favorably with - boys than with girls, and that this 'attention gap' is correlated with the gender gap in math scores on Chile's national standardized test.

[3] It is worth nothing that Ouazad and Page also finds that these beliefs are partially incorrect. In particular, female students invested more with male teachers despite the fact that female teachers were biased in favor of females. Ouazad and Page are unable to determine why this is, but hypothesize that female students either misperceive male teachers biases, or react to perceived biases against them by investing more.

3 Data

3.1 The National Evaluation of Teach for America

The data I use comes from the Mathematica Policy Research, Inc (henceforth MPR) National Evaluation of Teach for America (henceforth NETFA) Public Use File[4]. The NETFA was a field experiement conducted in elementary schools from six regions of the United States between 2001 and 2003. The full study consists of a pilot study, conducted in Baltimore during the 2001-2002 academic year, and a followup full-scale study conducted in Chicago, Los Angeles, Houston, New Orleans, and the Mississippi Delta during the 2002-2003 academic year. From each region except the Mississippi Delta, the participating schools come from a single school district. In the Mississippi Delta, the participating schools come from one of two school districts.

In each school district, schools that had at least one TFA teacher and at least one non-TFA teacher assigned to teach a class in the same grade were considered 'eligible' for the experiment. From the pool of eligible school-grade combinations, MPR selected a random sample to form an experimental group that was representative of the schools where TFA teachers tended to teach at the time[5]. If a school-grade combination was selected for inclusion in the experiment, students entering that school and grade were randomly assigned to the teachers allocated to that school and grade. Throughout the experimental year, MPR performed roster checks to enforce original classroom assignments.

After the random assignment to classrooms, but before the school year began, students in experimental classrooms took math and reading tests based on the last school grade they had completed, which I will refer to as pre-treatment tests. At the end of the school year (post-treatment), students again took math and reading tests based on the school grade they had just completed. For the vast majority of the students in the sample, the pre- and

[4] https://www.mathematica-mpr.com/-/media/publications/data-sets/2017/tfapublicuse.zip

[5] The Teach for America program has expanded significantly since the experiment. The sample is likely not representative of 'TFA schools' today.

post-treatment tests were the grade-appropriate Iowa Test of Basic Skills (ITBS). A small group of students took their tests in Spanish - for these students, the test was the Logramos test. Both tests are published by the same organization (Riverside Publishing), although they are normed relative to different groups.

The original purpose of the NETFA experiment was to evaluate the effectiveness of the Teach for America program. As a result, the sample is not representative of the U.S. school population - it is representative of the population of disadvantaged schools in high-poverty areas. While this prevents my results from generalizing to the broader school population, the students served by these schools are a subset of the student population on which policymakers have focused in the past.

3.2 Sample Statistics

The NETFA data includes detailed information on student and teacher characteristics. For students, I have class type (bilingual/monolingual), student demographic characteristics, class size, and math/reading scores both before and after treatment. For teachers[6], I have demographic characteristics, type of teacher certification (nontraditional/traditional), and years of experience. In addition, some teachers completed a survey that provides information about their teaching practices, their educational background, and their career goals. However, the combination of the relatively small sample and a significant number of missing survey answers renders the use of this additional information problematic. In addition to the baseline data, I construct a classroom-level indicator variable for the presence of at least one disruptive student[7].

The test score variables deserve some further discussion. The data does not contain

[6] Seven classrooms experienced teacher turnover during the experimental year. Following Antecol et al. (2015), I code the teacher as being the first teacher without missing data. In all but one case, this is equivalent to the longest-serving teacher.

[7] I use disciplinary data to proxy for this. Specifically, if a class contained at least one student who was suspended or expelled during the course of the school year, I code that classroom as having been disrupted. I should note, however, that some classes contained students that are not part of the research sample. I cannot be certain that a class coded as not disrupted did not contain a disruptive student, even if no such student appears in the data.

traditional test scores (percent of questions correctly answered). Instead, I have raw counts for number of correctly answered questions and number of questions attempted, and a battery of transformed scores. The transformed scores include standardized score, grade equivalent, national percentile rank, and normal curve equivalent scores. For my investigation, I use normal curve equivalent scores as both pre-treatment conditioning and post-treatment outcome variables. The primary reason for this choice is that normal curve equivalent scores have the same equal-interval property that a z-score does, which is critical for estimation techniques that average outcomes together as mine does. Normal curve equivalent (NCE) scores are defined as functions of the standard score (ss):

$$NCE(ss) = 50 + 21.063 \times ss$$

The choice of 21.063 as the multiplier ensures that, if the underlying standard scores are normally distributed, then a percentile rank of 1, 50, or 99 corresponds to a normal curve equivalent score of 1, 50, or 99 respectively. Close to 50, normal curve equivalent scores change more slowly than percentile ranks, while close to 1 or 99, they change much more rapidly[8].

Some students in the sample have raw scores (number of correct answers) of 99. These scores are invalid - the highest possible raw score in the sample is 44 in reading and 50 in math (Penner, 2016). Approximately 19 (21) percent of the initial math (reading) sample is lost due to students with missing or invalid data. This is a slightly larger loss than Antecol et al. (2015) because they retained invalid test scores in their main specification[9].

Table 1 reports summary statistics for the variables of interest. Note that the math estimation sample and the reading estimation sample are not identical. In general, this is because students who recorded an invalid test score in math or reading did not always record

[8] If the underlying test scores are normally distributed, a percentile rank between 89 and 95 will be transformed into a normal curve equivalent between 75.8 and 84.6. A percentile rank between 40 and 59 will be transformed into a normal curve equivalent between 44.7 and 54.8.

[9] In a supplementary specification, Antecol et al. (2015) removed the invalid scores and did not see a large change in their results.

an invalid test score in both subjects. In the interests of dropping as little data as possible, I retain students with invalid test scores in the 'wrong' subject when estimating the CATE for math or reading outcomes. This implicitly assumes that a student's propensity to record an invalid score is independent of whether they were taking a reading or math test, which seems plausible.

Table 2 reports the results of tests for mean differences between the full sample and the two estimation samples. I find very similar results to Antecol et al. (2015) in these tests. Sample attrition appears to be largely at random.

Table 1: Descriptive Statistics

	Definition	n=1938 Full Sample	n=1596 Math Sample	n=1551 Reading Sample
Student Characteristics				
Female	1 if student is female, 0 otherwise	0.49 (0.50)	0.49 (0.50)	0.50 (0.50)
Black	1 if student is non-Hispanic black, 0 otherwise	0.67 (0.47)	0.66 (0.48)	0.70 (0.46)
Hispanic	1 if student is Hispanic, 0 otherwise	0.26 (0.44)	0.28 (0.45)	0.24 (0.43)
Class Size	Number of students in the classroom at the end of the experiment	25.1 (5.6)	24.9 (5.5)	25.2 (5.6)
Pre-Treatment Math	Normal Curve Equivalent (NCE) score on math pre-test	29.7 (18.6)	31.2 (18.2)	29.4 (17.4)
Pre-Treatment Reading	Normal Curve Equivalent (NCE) score on reading pre-test	28.8 (19.3)	29.5 (19.4)	29.9 (18.4)
Disrupted Class	1 if student was in the same class as another student who was suspended or expelled	0.45 (0.50)	0.46 (0.50)	0.47 (0.50)
Teacher Characteristics				
Female	1 if teacher is female, 0 otherwise	0.76 (0.43)	0.77 (0.42)	0.76 (0.43)
Black	1 if teacher is non-Hispanic black, 0 otherwise	0.50 (0.50)	0.48 (0.50)	0.51 (0.50)
Hispanic	1 if teacher is Hispanic, 0 otherwise	0.09 (0.29)	0.10 (0.31)	0.08 (0.28)
TFA	1 if the teacher is a TFA teacher, 0 otherwise	0.44 (0.50)	0.43 (0.50)	0.44 (0.50)
Certification	1 if the teacher has a traditional teaching certification, 0 otherwise	0.53 (0.50)	0.56 (0.50)	0.53 (0.50)
Experience	Years of teaching experience	6.42 (8.5)	6.2 (8.0)	6.19 (8.0)

Table 2: Mean Differences between Full and Estimation Samples

	Full vs Math Estimation	Full vs Reading Estimation
Student Characteristics		
Female	-0.003	-0.007
Black	0.015	-0.026*
Hispanic	-0.023*	0.021†
Class Size	0.233	-0.096†
Pre-Treatment Math	-1.579*	0.248
Pre-Treatment Reading	-0.720	-1.122*
Disrupted Class	-0.014	-0.024†
Teacher Characteristics		
Female	-0.005	0.005
Black	0.018	-0.008
Hispanic	-0.001	0.010
TFA	0.006	-0.007
Certification	-0.024†	0.009
Experience	-0.024*	0.238

* denotes significance at the 5% level

† denotes significance at the 10% level

While there are some significant differences in means between the full and estimation samples, most are quantitatively small or only marginally significant. The only exceptions are in Pre-Treatment Math and Reading scores - and this is entirely due to the removal of invalid test scores[10].

Contrasting the estimation samples with only those students who have invalid test scores tells a somewhat different story. Black students are slightly more likely than average to have recorded an invalid math score, while being slightly less likely to record an invalid reading score. Hispanic students display the reverse pattern - they are slightly more likely to record an invalid reading score, and less likely to record an invalid math score. Finally, there is a

[10] Invalid raw scores of 99 were coded as normal curve equivalent scores of 0. Thus, removal of invalid scores will mechanically drive mean pre-treatment test scores up.

statistically significant difference in the mean class size between the math estimation sample and the sample of students with invalid math scores, suggesting that students in larger classes were slightly more likely to record an invalid score in math. These differences remain quantitatively small, and do not significantly impact the generalizability of my findings.

Since I will be estimating treatment effects conditional on pre-treatment test scores, it is worth looking at the distribution of those scores in the data. Figure 1 presents histograms of the pre-treatment math and reading scores across the relevant estimation samples. The red dashed line indicates the 90th quantile of the pre-treatment test score distribution for each sample.

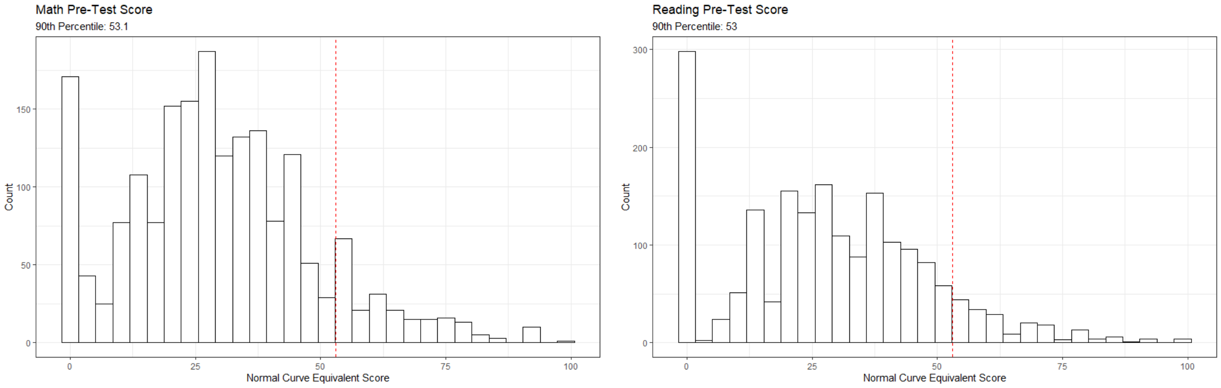


Figure 1: Pre-Treatment Test Score Distribution

If I were to estimate an average effect using pre-treatment test scores as a control, this uneven distribution would matter only in that it is informative as to how generalizable the resulting estimates are. However, since my estimates will rely on kernel-based local averaging of the pre-treatment test score, the relative lack of data in the upper half of the pre-treatment test score distribution has a direct impact on the variance of my estimates.

4 Estimation Strategy

4.1 Parameters of Interest

The parameter of interest in this investigation is the Conditional Average Treatment Effect (CATE). The CATE is defined as the value of the familiar Average Treatment Effect (ATE) parameter within a subpopulation defined by specific values of some covariates. While this parameter has been studied before (e.g. Heckman et al. (1997); Hahn (1998)), prior to Heckman and Vytlacil (2005) it served only as an intermediate estimand used in the estimation of the ATE. Lee and Whang (2009) and Hsu (2017) consider estimation and hypothesis testing of the CATE parameter when the conditioning covariate(s) are absolutely continuous. MaCurdy et al. (2011) also discusses the identification and estimation of the CATE parameter when conditioning on the entire set of covariates.

When seeking policy-relevant conclusions, conditioning on the entire set of available covariates is not ideal, and considering only absolutely continuous covariates is restrictive. I thus implement the CATE estimator proposed by Abrevaya et al. (2015), which allows me to include a large set of covariates, but condition on a strict subset of those covariates. In particular, I will estimate the CATE conditional on student gender and pre-treatment test scores, the latter serving as proxies for ability. It is plausible that teacher gender effects may vary with student ability, and understanding what that heterogeneity looks like can inform the process by which teachers are assigned to students.

In contrast with the quantile treatment effect (QTE), an older and more established parameter that captures heterogeneity, the CATE is advantageous in generating policy-relevant conclusions. The primary drawback of the QTE approach is that it allows for heterogeneity in the treatment effect across subpopulations that are not identifiable based on covariates. For instance, the QTE of assignment to a female teacher might be positive at the 60th quantile, but it may not be possible to determine *a priori* if a student would be in the 60th quantile or not. The CATE is defined explicitly in terms of *a priori* observable covariates - thus, if one were to know the true CATE function, it would be perfectly clear before treatment what effect a particular student would see.

4.2 The Conditional Average Treatment Effect

Suppose that one has a sample from a population of interest, consisting of $\{Y_i, D_i, X_i\}_{i=1}^{n}$. Y_i is the outcome of interest, D_i is a binary treatment indicator, and X_i is a vector of observed covariates. If treatment assignment is unconfounded conditional on the covariates X, the ATE is nonparametrically identified and can be recovered via a multitude of estimation procedures.

However, I am interested in how the ATE varies with a subset of the covariates. Formally, the CATE is defined as:

$$\tau(x_1) = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid X_1 = x_1\right] \tag{1}$$

where $Y_i(1)$ and $Y_i(0)$ are the *potential* outcomes[11] for observation i. In addition, X_1 is a strict subset of X - for example, X_1 is gender. If, after splitting the sample, unconfoundedness holds, $\tau(x_1)$ can be recovered simply by estimating the ATE in the two subsamples. However, if the conditioning covariate includes continuous variables (or variables that are discrete, but highly granular), sample splitting ceases to be a practicable option.

Abrevaya et al. (2015) provides an estimator that handles the cases where sample splitting alone is insufficient. Conditioning on X_1 alone will generally violate unconfoundedness, and therefore recovery of $\tau(x_1)$ requires estimating the CATE conditional on X and averaging out the unwanted components. Their estimator is a two-step estimator based on inverse

[11] $Y_i(1)$ is 'the outcome we would have observed if i received treatment', and $Y_i(0)$ is 'the outcome we would have observed if i did not receive treatment.'

probability weighting:

$$\hat{\tau}(x_1) = \frac{\frac{1}{nh^l}\sum_{i=1}^{n}\left(\frac{D_iY_i}{\hat{p}(X_i)} - \frac{(1-D_i)Y_i}{1-\hat{p}(X_i)}\right)K_1\left(\frac{X_{1i}-x_1}{h_1}\right)}{\frac{1}{nh^l}\sum_{i=1}^{n}K_1\left(\frac{X_{1i}-x_1}{h_1}\right)} \tag{2}$$

where $K_1(\cdot)$ and h_1 are, respectively, a kernel function and a bandwidth. l is the dimension of the vector X_1, and $\hat{p}(X_i)$ is an estimate of the propensity score[12]. Subject to mild regularity conditions on the first-stage propensity score estimation, Abrevaya et al. show that this estimator is asymptotically consistent for the CATE under the familiar unconfoundedness and sampling assumptions necessary for ATE estimation.

4.3 Estimating the effect of teacher gender conditional on student gender and test scores

For this investigation, I am interested in the heterogeneity in the effect of teacher gender across both student gender and pre-treatment test scores. Since student gender is a binary variable, sample splitting suffices for conditioning on gender. However, while pre-treatment test score is nominally discrete, sample splitting is not a realistic approach. Sample splitting on pre-treatment test scores would result in approximately 80 subsamples given my data. Thus, after splitting the sample by gender, I let X_1 be the pre-treatment test score and use the estimator described in (2) to estimate the CATE conditional on pre-test score in both the male and female subsamples.

The key identifying assumption underlying this approach is that a student's potential post-treatment test score outcomes are independent of the gender of the student's assigned teacher, conditional on some set of covariates X. Intuitively, this means that when comparing propensity-score weighted treated students to weighted control students, I am comparing

[12] Abrevaya et al. (2015) considers both parametric and nonparametric estimation of the propensity score, and provides consistency results for both cases. While the nonparametric approach offers potential efficiency gains, it does not handle discrete covariates well, and quickly runs into the curse of dimensionality when the set of covariates is of high dimension. As a result, I estimate the propensity score parametrically.

like with like. If, for instance, students with a high pre-treatment test score were more likely to be taught by a female teacher than a male teacher (and pre-treatment test scores predict post-treatment test scores), then non-parametric estimates that did not control for pre-treatment test scores would overstate the true effect of assignment to a female teacher.

Since students were assigned to teachers randomly, conditional on school and grade, a significant portion of the potential confounding mechanisms are rendered impossible - for instance, assignment of female teachers to students based on their pre-treatment test scores. However, some potential confounders remain. In particular, while students were randomly assigned to teachers, teachers were not randomly assigned to preparation pathways (i.e. TFA teachers and non-TFA teachers are likely to be different), nor were they randomly assigned to schools or grades (i.e. teachers in one school or grade may be different to teachers in another school or grade).

For TFA teachers, dealing with the latter issue is straightforward. Through correspondence with Teach For America, I have confirmed that TFA applicants at the time of the experiment were asked for regional preferences, as well as preferences for the level of education (primary school, middle school, or high school). Since the NETFA experiment was conducted only in primary schools, it is sufficient to include a region indicator in the propensity score estimation to control for non-random assignment of TFA teachers to schools or grades.

For non-TFA teachers, non-random assignment of teachers to schools or grades poses more of a problem. It is certainly possible that non-TFA teachers could select into different schools *within* a region, which would not be adequately controlled by a region indicator. It is even possible that teachers select into particular grades. However, it is hard to see why teachers would select differentially into schools within the population from which the sample was drawn. While teachers almost certainly select into or out of high-poverty schools, it is less clear that they select into different schools within the population of high-poverty schools, outside of simple geographic reasons (which are adequately controlled for by region indicators).

This would seem to suggest that the propensity score should be estimated as a function of region indicators (and perhaps school/grade indicators). However, this goes too far towards treating the data as coming from a perfectly randomized experiment. Notably, some schools in the sample have no male teachers - using school indicators when estimating the propensity score would result in students from those schools having estimated propensity scores of either 1, which is far from credible. Even if there is differential selection of teachers into schools, it is very difficult to see how it could produce certain schools that would *never* have male teachers. The existence of schools with only female teachers is far more likely to be a result of the relative proportion of female primary school teachers in general, rather than evidence of a strong selection mechanism that eliminates male teachers entirely from some schools.

Additionally, for the purpose of estimating treatment effects, the goal of the propensity score estimation step is *not* to produce optimal estimates of the propensity score. Rather, the goal is "to obtain estimates of the propensity score that balance the covariates between treated and control samples" (Imbens and Rubin, 2015). In finite samples[13] it is thus important to include not only covariates that potentially explain treatment assignment, but covariates that explain the outcome of interest - even if they are known not to play a role in treatment assignment. I thus estimate the propensity score with the following logistic regression:

$$\log \frac{P(FEMTEACH_i = 1)}{1 - P(FEMTEACH_i = 1)} = \beta_0 + \beta_1 SC_i^{'} + \beta_2 TC_i^{'} + \beta_3 R_i^{'} + \beta_4 TFA_i + \beta_5 CS_i \qquad (3)$$

where $SC^{'}$ is a vector of student covariates (race/ethnicity, pre-treatment test score, and an indicator for the presence of a disruptive student in the classroom), $TC^{'}$ is a vector of teacher characteristics (race/ethnicity and years of teaching experience), $R^{'}$ is a vector of region dummy variables, TFA is an indicator for whether the teacher was a TFA teacher or not, and CS_i is the size of student i's class. In the appendix, I consider alternative specifications for the propensity score.

[13] With a sufficiently large sample, correctly specifying the propensity score model suffices to achieve covariate balance. However, in any finite sample, even one from a perfectly randomized experiment, there is no guarantee that weighting by the true propensity score will balance important covariates.

One potential issue facing any investigation that uses inverse propensity weighting is the effect of very large or very small propensity scores. It is clear from equation (2) that if $\hat{p}(X_i)$ is very close to 0 (1) for treated (untreated) students, the outcomes for those students will be inflated significantly by the weighting procedure. Weights such as these lead to highly variable estimates, and may indicate a failure of the overlap condition. In the above specification, this does not prove to be a significant issue. To deal with the minority of students with extreme propensity scores, I set propensity scores above 0.95 (below 0.05) to 0.95 (0.05). My results are robust to different trimming behavior - in particular, dropping students with extreme propensity scores does not have a noticeable effect on the results.

4.4 Choice of Smooting Parameters

The IPW-based estimator in (2) requires the choice of two smoothing parameters - the kernel and the bandwidth. Following Abrevaya et al. (2015), I set bandwidth to be a multiple of the sample standard deviation in the conditioning covariate (pre-treatment test score). In my main specification, the bandwidth is set to be half the sample standard deviation (approximately 9 for male students in math, for example). I use a Gaussian kernel:

$$K_g(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \tag{4}$$

In the appendix, I report results for different bandwidths and kernels. As is often the case with kernel-based local averaging, bandwidth choice strongly influences the resulting estimates, while kernel choice generally does not have a strong effect. Smaller bandwidths produce more variable CATE estimates, which are often non-monotonic and can have extreme ranges. Larger bandwidths produce flatter CATE estimates, and mechanically force the estimated CATE function towards monotonicity. As bandwidth increases, the CATE estimator quickly becomes uninformative as to heterogeneity, essentially recovering an estimate of the ATE.

While overfitting is a valid concern, my main goal is not to provide another estimate of the average effect of teacher gender. Heterogeneity in that effect is my primary concern, and I thus err on the side of choosing a bandwidth that is too small for my main specification.

5 Results

5.1 Conditioning on Pre-Treatment Test Score

Figure 2 depicts the estimated CATE function for female students. Post-treatment math test scores are the outcome of interest, and the conditioning covariate is the student's pre-treatment normal curve equivalent test score in math. Pointwise valid confidence bands are constructed using the asymptotic approximations from Abrevaya et al. (2015). As one would expect, given the distribution of pre-treatment test scores in the sample (Figure 1), the size of the confidence intervals grows rapidly once the pre-test score exceeds approximately 50, due to lack of data. Notably, the confidence interval for a pre-test score of 1 is relatively

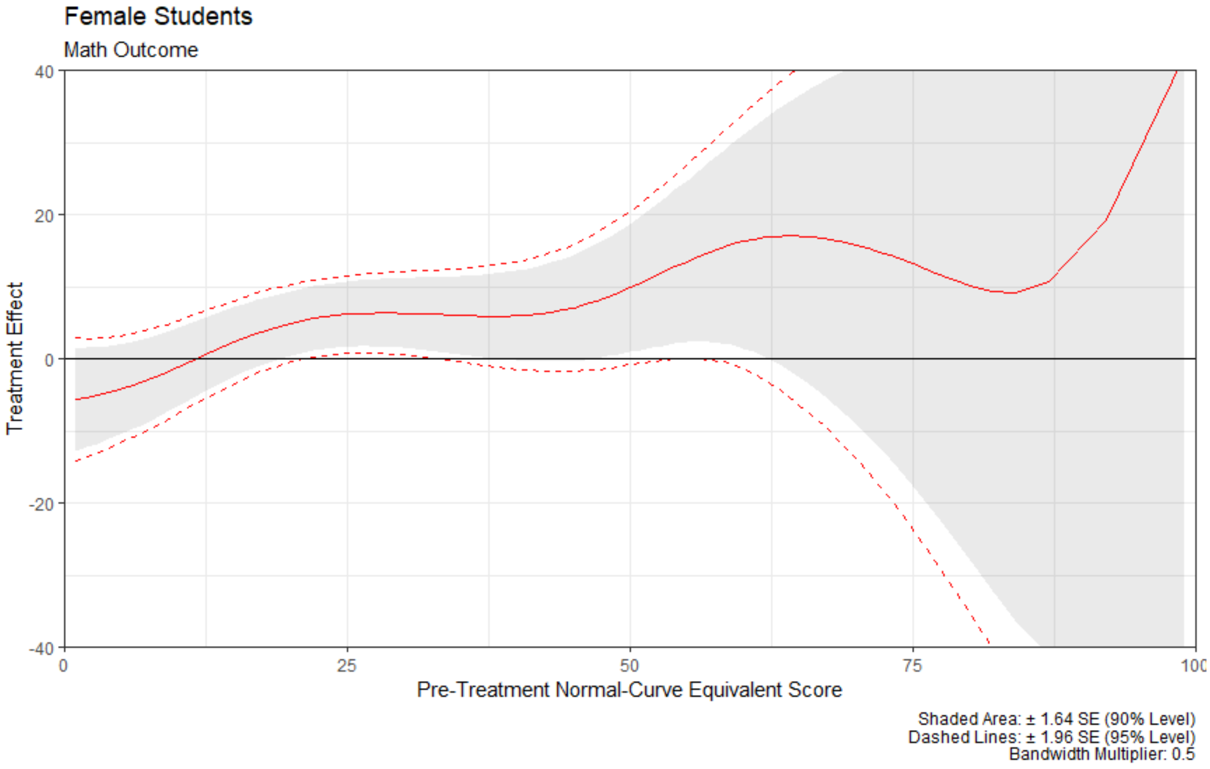


Figure 2: CATE for female students

small, despite being a boundary point. This is largely due to the significant mass of students scoring 1 on the pre-test (also seen in Figure 1).

For the majority of students in this sample, I cannot reject the hypothesis that the true effect of being assigned a female teacher is zero. Indeed, while the confidence intervals here are pointwise valid, it is highly likely that uniformly valid confidence bands would be wider, and thus would not reject the hypothesis that the true effect of assignment to a female teacher is a *constant* zero across the pre-test distribution. The implied average treatment effect[14] is around 0.25 standard deviations, or 4.5 points on the normal curve equivalent scale. While this is quite high, especially in comparison to Antecol et al. (2015), note that formally assessing the statistically significance of the implied ATE remains an open question. In light of the confidence intervals and the size of the implied ATE, it seems unlikely that the implied ATE would be statistically significant[15]. Restraining the calculation to consider only point estimates below 55, the implied ATE decreases to around 0.19 standard deviations (3.4 on the normal curve scale).

Qualitatively, while the majority of the point estimates are insignificant, the confidence intervals themselves suggest that if the true effect is not zero, female students at the very bottom of the ability distribution in math see less benefit from assignment to a female teacher. Outside of the very bottom of the ability distribution, there does not appear to be much, if any, heterogeneity in the effect of teacher gender on math test scores for female students. My results are reasonably consistent with the true CATE having a monotonic relationship between pre-test scores and the treatment effect, which is not immediately unreasonable. Indeed, particularly for TFA teachers, it is entirely plausible that students with higher ability are easier to teach effectively[16].

[14] The implied ATE is calculating by taking a weighted average of the CATE point estimates, where the weight on $\hat{\tau}(x_1)$ is equal the proportion of the sample with $X_1 = x_1$. It is the point estimate of the average treatment effect we would expect to see if the CATE point estimates are correct.

[15] I performed a standard non-parametric bootstrap for the implied ATE, and subject to the caveat that such a procedure is not currently known to be valid, the bootstrap results support this claim.

[16] Since TFA is a *highly* selective program and primarily accepts the highest-achieving applicants, it is likely that those applicants were high-achievement students in primary school as well. Since they receive a relatively small amount of accelerated training in teaching, they may have an easier time understanding

Figure 3 depicts the estimated CATE function for male students, again with math scores as the outcome of interest and conditioning covariates. The increase in the size of the confidence intervals starts even earlier than in Figure 2, primarily because the male pre-test score distribution is more skewed to the left than that of the full sample (which is in line with male students generally performing worse than female students in school). In addition, since no male in the sample scored higher than 92 on the pre-test, CATE estimates for pre-treatment test scores above 92 cannot be constructed. In contrast to Figure 2, for the majority of the students in this sample the effect of assignment to a female teacher is at least marginally significant and positive. This is in stark contrast to what one would expect if the bias from Cappelen et al. (2019) was present. If anything, my results so far would be consistent with a bias in the opposite direction - against low-performing or low-ability *female* students.

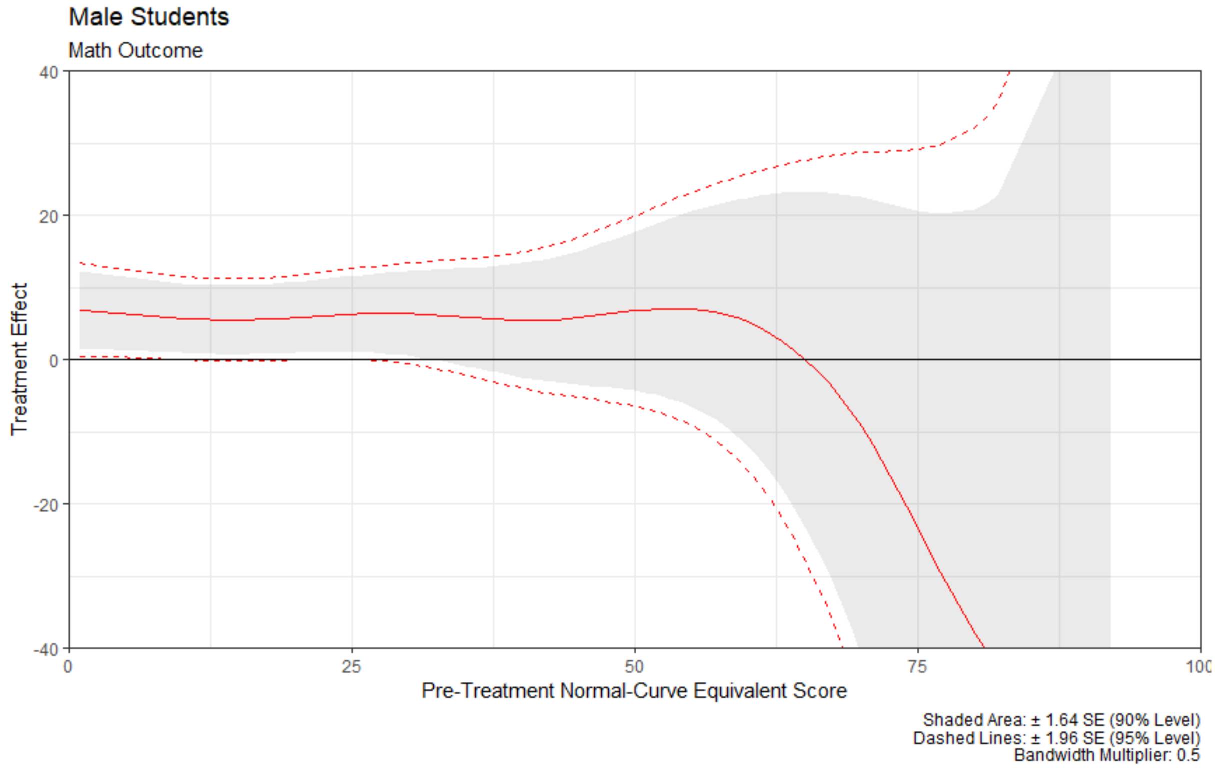


Figure 3: CATE for male students

the difficulties faced by high-achieving students in their classrooms while struggling to understand those difficulties faced by the lowest ability students.

The implied ATE is approximately 0.25 standard deviations (4.7 on the normal curve scale). Considering only pre-test scores below 55, as before, raises the implied ATE significantly to 0.33 standard deviations (6.0 on the normal curve scale). As before, it seems unlikely that the implied ATE would be statistically significant. Using the same rough rule of thumb that uniformly valid confidence bands would be larger, it is also likely that I would be able to reject the hypothesis that the true effect was a constant zero.

It is notable that, discounting the extreme point estimates arising from lack of data at the very top of the pre-treatment test score distribution, there is essentially no evidence of heterogeneity in the effect of teacher gender on male students. A male who scored 1 on the pre-test has nearly the same estimated CATE as one who recorded a score between 2 and 55. The only change is an increase in the size of the confidence intervals, which may be entirely due to the decrease in available data as test scores increase. The size of the positive effect is roughly the same as for female students in the middle of the pre-treatment test score distribution.

Figures 4 and 5 depict the estimated CATE functions for female and male students, respectively, with reading test scores as the outcome of interest and conditioning covariates. The first-stage propensity score model is the same as before except for the change from math to reading test score variables. For female students, there is noticeably more heterogeneity in the estimated CATE function, and it is no longer consistent with a monotonic relationship between treatment effects and pre-treatment test scores. The implied ATE is around 0.09 standard deviations (1.7 on the normal curve scale). A much smaller effect on reading than in math is consistent with previous literature studying the effect of teacher gender. Restricting attention to pre-test scores below 55 has almost no impact on the implied ATE. In contrast to previous literature suggesting that effects on reading are non-existent, I find that female students with pre-treatment test scores in the middle of the distribution have see a significant and large positive treatment effect.

For male students, the story appears largely the same as before. There is limited

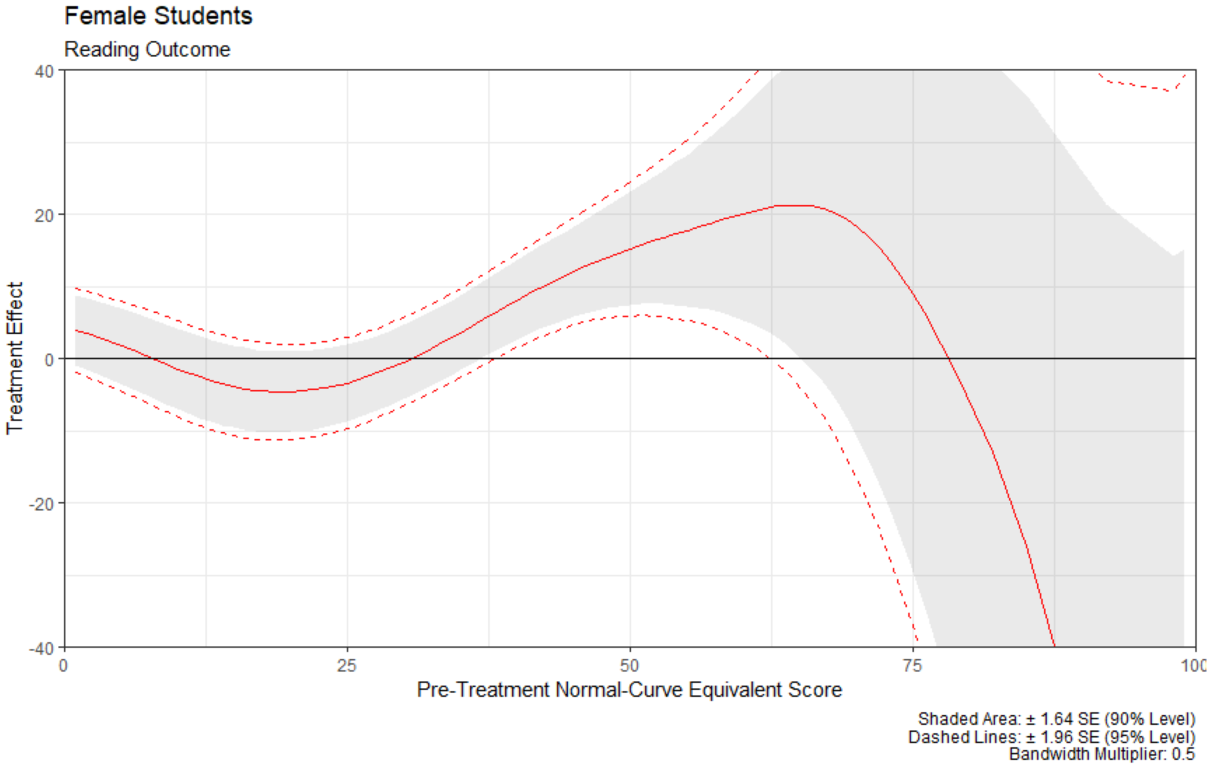


Figure 4: CATE for female students

heterogeneity (although potentially *slightly* more than in math). The estimated CATE is positive for all pre-treatment test scores below 55, as before, and the change in the CATE within that range is limited. As was the case with math results, the implied ATE for male students is relatively large - approximately 0.31 standard deviations (5.5 on the normal curve scale) for the full sample, and around 0.29 standard deviations (5.2) for students scoring less than 55 on the pre-test. Again, it is unlikely that the implied ATE is statistically significant.

5.2 Conditioning on Class Rank

To this point, I have been agnostic as to what might drive heterogeneity in the effect of teacher gender. Most of the standard mechanisms for teacher gender effects could plausibly include heterogeneous behavior. Role model effects, for instance, might be stronger for high-ability students (particularly if the teacher was a high-ability student), or stereotype threat effects on women in math may be more powerful at the low end of the ability distribution. It is also

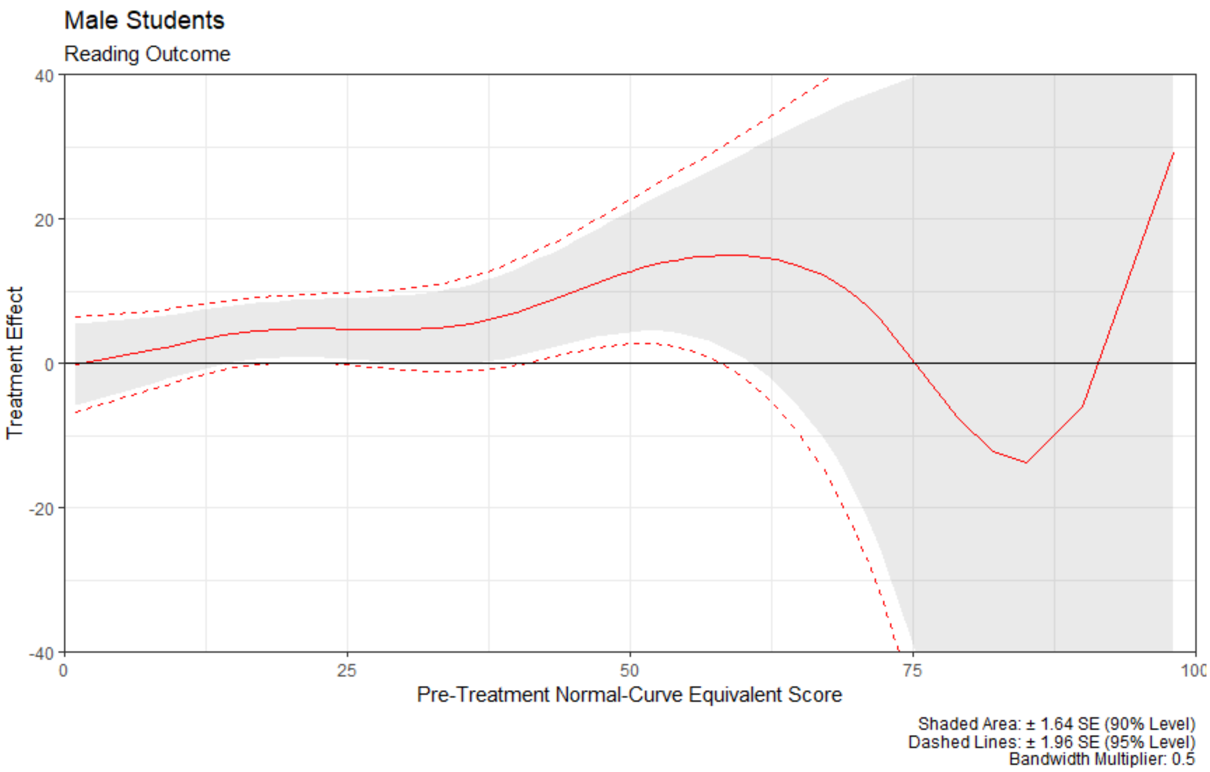


Figure 5: CATE for male students

possible that teacher behavior differs for students of different *perceived* ability - e.g. teachers may invest different amounts of effort in students they perceive as struggling or excelling.

Perceived ability may not closely track 'objective' ability as measured by pre-treatment test scores, or it may be that teachers care more about the ability of a student relative to the rest of the class, rather than relative to a national norm group. To investigate this possibility, I estimate the CATE functions as before, but replace the pre-treatment test score with a class rank variable constructed from the data[17]. Figure 6 presents the estimated CATE functions conditional on class rank for the four subsamples.

The class rank variable is scaled into a 'percentile' rank, with 0 being the worst student in the class and higher values reflecting higher within-class rankings, so the interpretation of the graphs is similar to before - and the results suggest that within-class performance is not

[17] Unfortunately, since some classes contain students not in the research sample, the accuracy of this variable is likely imperfect. If there is a correlation between student ability and whether a student was in the research sample, identification of the CATE may fail for this specification.

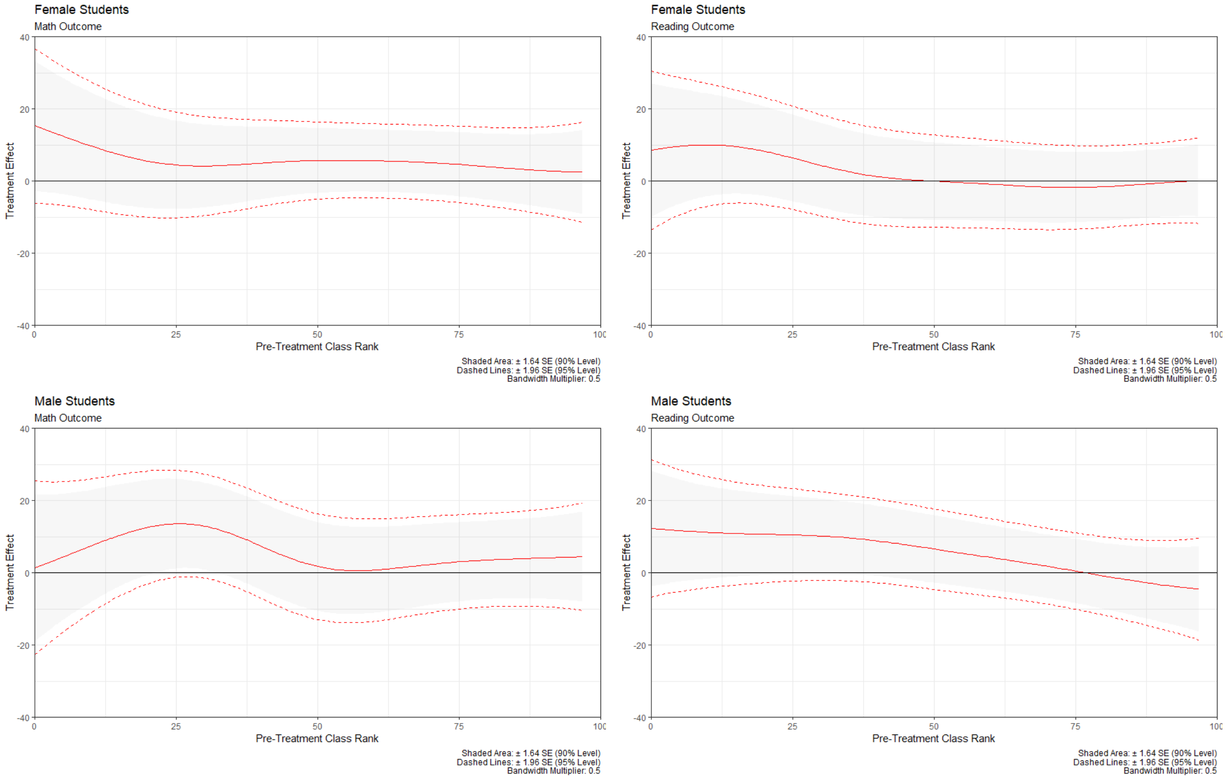


Figure 6: Conditioning on Class Rank

correlated with the size of the teacher gender effect. Even with pointwise valid confidence bands, the hypothesis that the true effect conditional on class rank is a constant zero cannot be rejected in any subsample.

6 Discussion

Somewhat surprisingly, the overriding takeaway from this investigation is that there is very little heterogeneity in the effect of teacher gender on students of different levels of ability. Assignment to a female teacher is either neutral or positive for all students, and the heterogeneity is largely confined to the different effects for male and female students. In math, male students see a uniformly positive effect from assignment to a female teacher, as do female students outside of the very bottom of the pre-treatment test score distribution. In reading, I find that students of either gender with pre-treatment test scores that are average compared to the national norm see positive effects from assignment to a female teacher, and the remainder of students see no significant effect.

The presence of significant effects on reading is surprising in light of the existing literature. It may be that, for relatively well prepared students, female teachers are more effective in teaching reading because they have internalized stereotypes labelling reading as an area where women are better. It may also be the students who have internalized such a stereotype, and exert more effort or are more engaged in reading when taught by a woman.

Differential teacher behavior could also explain why I find a positive effect on male students in math, but no significant effect for female students. Female teachers who view math as a 'male' subject might view low achievement from a male student as a sign that help is needed, while viewing low achievement in math from a female student as being expected. Unlike with the reading effects, it is difficult to see how traditional gender stereotypes about math might drive male students to be more engaged when taught by a woman.

In terms of policy implications, the most important implication is that male students benefit from assignment to female teachers, while female students appear largely unaffected. Primary school teaching is already an occupation dominated by women, and my results suggest that, if anything, this has benefited male students.

Since classes are generally not split by gender, consideration of teacher gender when assigning teachers to classes is unlikely to generate benefits overall. That said, **?** finds that male teachers are more likely to be assigned to classes with lower average math and reading scores. This kind of sorting is likely to have a negative overall effect on student achievement - while the very worst-performing female students might benefit from assignment to a male teacher, my results suggest that male students will be harmed, and female students with higher scores may also be harmed relative to being assigned a female teacher. If anything, my results suggest that, all else equal, women should be preferred when seeking a teacher for a classroom of low-achieving students.

In terms of average effects, my results are rather different from those of Antecol et al.

(2015), who find a negative association between assignment to a female teacher and a female student's test scores in math. Partially, this is due to consideration of different parameters. Antecol et al. (2015) consider estimates of could be thought of as the effect of being a female student, and how that changes with teacher gender. In their specification, the estimated effect of being assigned to a female teacher is insignificant at conventional levels for all students, which is at least somewhat consistent with my results. More generally, the relative treatment effects for male and female students display the same relationship - males benefit more (or are harmed less) by assignment to a female teacher. Antecol et al. also provide suggestive evidence that the mechanism underlying their results is that of stereotype threat, which falls in line with the hypothesis of differential teacher behavior proposed above.

As my sample is not representative of the U.S. student and teacher populations, it is possible that my results are driven by the difference between the population of disadvantaged schools and the broader U.S. school population. It is plausible, for instance, that teachers working in the most disadvantaged schools are less likely to be biased against (or more aware of their potential biases against) low-ability student. They may receive specialized training to help them effectively teach low-ability students that a teacher in a less disadvantaged school would not receive. The level of schooling may also play a role - as my sample consists entirely of primary school students between first and fifth grade. Different levels of schooling, and students/teachers more representative of the U.S. school population overall, provide exciting avenues to extend this research.

7 Conclusion

I estimate the Conditional Average Treatment Effect of assignment to a female teacher on students of different abilities, using data from the National Evaluation of Teach for America, a field experiment run between 2001 and 2003. I find little evidence of heterogeneity across students of different abilities, and a small degree of heterogeneity across students of different genders. Male students see a uniformly positive, but marginally significant, effect from being assigned to a female teacher in math, while female students see effects that are generally insignificant. In reading, students that are average relative to the national norm group see positive and significant effects from assignment to a female teacher, while the remainder of students see insignificant effects.

Overall, my results suggest that teacher gender effects in math do not significantly change with student ability, with what little heterogeneity there is being primarily on the gender axis. In reading, there is some evidence of heterogeneity along the ability axis, but much less difference between students of different genders. My results are most consistent with teachers internalizing traditional gender stereotypes regarding math and reading, and not consistent at all with the bias identified by Cappelen et al. (2019).

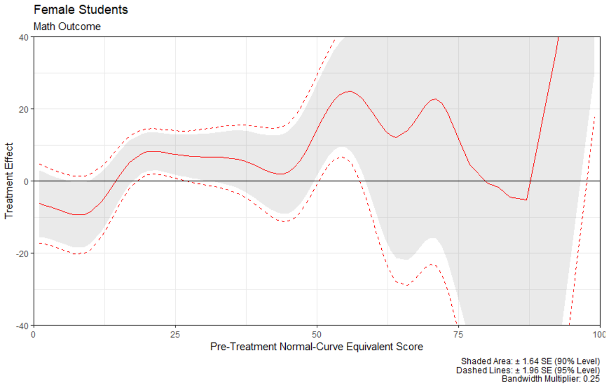
8 Appendix

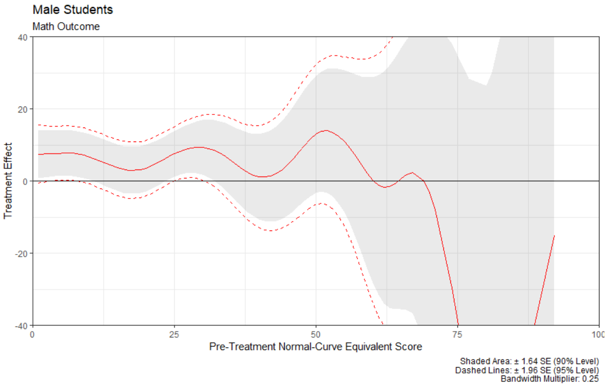
8.1 Robustness Checks

8.1.1 Bandwidth Choice

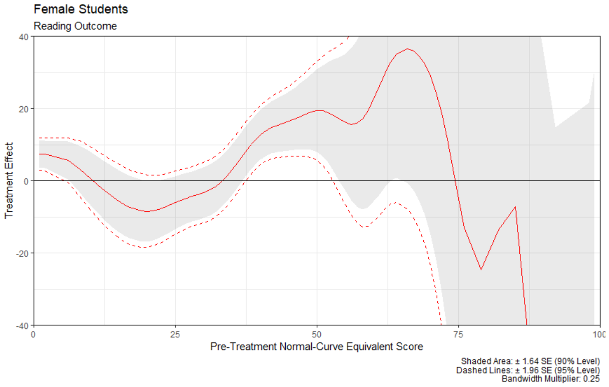
Following Abrevaya et al. (2015), the bandwidth for my estimates was selected as a multiple of the sample standard deviation in the conditioning covariate. I consider four different multipliers - 0.25, 0.5, 1, and 2. While the range of these multipliers is much smaller than that considered by Abrevaya et al. in their empirical illustration, it will quickly become clear that even the medium bandwidth of 1 causes the CATE estimator to oversmooth to the extent that it becomes no more informative than an ATE estimator.

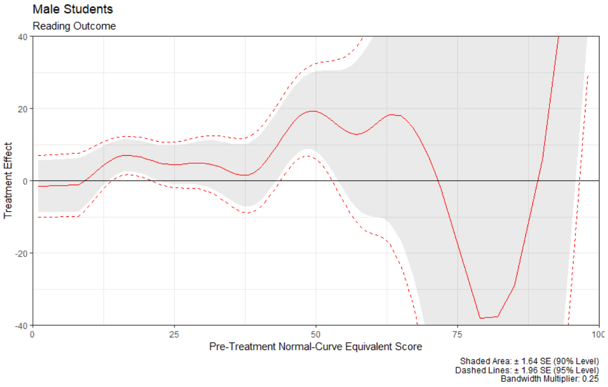
Recall that my main specification sets the bandwidth multiplier to 0.5. Setting the bandwidth multiplier to 0.25 causes the estimated CATE function to be significantly less smooth. Qualitatively, however, the story is largely unchanged. The worst-performing female students see a negative effect of assignment to a female teacher, while male students see significantly less heterogeneity and no significant negative effects.



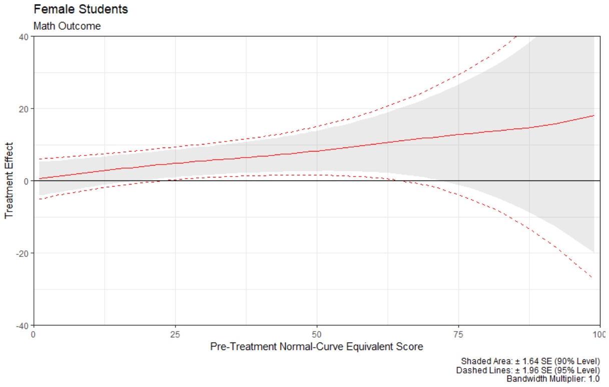


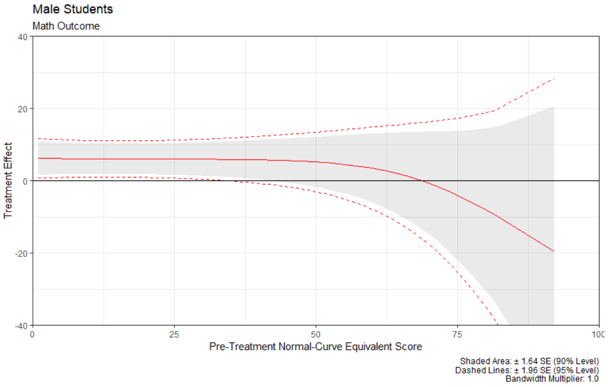
The effect of reducing the bandwidth multiplier is nearly identical for reading outcomes. The qualitative story of the estimated CATE function is largely unchanged - significant effects are observed in roughly the same places, and the general shape of the function is similar. Again, there appears to be significantly less heterogeneity for male students than for female students.

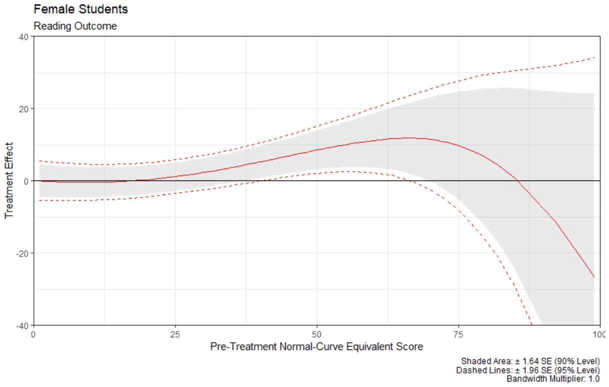


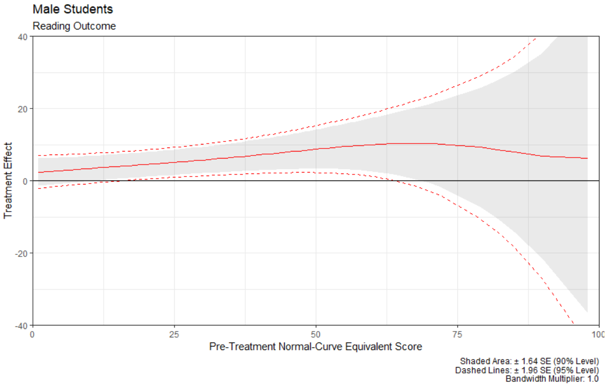


Moving in the other direction and increasing the bandwidth multiplier pushes the estimated CATE function strongly towards monotonicity, and towards a flat slope. With a bandwidth multiplier of 1, almost every estimated CATE function is strictly monotonic, and the vast majority of the variation occurs for estimates conditional on the highest test scores, where very little data is available. Given the heterogeneity present for smaller bandwidths, it seems reasonable to say that at this bandwidth the estimator is clearly oversmoothing. However, note that even with this bandwidth, female students still see notably more heterogeneity than male students in reading, although the difference largely vanishes for math.

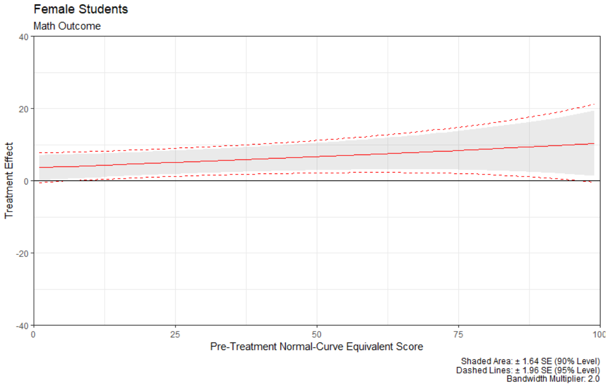


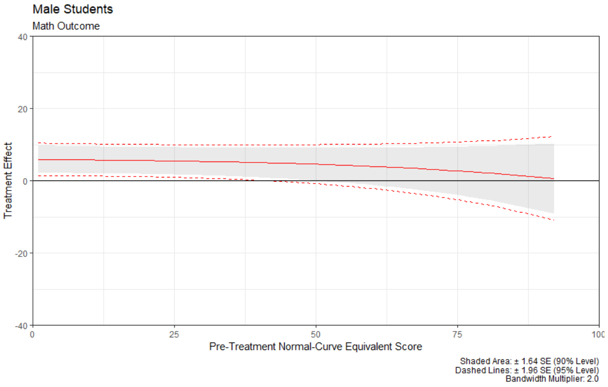


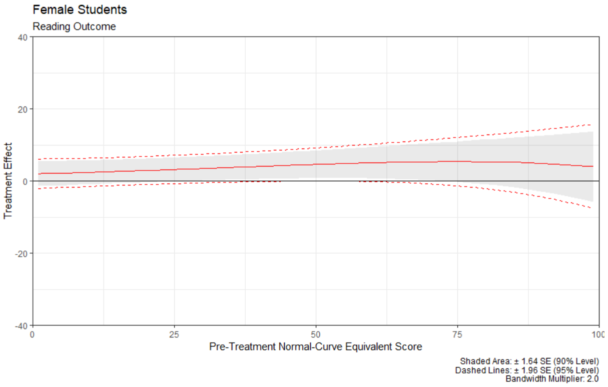


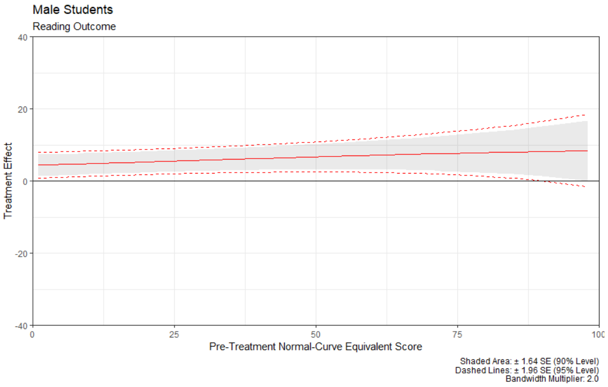


Increasing the bandwidth multiplier even further, to 2, forces near-constancy on almost all estimated CATE functions:









8.1.2 Kernel Choice

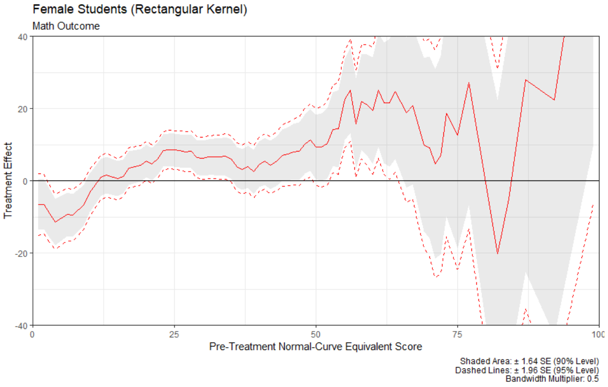
As tends to be the case with kernel-based local averaging estimators, the choice of kernel does not have a huge impact on the resulting estimates - bandwidth choice is dramatically more important. I consider two different kernels - the rectangular (uniform) kernel K_r and the Epanechnikov kernel K_e:

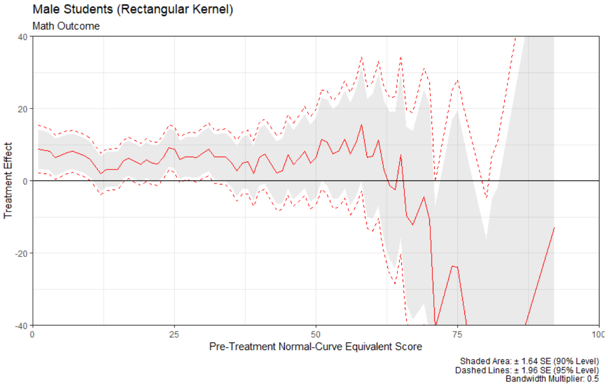
$$K_r(u) = \begin{cases} \frac{1}{2} \text{ if } |u| \leq 1 \\ 0 \text{ otherwise} \end{cases}$$

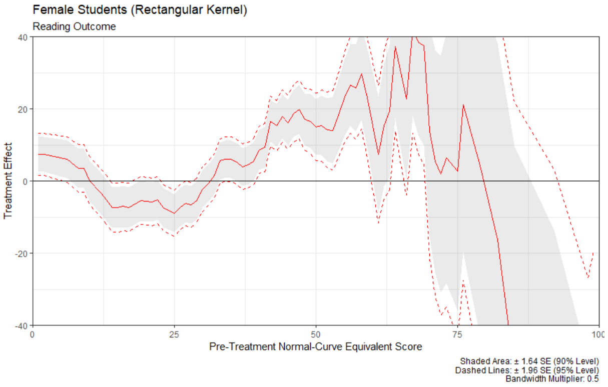
$$K_e(u) = \begin{cases} \frac{3}{4}\left(1 - u^2\right) \text{ if } |u| \leq 1 \\ 0 \text{ otherwise} \end{cases}$$

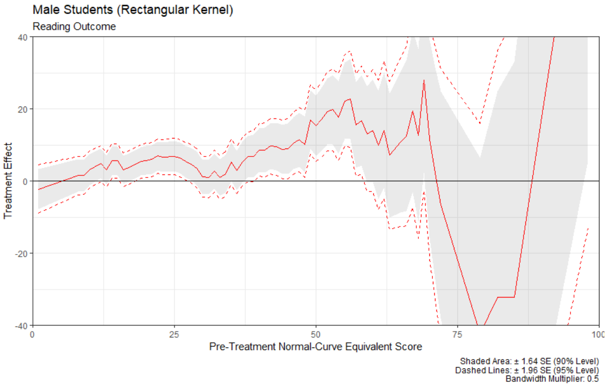
The primary difference between these kernels and the Gaussian kernel is that weights decrease towards zero more rapidly, particularly with the rectangular kernel. This results in less smooth estimates of the CATE function, but the qualitative story is largely unchanged. The effect of bandwidth choice is essentially identical for all kernels, so I report only the results for the intermediate bandwidth multipliers of 0.5 and 1 for these alternative kernels. The effects of other relatively efficient kernels, such as quartic or triweight kernels, are very similar to the effect of the Epanechnikov kernel.

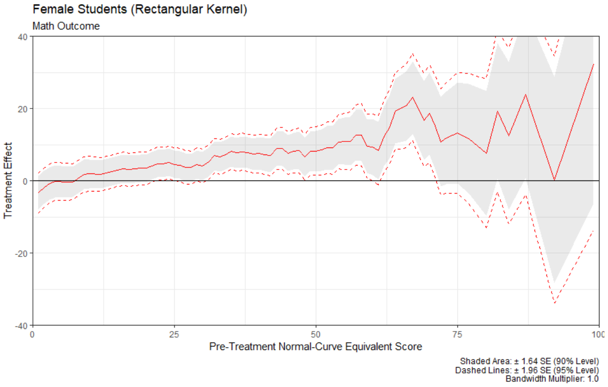
Selection of a rectangular kernel generates the least smooth estimates for any given bandwidth:

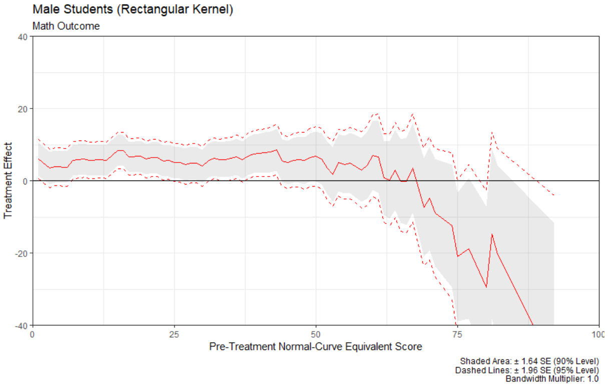


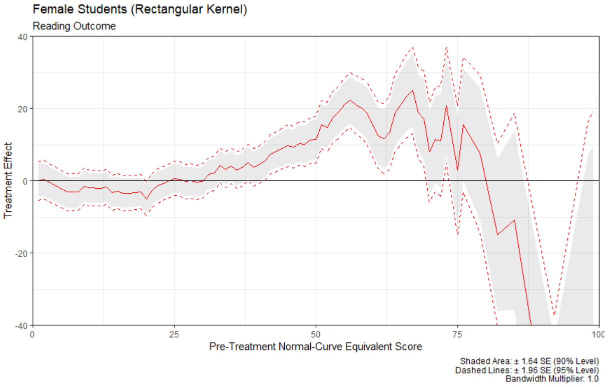


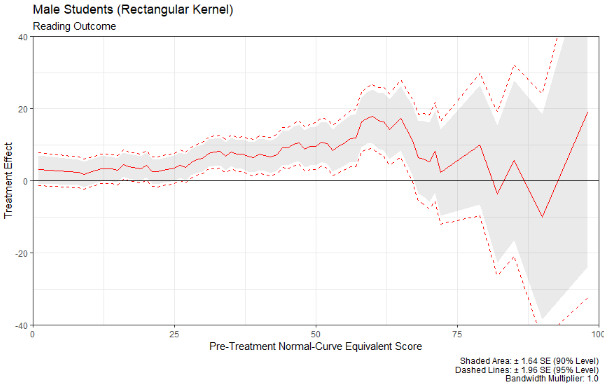




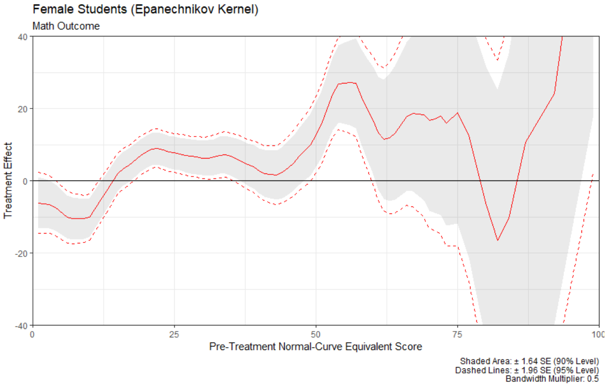


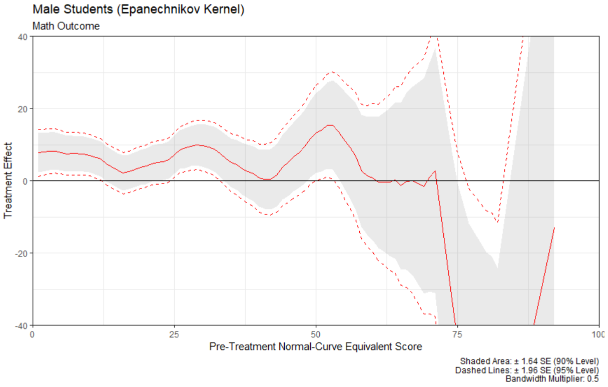


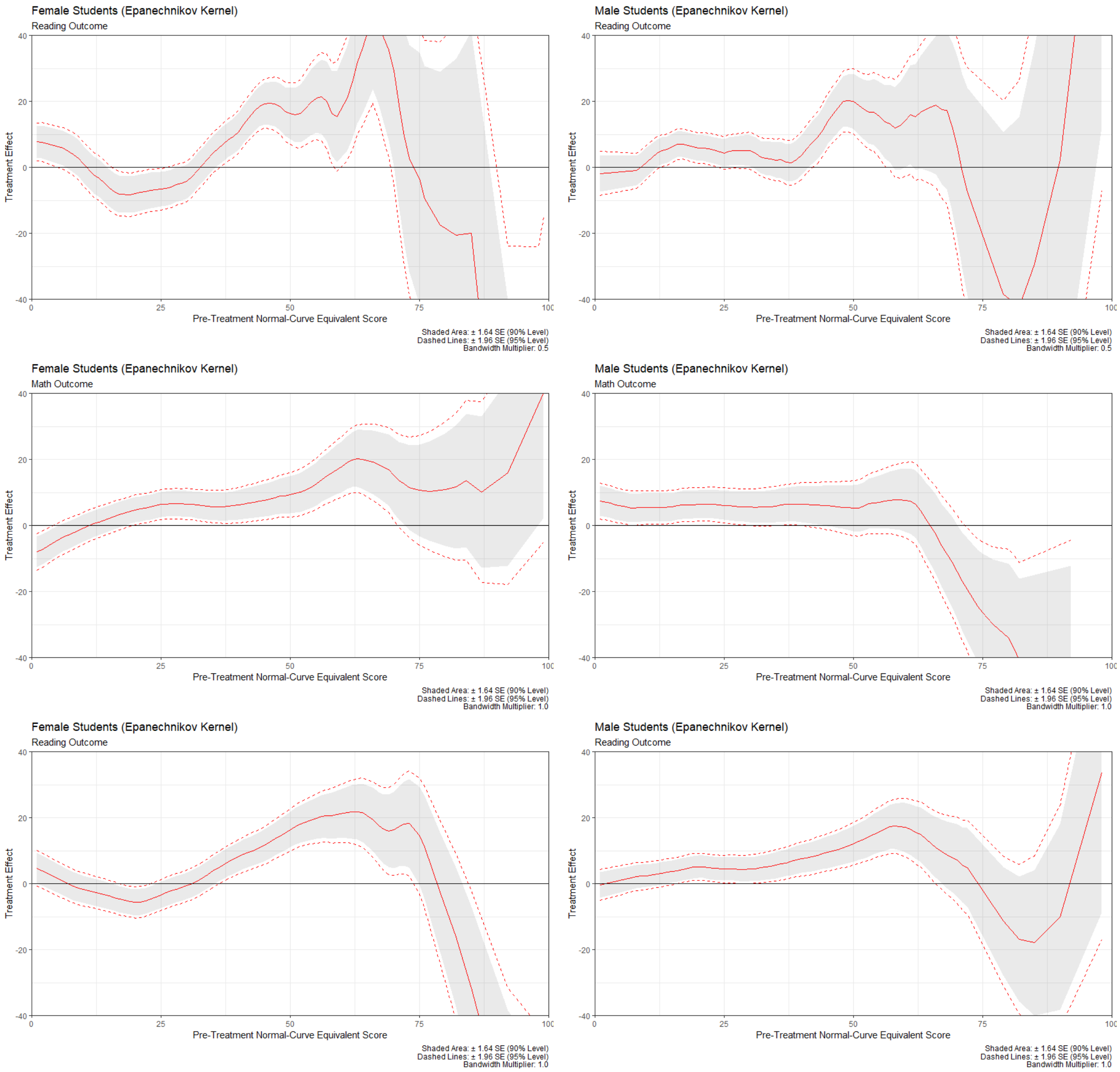




The Epanechnikov kernel likewise does not significantly change the qualitative results:

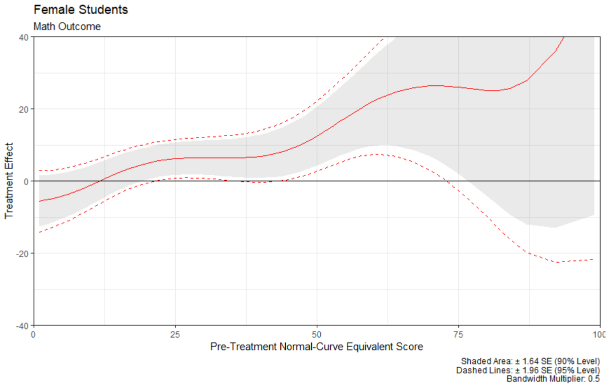


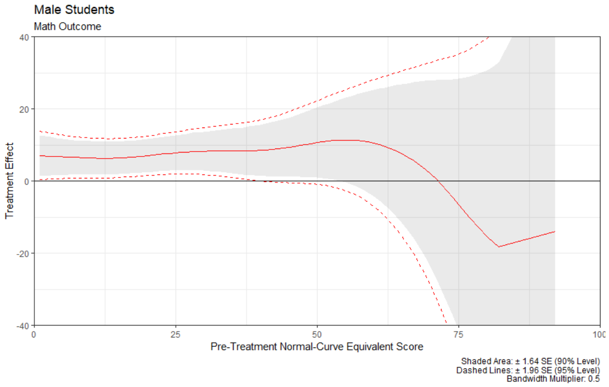


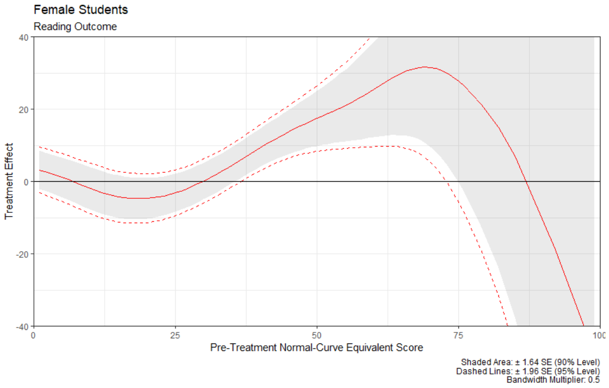


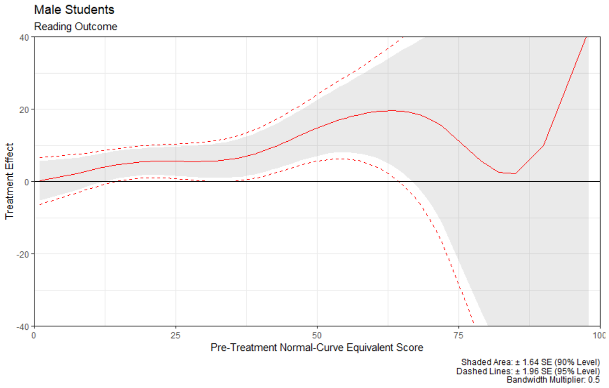
8.1.3 Different Specifications of the Propensity Score

First, I consider the addition of the indicator for a traditional teacher certification. The main specification excludes this variable because previous research (e.g. Staiger and Rockoff (2010)) suggests that teacher certifications are not good predictors of teacher quality, and thus balancing of samples on teacher certification would be harmful unless such balance could be achieved without cost to balance on another covariate (which is not the case). The results of including teacher certification in the propensity score model largely bear this claim out - the qualitative story is almost identical, and the only real change is an increase in the size of the confidence intervals. This is consistent with the expected effects of including an irrelevant covariate in the propensity score model.



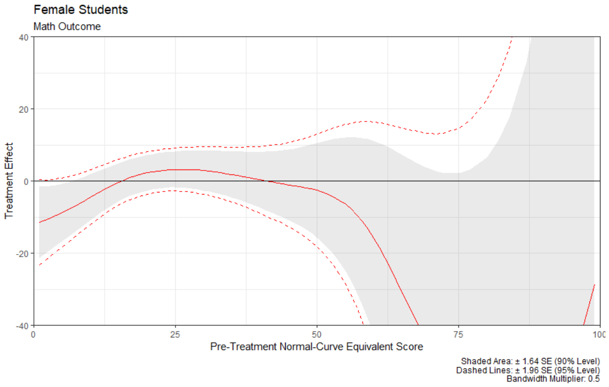


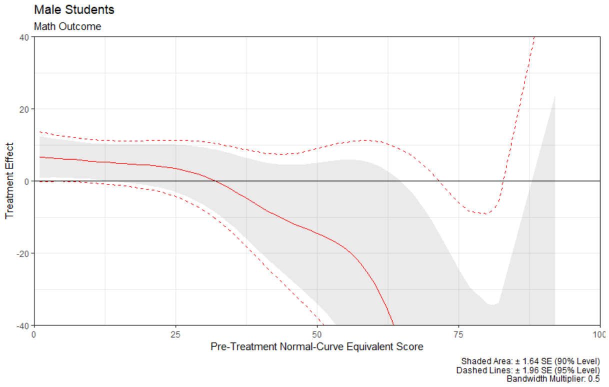


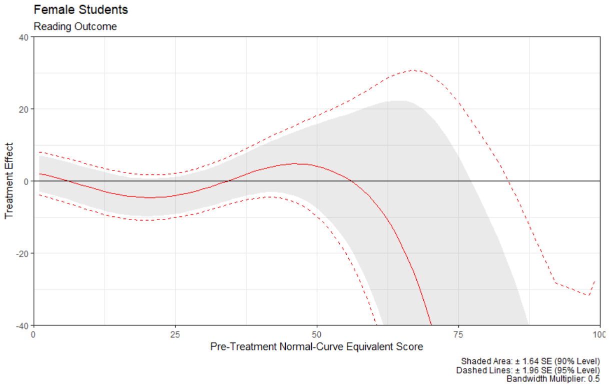


I omit reports for other bandwidth multipliers because the results of that exercise are identical - larger confidence bands, with no significant change to the underlying function.

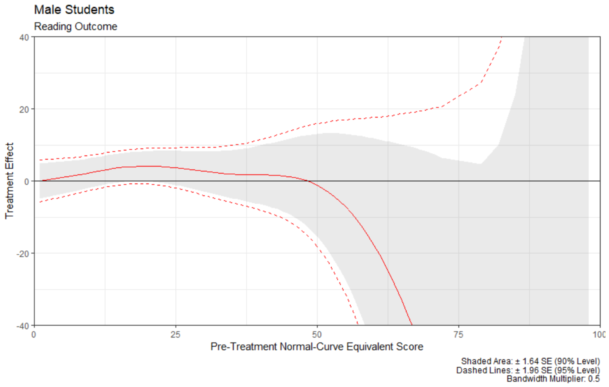
Finally, I consider a much simpler propensity score specification, dropping the teacher and student demographic variables to leave only pre-test score, class size, teacher experience, and indicators for disrupted class, assignment to a TFA teacher, and region. While this specification clearly excludes potentially relevant covariates, it also results in a complete elimination of numerically 0 or 1 propensity scores, and far fewer extreme propensity scores. If the effect of student or teacher demographics is limited, this specification may make a profitable bias/variance tradeoff. In particular, if sorting of teachers into schools was in fact random, or at least uncorrelated with teacher or school characteristics, this specification would be preferable.






Female Students
Reading Outcome

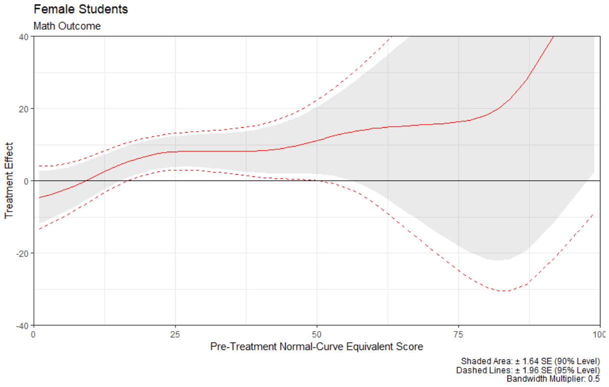
Shaded Area: ± 1.64 SE (90% Level)
Dashed Lines: ± 1.96 SE (95% Level)
Bandwidth Multiplier: 0.5


Male Students
Reading Outcome

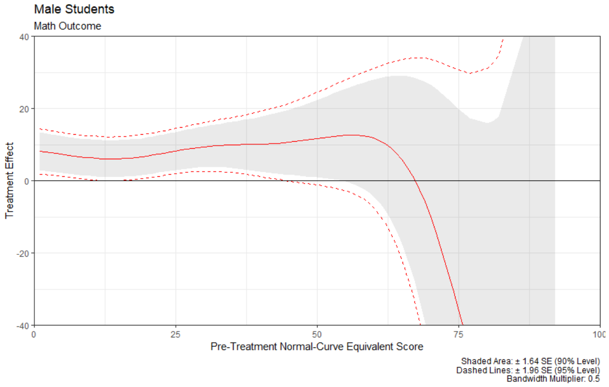
Shaded Area: ± 1.64 SE (90% Level)
Dashed Lines: ± 1.96 SE (95% Level)
Bandwidth Multiplier: 0.5

While the results for male students are very marginally consistent with the results from my main specification, particularly in math, it is clear that (as prior research would suggest) the demographic variables excluded in this specification are relevant. If they were irrelevant or had a sufficiently minor impact on outcomes, one would expect to see smaller confidence intervals but a largely similar underlying function from this specification.

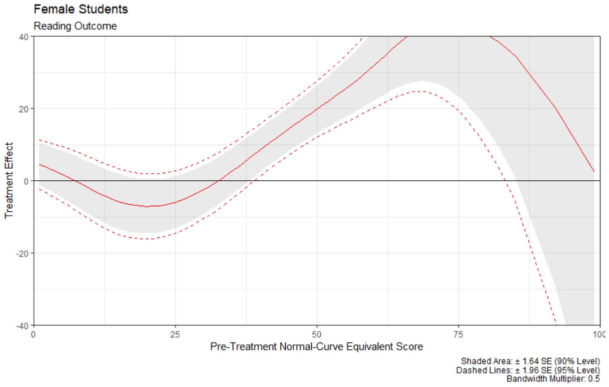
Finally, I consider the addition of a school fixed effect to the propensity score model. Since a significant minority of schools contain only female teachers, this causes the trimming behavior to play a larger part in the results - many more students receive propensity scores close to 1 or 0 and are thus subject to the trimming behavior. With my default trimming behavior (setting extreme propensity scores to 0.95 or 0.05), the results are again reasonably similar in terms of qualitative story.


Female Students
Math Outcome

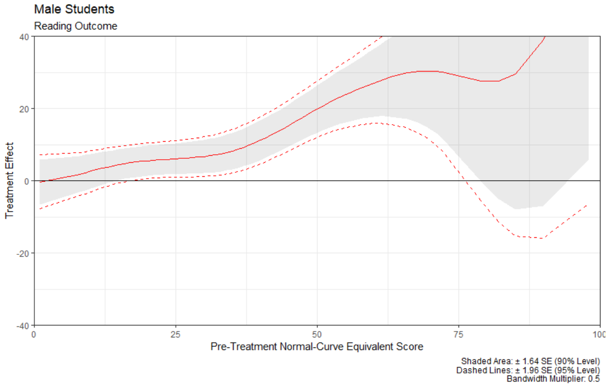
Shaded Area: ± 1.64 SE (90% Level)
Dashed Lines: ± 1.96 SE (95% Level)
Bandwidth Multiplier: 0.5


Male Students
Math Outcome

Shaded Area: ± 1.64 SE (90% Level)
Dashed Lines: ± 1.96 SE (95% Level)
Bandwidth Multiplier: 0.5

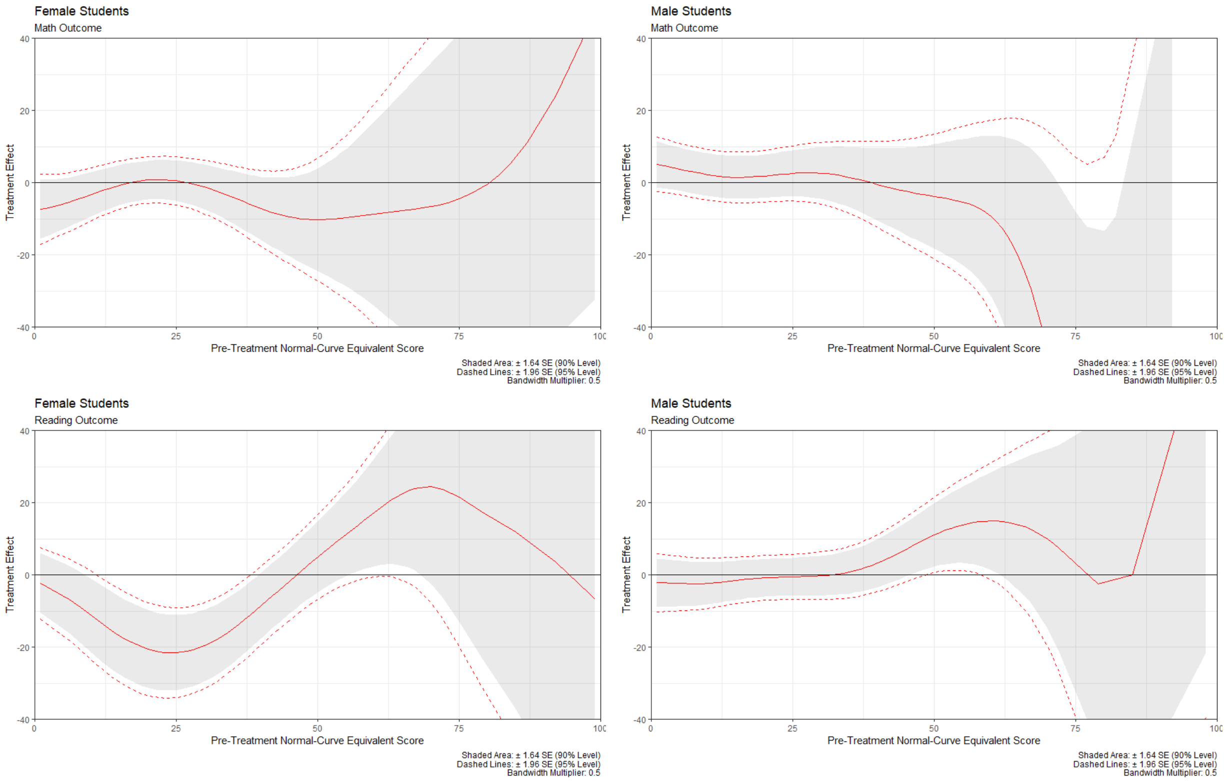

Female Students
Reading Outcome

Shaded Area: ± 1.64 SE (90% Level)
Dashed Lines: ± 1.96 SE (95% Level)
Bandwidth Multiplier: 0.5


Male Students
Reading Outcome

Shaded Area: ± 1.64 SE (90% Level)
Dashed Lines: ± 1.96 SE (95% Level)
Bandwidth Multiplier: 0.5

However, for female students in math and male students in reading, these results are no longer robust to changes in the trimming behavior. Dropping students with extreme propensity scores generates the following results



These results suggest that non-TFA teachers are not sorting differentially into schools within a region, which was the only potential source of endogeneity in my main specification. A conservative reading of these robustness checks would suggest that the positive treatment effect I find on male students is potentially uncertain, but conclusions related to the heterogeneity in the effect of teacher gender on students of differing abilities are unaffected.

References

Abrevaya, J., Hsu, Y.-C., and Lieli, R. P. (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505.

Ambady, N., Shih, M., Kim, A., and Pittinsky, T. L. (2001). Stereotype susceptibility in children: Effects of identity activation on quantitative performance. *Psychological Science*, 12(5):385–390. PMID: 11554671.

Antecol, H., Eren, O., and Ozbeklik, S. (2015). The Effect of Teacher Gender on Student Achievement in Primary School. *Journal of Labor Economics*, 33(1):63–89.

Autor, D. and Wasserman, M. (2013). Wayward sons: The emerging gender gap in labor markets and education. Technical report, Third Way.

Bassi, M., Mateo Díaz, M., Blumberg, R. L., and Reynoso, A. (2018). Failing to notice? uneven teachers' attention to boys and girls in the classroom. *IZA Journal of Labor Economics*, 7(1):9.

Beilock, S. L., Gunderson, E. A., Ramirez, G., and Levine, S. C. (2010). Female teachers' math anxiety affects girls' math achievement. *Proceedings of the National Academy of Sciences*, 107(5):1860–1863.

Bettinger, E. P. and Long, B. T. (2005). Do faculty serve as role models? the impact of instructor gender on female students. *The American Economic Review*, 95(2):152–157.

Bitler, M. P., Gelbach, J. B., and Hoynes, H. W. (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review*, 96(4):988–1012.

Cappelen, A. W., Falch, R., and Tungodden, B. (2019). The boy crisis: Experimental evidence on the acceptance of males falling behind. Discussion Paper Series in Economics 6/2019, Norwegian School of Economics, Department of Economics.

Carlana, M. (2019). Implicit Stereotypes: Evidence from Teachers' Gender Bias*. *The Quarterly Journal of Economics*.

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

Dee, T. S. (2004). Teachers, race, and student achievement in a randomized experiment. *The Review of Economics and Statistics*, 86(1):195–210.

Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review*, 95(2):158–165.

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. *Journal of Human Resources*, XLII(3):528–554.

Downey, D. B. and Pribesh, S. (2004). When race matters: Teachers' evaluations of students' classroom behavior. *Sociology of Education*, 77(4):267–282.

Egalite, A. J., Kisida, B., and Winters, M. A. (2015). Representation in the classroom: The effect of own-race teachers on student achievement. *Economics of Education Review*, 45:44 – 52.

Ehrenberg, R., Goldhaber, D., and Brewer, D. (1995). Do teachers? race, gender, and ethnicity matter? evidence from the national education longitudinal study of 1988. *Industrial and Labor Relations Review*, 48.

Fairlie, R. W., Hoffmann, F., and Oreopoulos, P. (2014). A community college instructor like me: Race and ethnicity interactions in the classroom. *The American Economic Review*, 104(8):2567–2591.

Fortin, N. M., Oreopoulos, P., and Phipps, S. (2015). Leaving boys behind: Gender disparities in high academic achievement. *Journal of Human Resources*, 50(3):549–579.

Gershenson, S., Holt, S. B., and Papageorge, N. W. (2016). Who believes in me? the effect of student/teacher demographic match on teacher expectations. *Economics of Education Review*, 52:209 – 224.

Gong, J., Lu, Y., and Song, H. (2018). The effect of teacher gender on students? academic and noncognitive outcomes. *Journal of Labor Economics*, 36(3):743–778.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64(4):605–654.

Heckman, J. J. and Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, 73(3):669–738.

Hedges, L. V. and Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science*, 269(5220):41–45.

Hess, F. and L. Leal, D. (1997). Minority teachers, minority students, and college matriculation: a new look at the role-modeling hypothesis. *Policy Studies Journal - POLICY STUD J*, 25:235–248.

Hill, C., Corbett, C., and Rose, A. S. (2010). Why so few? women in science, technology, engineering, and mathematics. Report, American Association of University Women.

Hilmer, C. and Hilmer, M. (2007). Women Helping Women, Men Helping Women? Same-Gender Mentoring, Initial Job Placements, and Early Career Publishing Success for Economics PhDs. *American Economic Review*, 97(2):422–426.

Hoffmann, F. and Oreopoulos, P. (2009). A professor like me: The influence of instructor gender on college achievement. *The Journal of Human Resources*, 44(2):479–494.

Holt, S. B. and Gershenson, S. (2017). The impact of demographic representation on absences and suspensions. *Policy Studies Journal*, 0(0).

Hsu, Y.-C. (2017). Consistent tests for conditional treatment effects. *The Econometrics Journal*, 20(1):1–22.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press.

Joensen, J. S. and Nielsen, H. S. (2016). Mathematics and gender: Heterogeneity in causes and consequences. *The Economic Journal*, 126(593):1129–1163.

Krueger, A. (2017). Where have all the workers gone? an inquiry into the decline of the u.s. labor force participation rate. *Brookings Papers on Economic Activity*, 48(2 (Fall)):1–87.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? evidence from a natural experiment. *Journal of Public Economics*, 92(10):2083 – 2105.

Lavy, V. and Sand, E. (2018). On the origins of gender gaps in human capital: Short- and long-term consequences of teachers' biases. *Journal of Public Economics*, 167:263 – 279.

Lee, S. and Whang, Y.-J. (2009). Nonparametric Tests of Conditional Treatment Effects. Cowles Foundation Discussion Papers 1740, Cowles Foundation for Research in Economics, Yale University.

MaCurdy, T., Chen, X., and Hong, H. (2011). Flexible estimation of treatment effect parameters. *The American Economic Review*, 101(3):544–551.

Maria Villegas, A., Strom, K., and Lucas, T. (2012). Closing the racial/ethnic gap between students of color and their teachers: An elusive goal. *Equity & Excellence in Education*, 45:283–301.

Mechtenberg, L. (2009). Cheap Talk in the Classroom: How Biased Grading at School Explains Gender Differences in Achievements, Career Choices and Wages. *Review of Economic Studies*, 76(4):1431–1459.

Neumark, D. and Gardecki, R. (1998). Women helping women? role model and mentoring effects on female ph.d. students in economics. *The Journal of Human Resources*, 33(1):220–246.

Ouazad, A. (2014). Assessed by a teacher like me: Race and teacher assessments. *Education Finance and Policy*, 9(3):334–372.

Ouazad, A. and Page, L. (2012). Students' Perceptions of Teacher Biases: Experimental Economics in Schools. CEE Discussion Papers 0133, Centre for the Economics of Education, LSE.

Penner, E. K. (2016). Teaching for all? teach for america's effects across the distribution of student achievement. *Journal of Research on Educational Effectiveness*, 9(3):259–282.

Staiger, D. O. and Rockoff, J. E. (2010). Searching for effective teachers with imperfect information. *Journal of Economic Perspectives*, 24(3):97–118.

Steele, C. M. (1997). A threat in the air. how stereotypes shape intellectual identity and performance. *The American psychologist*, 52 6:613–29.

Steele, J. (2003). Children's gender stereotypes about math: The role of stereotype stratification1. *Journal of Applied Social Psychology*, 33(12):2587–2606.

Terrier, C. (2016). Boys Lag Behind: How Teachers' Gender Biases Affect Student Achievement. IZA Discussion Papers 10343, Institute of Labor Economics (IZA).

Vesterlund, L. and Niederle, M. (2010). Explaining the gender gap in math test scores: The role of competition. *Journal of Economic Perspectives*, 24:129–44.

Vincent-Lancrin, S. (2008). The reversal of gender inequalities in higher education: An on-going trend. 1.

Winters, M. A., Haight, R. C., Swaim, T. T., and Pickering, K. A. (2013). The effect of same-gender teacher assignment on student achievement in the elementary and secondary grades: Evidence from panel data. *Economics of Education Review*, 34:69 – 75.