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

Niklaus Julius

January 22, 2020

Abstract

Gender dynamics may play an important role in the determination of student outcomes in education. To date, the study of student/teacher gender dynamics has focused mostly on average effects. Exploiting random assignment of students to teachers in a field experiment, I study heterogeneity in the impact of teacher gender on the math and reading test scores for primary school students of differing ability. I find that assignment to a female teacher is generally positive for male students while having no significant effect for female students. In addition, I find very little heterogeneity in the effect of teacher gender on the ability axis, suggesting that average effect estimates do not mask significant heterogeneity. My results are consistent with differential teacher behavior based on gender stereotypes, and somewhat inconsistent with differential student behavior based on gender stereotypes.

Keywords: teacher gender, student achievement, heterogeneity

JEL Codes: I21, I24, I26

1 Introduction

Achievement on school tests has important implications for students in both the short and the long run. In the short run, test scores serve as signals to students about their ability, and can induce students to choose different educational paths (Mechtenberg, 2009; Lavy, 2008; Lavy and Sand, 2018; Terrier, 2016). In the long run, these choices can have major implications for lifetime earnings and health outcomes (Joensen and Nielsen, 2016; Autor and Wasserman, 2013; Krueger, 2017). Gender dynamics between students and teachers can play a significant role in determining student test score outcomes (Dee, 2005; Lavy, 2008; Antecol et al., 2015; Terrier, 2016).

To date, the study of gender dynamics in the classroom has mostly considered average effects, which can mask significant heterogeneity (Bitler et al., 2006). It is possible that the effect of teacher gender on students might depend significantly on student ability, which would have important implications for policy - particularly with regard to addressing inequality. For instance, male and female teachers may internalize different gender stereotypes and thus react differently to low- or high-performing male or female students (Williams and Ceci, 2015), or students may internalize different gender stereotypes and thus be more or less receptive to teaching from teachers of a particular gender (Ouazad and Page, 2012).

In this paper, I address this gap in the literature by studying how the effect of assignment to a female teacher changes with both the gender and ability of the student using data from a field experiment conducted to evaluate the Teach for America (TFA) Program. I estimate the Conditional Average Treatment Effect (CATE) of assignment to a female teacher, conditioning on student gender and on pre-treatment test score as a proxy for ability. The CATE parameter is ideal for this study because it is a policy-relevant parameter that directly addresses the question of how student ability changes the effect of teacher gender on student outcomes. My estimates show how the effect of being assigned to a female teacher changes with both student gender and student ability.

Exploiting random assignment of students to teachers in the data allows me to deploy

non-parametric techniques that require the strong assumption of unconfoundedness rather than imposing functional form restrictions. While the data is not representative of the U.S. primary school student population overall, it is representative of the most disadvantaged students and schools - a subset of particular importance to policymakers. Students in these schools are less likely to continue on to higher education, and thus more likely to face the challenges facing individuals without a college education in modern society[1].

I find very limited heterogeneity in the effect of teacher gender on students with different levels of prior achievement. For male students, assignment to a female teacher has a nearly uniform positive impact on math test scores. For reading, there is a small positive relationship between student ability and the effect of assignment to a female teacher. For female students, there is notably more heterogeneity in the effect of teacher gender. In math, there is a stronger positive relationship between ability and the effect of teacher gender than for male students, and some indication that the lowest-performing female students might be harmed by assignment to a female teacher. In reading, there is a non-monotonic relationship between student ability and the effect of teacher gender.

My results echo much of the previous economics literature in finding no significant *average* effect of teacher gender on students. Outside of the bottom of the pre-treatment test score distribution, the effect of assignment to a female teacher does not significantly differ with student gender. At the very bottom of that distribution, female students may benefit less than male students from assignment to female teachers in math. Notably, for all students, the effect of assignment to a female teacher is either positive or insignificant, which suggests that biases such as those found by Lavy (2008), Terrier (2016), or Cappelen et al. (2019) are not present in primary school.

The remainder of the paper is organized as follows. Section 2 reviews related literature.

[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.

Section 3 discusses the data, the institutional background, and the experiment itself. Section 4 briefly introduces the theoretical framework for the CATE estimator and sets out my estimation strategy. Section 5 presents the main results. Section 6 considers possible mechanisms and policy implications. Finally, Section 7 concludes.

2 Related Literature

This paper contributes directly to the literature that studies student/teacher dynamics based on demographic features, and indirectly to a related strand of literature that considers the underlying mechanisms.

Reduced form estimates of the effect of demographic matching between students and teachers go back to Ehrenberg et al. (1995), who found that demographic matching had little impact on student learning, but a significant impact on teacher perceptions of students using NELS:88[2] data. Dee (2004) used Project STAR data to investigate the effect of teacher race on students, finding a positive effect of same-race teachers on math and reading for students. Dee (2005) exploited a unique feature of the NELS:88 data to control for student fixed effects, again finding that student/teacher demographic dynamics had significant effects on teacher perceptions. Dee (2007), restricting attention to gender dynamics, found that assignment to a same-gender teacher significantly improved student test scores, teacher perceptions of the student, and student engagement.

Bettinger and Long (2005) and Hoffmann and Oreopoulos (2009) studied the effect of instructor gender on undergraduate students using administrative data from different universities[3]. Hoffmann and Oreopoulos (2009) found that assignment to a same-sex instructor boosted relative student performance and likelihood of course completion, but had little impact on upper-year course selection. Bettinger and Long (2005) found very mixed results -

[2]The National Educational Longitudinal Study of 1988 consists of a representative sample of students that were in 8th grade in 1988.

[3]Bettinger and Long (2005) uses data on full-time undergraduate students in Ohio during 1998 and 1999. Hoffmann and Oreopoulos (2009) uses data on students at the University of Toronto.

their primary conclusion is that the effect of instructor gender changes dramatically based on the subject in question. For instance, they found strong positive effects on female students in math and statistics, and a weak effect in economics. They also add to the growing number of studies that find negligible effects of instructor gender on male students. Carrell et al. (2010), using administrative data from the U.S. Air Force Academy and exploiting random assignment of students to teachers, found limited impacts of instructor gender on male students, but significant positive impacts on female students in math and science. In contrast to Hoffmann and Oreopoulos (2009), Carrell et al. (2010) finds significant impacts for upper-year course selection. Fairlie et al. (2014), using administrative data from a community college, found similar effects for instructor race - in particular, assignment to an instructor from an underrepresented minority group shrinks the performance gap between white and minority students.

In postgraduate education, Neumark and Gardecki (1998) found that job placement outcomes for female graduate students in economics were not significantly impacted by the addition of female faculty members or having a female dissertation chair, while finding limited evidence for positive effects on graduation time and graduation likelihood. Hilmer and Hilmer (2007) studied top-30 economics doctoral programs between 1990 and 1994, and found that female students with male advisors were significantly more likely to accept a research-oriented first job, but found little effect on early career publication success.

In recent years, some large additional datasets have become available to researchers. Egalite et al. (2015) uses administrative from the Florida public school system to find small but significant effects of teacher race/ethnicity on students. Winters et al. (2013), also using Florida public school data, find that assignment to a female teacher positively impacts both male and female students in math, primarily between the 6th and 10th grade levels.

One implication of this is that teacher gender may not have an effect on student outcomes before middle school. However, there remains some uncertainty about when children begin to understand or internalize gender stereotypes. Ambady et al. (2001) suggests that it begins

around 10 years of age, while Steele (2003) finds evidence suggesting that it begins as early as 7 years of age. Antecol et al. (2015), using the same data as this paper, finds that female teachers have a negative impact on female students in math, and no impact elsewhere. They offer suggestive evidence that the underlying mechanism is math anxiety among female teachers.

The mechanisms underlying student/teacher gender or race dynamics remain an area of ongoing research. One of the most commonly proposed theories is that teachers serve as role models for demographically similar students (Hess and L. Leal, 1997), potentially increasing student motivation and ambition (Maria Villegas et al., 2012), or reducing the effect of stereotype threat[4] (Steele, 1997; Beilock et al., 2010).

An alternative theory is that demographic dynamics affect teacher expectations of students, and that these expectations have material influence on relevant student outcomes. Prior research has found that teacher expectations are influenced by demographic matching (Ouazad and Page, 2012; Ouazad, 2014; Gershenson et al., 2016). The impact of teacher expectations on students appears to be largely uncontroversial, but Mechtenberg (2009) develops a model of cheap-talk grading that generates the same kind of achievement gaps observed empirically.

Finally, it could be that teachers are less likely to exhibit biases against demographically similar students, either directly through biased grading behavior (Terrier, 2016; Lavy, 2008; Lavy and Sand, 2018) or through moderated responses to student misbehavior (Downey and Pribesh, 2004; Holt and Gershenson, 2017).

Ouazad and Page (2012) offers suggestive evidence that the effect of teacher gender on students may depend on the students as well. In an experiment designed to elicit student beliefs about teacher biases, they found that male students correctly expected female teachers to be biased against them, while female students incorrectly expected male teachers to be biased in their own favor.

[4]Stereotype threat posits that when an individual feels that they run the risk of confirming stereotypes about their social group, they become more anxious about their performance, and this may hinder their performance at a particular task.

Pinning down the active mechanisms is a significant empirical challenge. The data necessary to distinguish between different mechanisms is difficult to acquire. For instance, determining whether teachers demonstrate biases in grading behavior requires access to both teacher grades and anonymous grades, as in Lavy (2008), Terrier (2016), or Lavy and Sand (2018). Carlana (2019) uses the Gender-Science Implicit Association Test to measure teacher biases directly, and finds that biased teachers increase the gender gap in math performance in their classes. Bassi et al. (2018), using video of teachers in Chilean schools, finds that teachers pay more attention to, and interact more favorably with, boys than with girls. They find that this 'attention gap' is correlated with the gender gap in math scores in Chile.

3 Data

3.1 The National Evaluation of Teach for America

The data comes from the Mathematica Policy Research, Inc (MPR) National Evaluation of Teach for America (NETFA) Public Use File[5]. The NETFA was a field experiment 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 follow-up full-scale study conducted in Chicago, Los Angeles, Houston, New Orleans, and the Mississippi Delta during the 2002-2003 academic year. In total, 17 schools containing 98 classes and 1938 students took part in the experiment.

In each region, 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[6]. If a school-grade combination was selected for inclusion in the experiment,

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

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

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 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 schools that usually participate in the TFA program. 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, it includes class type (bilingual/monolingual), student demographic characteristics, class size, and math/reading scores both before and after treatment. For teachers, it includes demographic characteristics, type of teacher certification (nontraditional/traditional), and years of experience[7]. In addition to the baseline data, I construct a classroom-level indicator

[7]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.

variable for the presence of at least one disruptive student[8].

The test score variables deserve some further discussion. The data does not contain traditional test scores. Instead, there are 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. 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 more rapidly[9].

Some students in the sample have raw scores 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[10].

Table 1 reports summary statistics for the variables of interest. Note that the math

[8]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. Some classes contained students that are not part of the research sample, so some classes may be incorrectly coded as not disrupted.

[9]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.

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

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 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.

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.

While there are some significant differences in means between the full and estimation samples, most are quantitatively small. The only exceptions are in pre-treatment math and reading scores - and this is entirely due to the removal of invalid test scores[11].

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 statistically significant difference in the mean class size between the math estimation sample and the sample of students with invalid math scores. This is likely because larger classes have more chances to draw an invalid score, rather than there being a causal relationship between class size and invalid scores.

[11]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. The full-sample mean pre-treatment scores after removing invalid scores are essentially identical to the estimation sample means.

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)

Source: Author's calculations using Mathematica Policy Research, Inc National Evaluation of Teach for America Public Use File, 2001-2003.
Note: Table presents sample means with standard errors reported in parentheses. The full sample does not include two classrooms for which teacher gender is unavailable. The math and reading samples respectively exclude students with missing or invalid math and reading test scores in either the pre- or post-treatment periods.

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

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.

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 what population the resulting estimates applied to. However, since I will be estimating treatment effects conditional on pre-treatment test scores, 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.

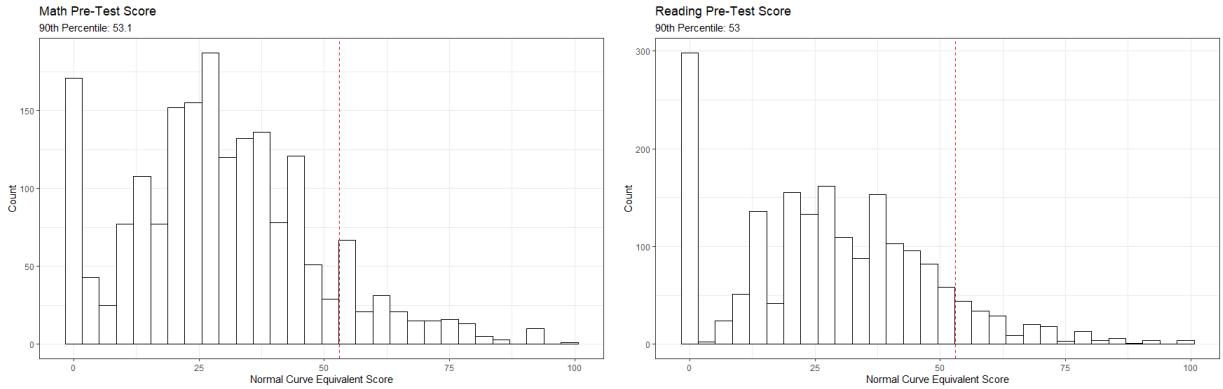


Figure 1: Pre-Treatment Test Score Distribution

4 Estimation Strategy

Capturing the heterogeneity of a treatment effect has traditionally been done through the estimation of quantile treatment effects (QTEs), which describe the difference between quantiles of the outcome distribution for untreated and treated individuals. QTE estimation, however, allows for heterogeneity in the treatment effect across sub-populations that are not identifiable given covariates. For example, the QTE of assignment to a female teacher might be positive for students in the 60th quantile, negative for students in the 40th quantile, and zero for those in the 50th quantile - but it may not be possible to determine *a priori* whether a particular student was in any of those quantiles. In the context of my investigation, this is undesirable - I am interested in how the treatment effect of assignment to a female teacher changes with specific covariates (gender and pre-treatment test scores).

Thus, instead of the QTE, I estimate the Conditional Average Treatment Effect (CATE) function. The CATE is defined as the value of the Average Treatment Effect (ATE) within a sub-population defined by specific covariate values. While the CATE is not an entirely new parameter, often appearing as an intermediate estimand for ATE estimation (Heckman et al., 1997; Hahn, 1998), treatment of the CATE as a parameter of interest is relatively recent.

The chief difficulty in identifying the CATE is that unconfoundedness probably does not hold when conditioning on a strict subset of the available covariates. In the context of this investigation, it is unlikely that unconfoundedness holds conditional on only student gender

and pre-treatment test scores. Abrevaya et al. (2015) provides a semi-parametric estimation procedure that accounts for this issue and allows for consistent estimation of the CATE parameter when conditioning on a subset of the covariates for which unconfoundedness does not hold.

I implement the Abrevaya et al. (2015) estimator and estimate the CATE of assignment to a female teacher conditional on pre-treatment test scores after splitting the sample by student gender, which recovers the CATE conditional on both student gender and pre-treatment test scores.

4.1 The Abrevaya et al. (2015) CATE Estimator

For compactness of notation, let Y_i be the post-treatment test score for student i, X_i be a vector of control covariates, and D_i be a binary indicator for the gender of student i's teacher ($D_i = 1$ if i was assigned to a female teacher, $D_i = 0$ otherwise). Let X_{1i} be a strict subset of X_i, containing only i's pre-treatment test score, and an indicator for i's gender. Formally, the CATE is defined as

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

This parameter captures how the average treatment effect $\mathbb{E}\left[Y_i(1) - Y_i(0)\right]$ depends on the covariates contained in X_1 - in this context, how the effect of assignment to a female teacher changes with student gender and pre-treatment test scores. The Abrevaya et al. (2015) estimator of the CATE is

$$\hat{\tau}(x_1) = \frac{\frac{1}{nh^l}\sum_{i=1}^{n}\left(\frac{D_i Y_i}{\hat{p}(X_i)} - \frac{(1-D_i)Y_i}{1-\hat{p}(X_i)}\right) K_1\left(\frac{X_{1i}-x_1}{h_1}\right)}{\frac{1}{nh^l}\sum_{i=1}^{n} K_1\left(\frac{X_{1i}-x_1}{h1}\right)} \tag{2}$$

where $K_1(\cdot)$ and h_1 are respectively a kernel function and a bandwidth, l is the dimension of the vector X_1 (in this case, $l = 1$ because I condition on gender by splitting the sample,

leaving only pre-treatment test score in 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. (2015) show that this estimator is asymptotically consistent for the CATE under the familiar unconfoundedness and sampling assumptions necessary for ATE estimation.

4.2 Identification Strategy

Intuitively, the identifying assumptions require that students who are assigned to a female teacher are comparable to students assigned to male teachers, conditional on pre-treatment test scores, student gender, and other covariates. If, for instance, students in one region had much stronger gender stereotypes and were also more likely to be assigned to a female teacher, unconfoundedness would likely fail. Without controlling for region effects in the estimation, the estimated effect of assignment to a female teacher would be biased downwards.

A major upside of the data used is that conditional on randomization block, students were assigned to teachers totally at random. This means that a number of potential confounders, like better students being assigned to female teachers[13], are not concerns. However, since the randomization is not unconditional, some potential sources of confounding remain. In particular, while students were randomly assigned to teachers, teachers were not randomly assigned to preparation pathways (i.e. TFA and non-TFA teachers are likely to be different), nor were they randomly assigned to schools or grades (i.e. it may be that teachers in one school are different from those in another school).

For TFA teachers, dealing with these issues is straightforward. TFA applicants in the experiment provided regional preferences, which allows for teachers to differ across regions, but not across schools within a region[14].

[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 requires complicated transformations of discrete variables. In addition, it 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.

[13] Clotfelter et al. (2006) finds that male teachers are more likely to be assigned students with lower math and reading scores, so this would be a real concern with purely observational data.

[14] To be more specific, TFA applicants reported regional preferences as well as preferences for level of

For non-TFA teachers, non-random assignment of teachers to schools or grades poses a more difficult 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 0 or 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 "to obtain estimates of the propensity score that balance the covariates between treated and control samples" (Imbens and Rubin, 2015). In finite samples[15] 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

education (e.g. primary/middle/high school levels). Since the experiment considers only primary school students, the latter preferences cannot introduce confounding. I thank the TFA administrators for a thorough explanation of the application process at the time of the experiment.

[15]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.

role in treatment assignment. I thus estimate the propensity score with the following logistic regression:

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

where $FTEACH_i$ is an indicator for assignment to a female teacher, $SC^{'}$ is a vector of student covariates, $TC^{'}$ is a vector of teacher characteristics, $R^{'}$ 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. Full details of this specification can be found in the appendix, where I also consider some alternative specifications for the propensity score.

One potential issue facing any investigation that uses inverse probability 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 importance of 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 is not 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). The main specification is robust to different trimming behavior - in particular, dropping students with extreme propensity scores instead of changing their propensity scores does not have a noticeable effect on the results. One alternative specification in the appendix depends more strongly on trimming behavior.

4.3 Choice of Smoothing 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 intervals are constructed using the asymptotic approximations from Abrevaya et al. (2015)[16]. 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 small, despite being a boundary point. This is largely due to the significant mass of students

[16]To the best of my knowledge, construction of uniformly valid confidence intervals for the Abrevaya et al. (2015) estimator is an open problem.

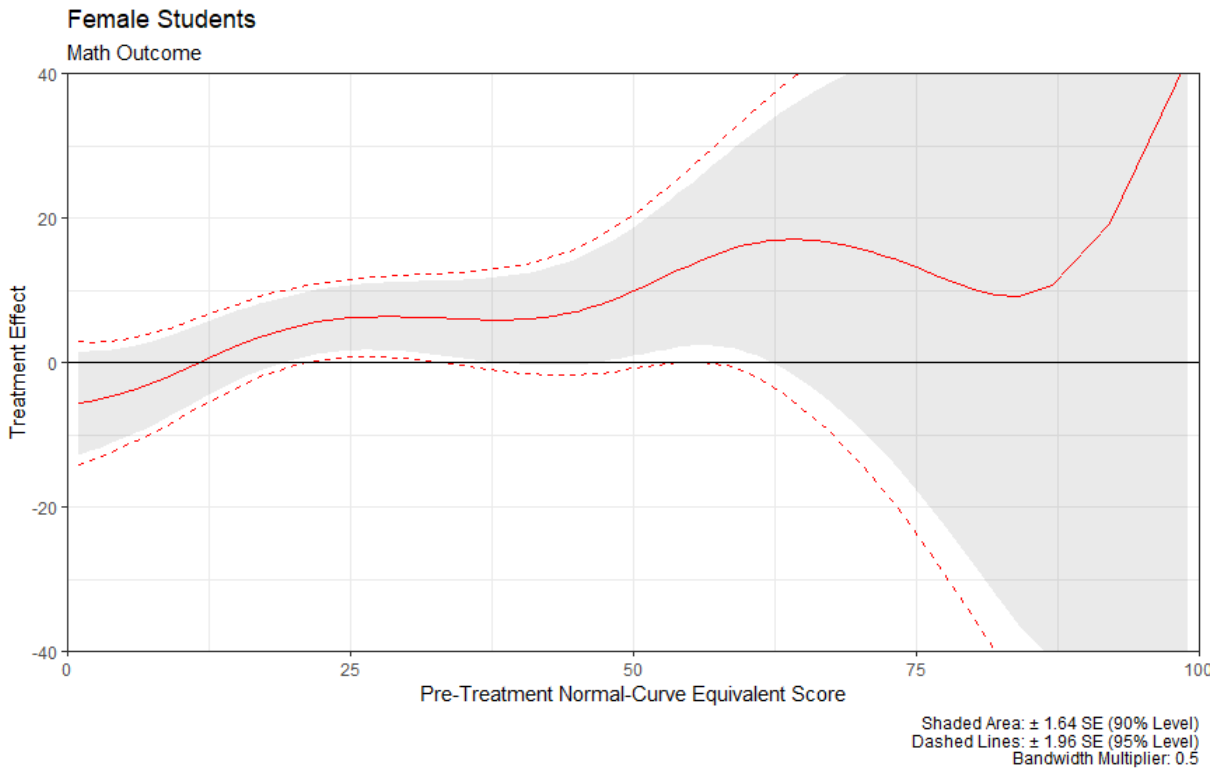


Figure 2: CATE (Math) for female students

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 likely that uniformly valid confidence bands would be wider, and might not reject the hypothesis that the true effect of assignment to a female teacher is a *constant* zero across the pre-treatment test score distribution.

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 than female students of higher ability. 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.

Indeed, particularly for TFA teachers, a possible conjecture is that students with higher ability are easier to teach effectively[17].

The implied average treatment effect[18] 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 statistical 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[19]. Restraining the calculation to consider only point estimates below 55, thus excluding potentially extreme point estimates driven by lack of data, the implied ATE decreases to around 0.19 standard deviations (3.4 on the normal curve scale).

Figure 3 depicts the estimated CATE function for male students, again with math scores as the outcome of interest and conditioning covariate. The increase in the size of the confidence intervals starts earlier than in Figure 2, primarily because the male pre-test score distribution is skewed to the left relative to 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.

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

[18]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.

[19]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.

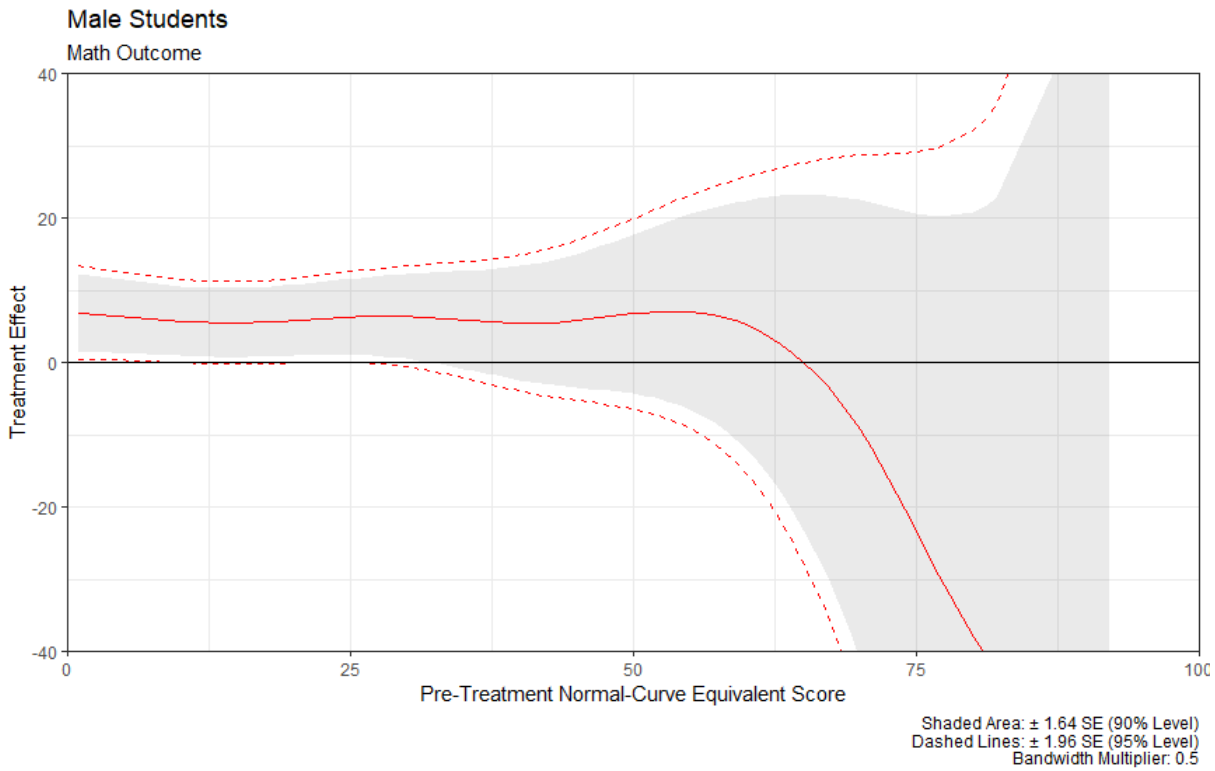


Figure 3: CATE (Math) for male students

The implied ATE is approximately 0.25 standard deviations (4.7 on the normal curve scale). Considering only pre-test scores below 55 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 unlikely 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 covariate. 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

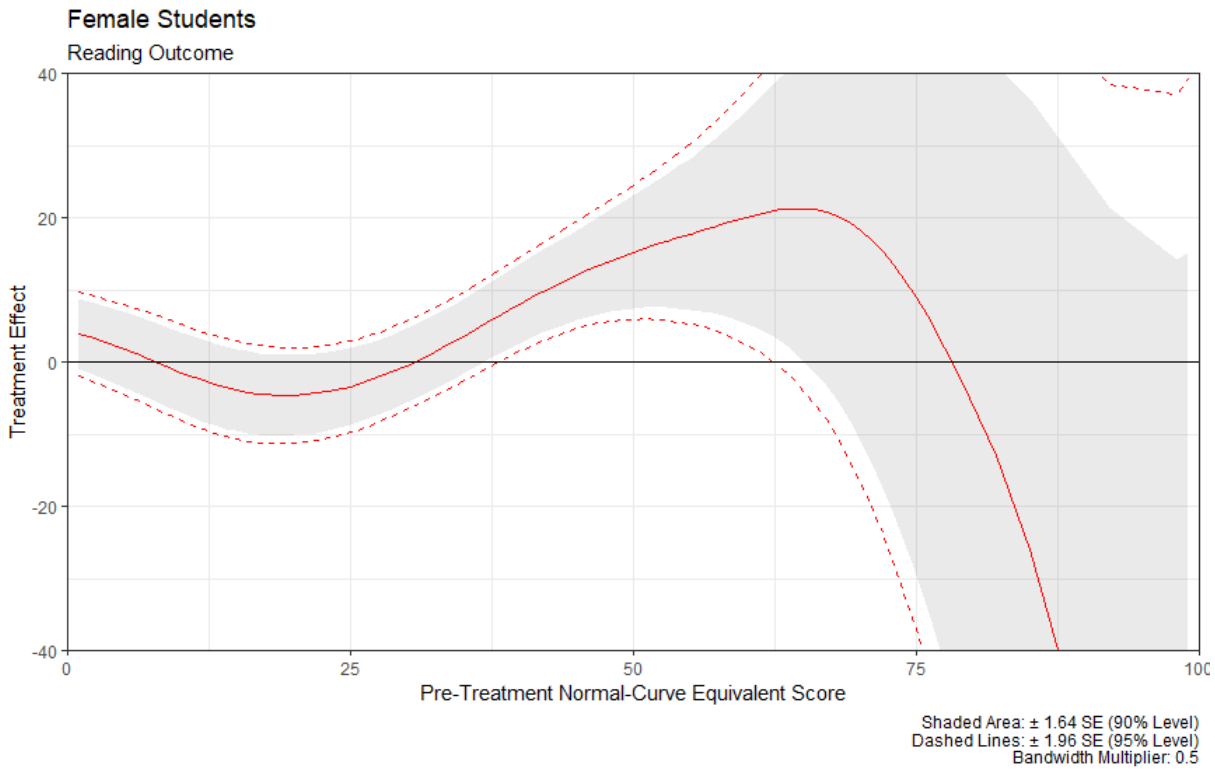


Figure 4: CATE (Reading) for female students

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 see a significant and large positive treatment effect.

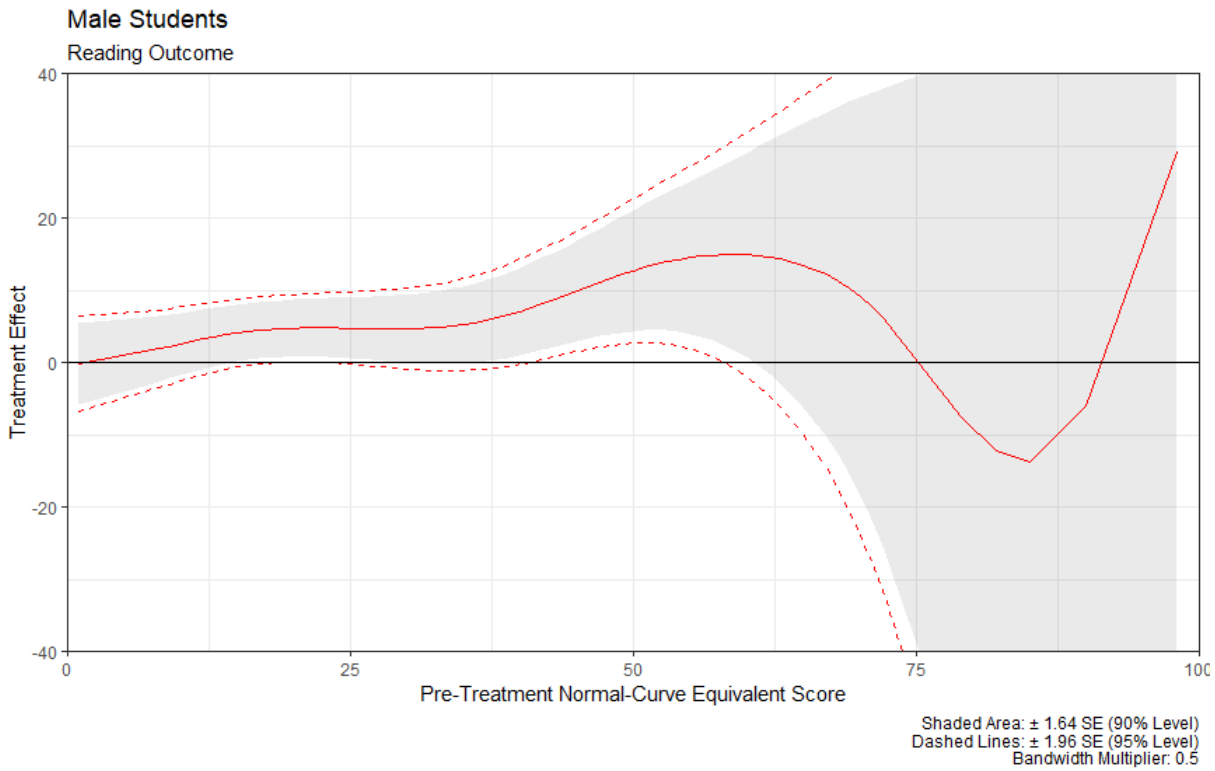


Figure 5: CATE (Reading) for male students

For male students, the story appears largely the same as before. There is limited heterogeneity (although potentially more than in math). The estimated CATE is positive for almost 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 heterogeneity on ability. Role model effects, for instance, might be stronger for high-ability students, or stereotype threat effects on women in math may be more powerful

at the low end of the ability distribution. However, it is also 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[20]. Figure 6 presents the estimated CATE functions conditional on class rank for the four subsamples.

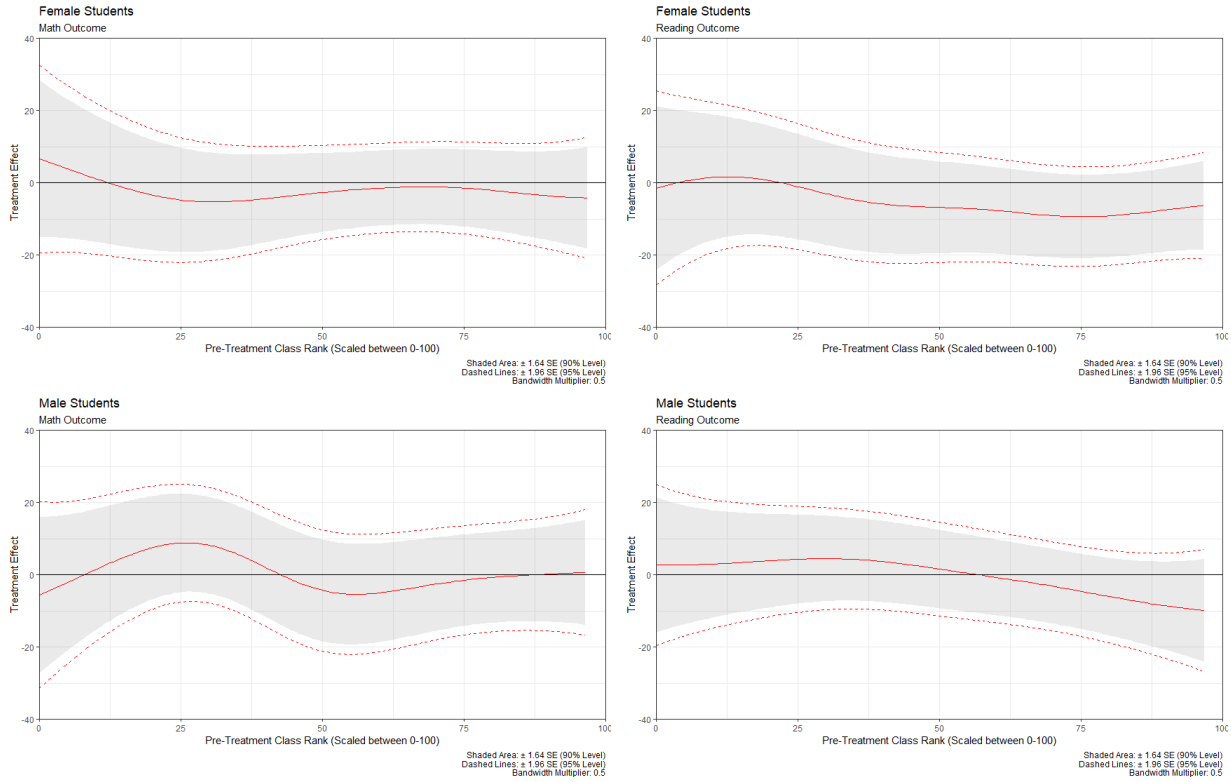


Figure 6: Conditioning on Class Rank

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

[20]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.

the graphs is similar to before - and the results suggest that within-class performance is not correlated with the size of the teacher gender effect. Even with point-wise valid confidence bands, the hypothesis that the true effect conditional on class rank is a constant zero cannot be rejected in any sub-sample at the 95% level.

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 labeling 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 women.

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, Clotfelter et al. (2006) 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 differ 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 what can 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. (2015) also provide suggestive evidence that the mechanism underlying their results is powered by 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 students. 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. This may be too early for gender stereotypes to strongly affect gender dynamics between students and teachers, although Antecol et al. (2015) suggests otherwise. Different levels of schooling, and a sample 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 at all consistent with the bias found in Cappelen et al. (2019).

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.

Clotfelter, C. T., Ladd, H. F., and Vigdor, J. L. (2006). Teacher-Student Matching and the Assessment of Teacher Effectiveness. *Journal of Human Resources*, 41(4).

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.

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.

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.

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.

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).

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.

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).

Williams, W. M. and Ceci, S. J. (2015). National hiring experiments reveal 2:1 faculty preference for women on stem tenure track. *Proceedings of the National Academy of Sciences*, 112(17):5360–5365.

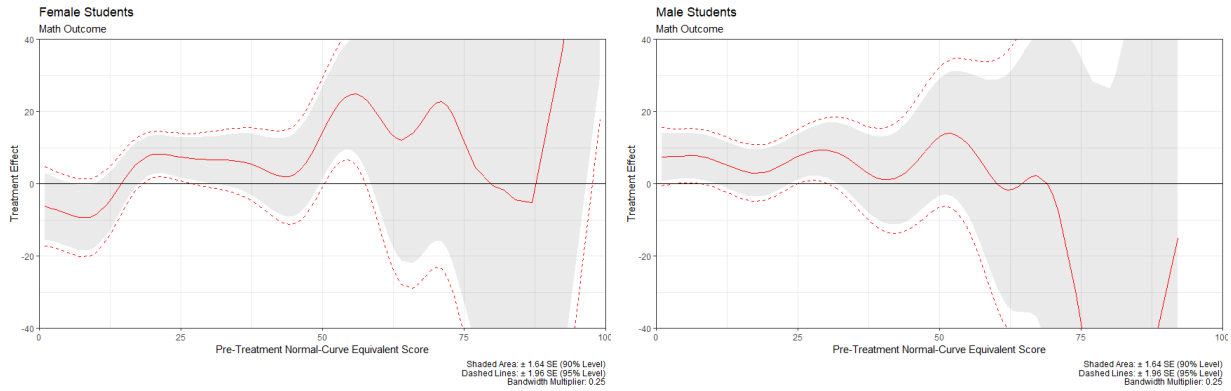
Winters, M. A., Haight, R. C., Swaim, T. T., and Pickering, K. A. (2013). The effect of same-gender teacher assignment on student achievement in the elementary and secondary grades: Evidence from panel data. *Economics of Education Review*, 34:69 – 75.

8 Appendix

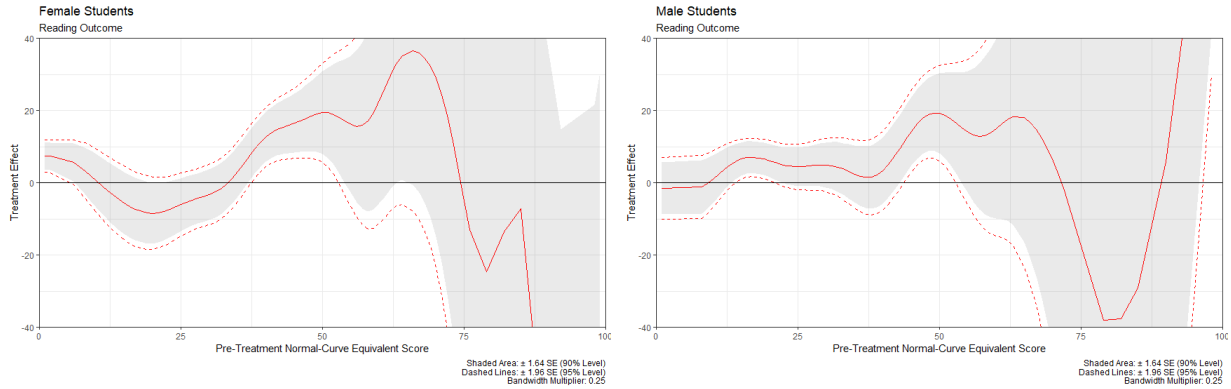
8.1 Bandwidth Choice

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 over-smooth to the extent that it becomes no more informative than an ATE estimator.

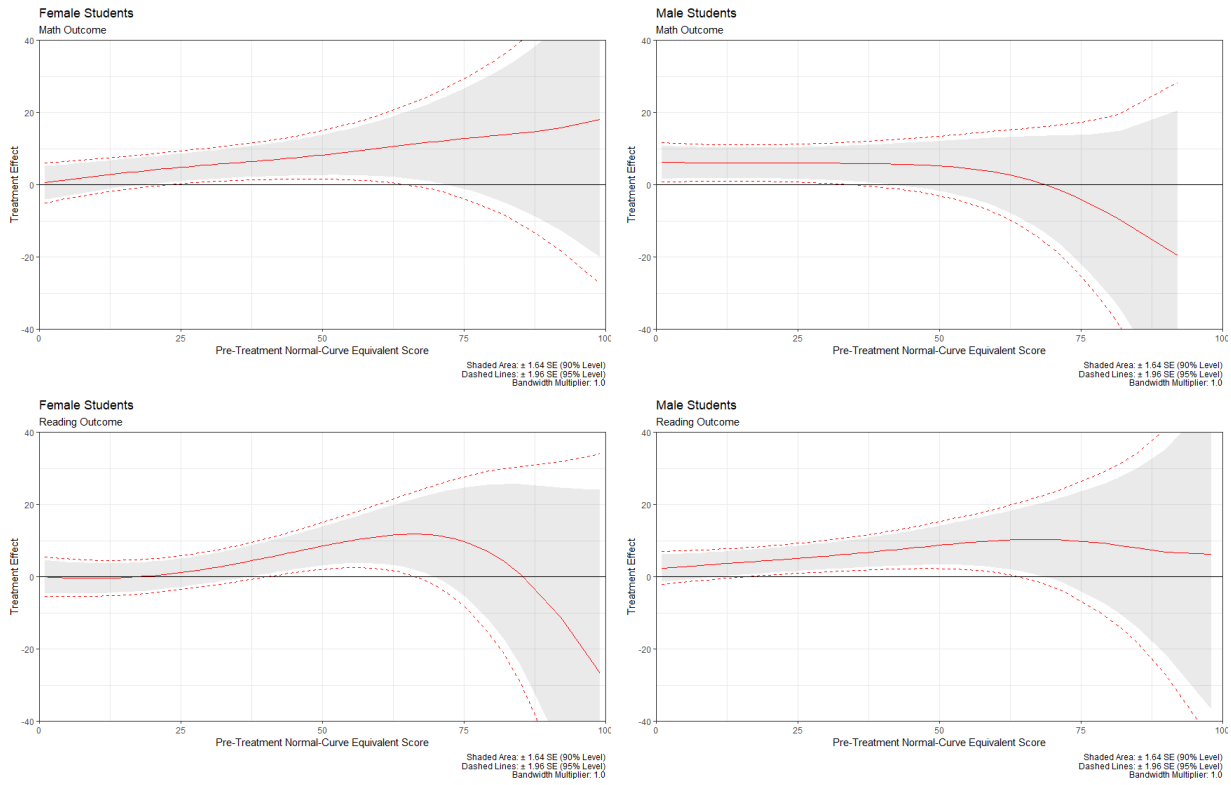
Recall that my main specification sets the bandwidth multiplier to 0.5. Setting the bandwidth multiplier to 0.25 causes the estimated CATE function to be significantly less smooth. Qualitatively, however, the story is largely unchanged. The worst-performing female students see a negative effect of assignment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



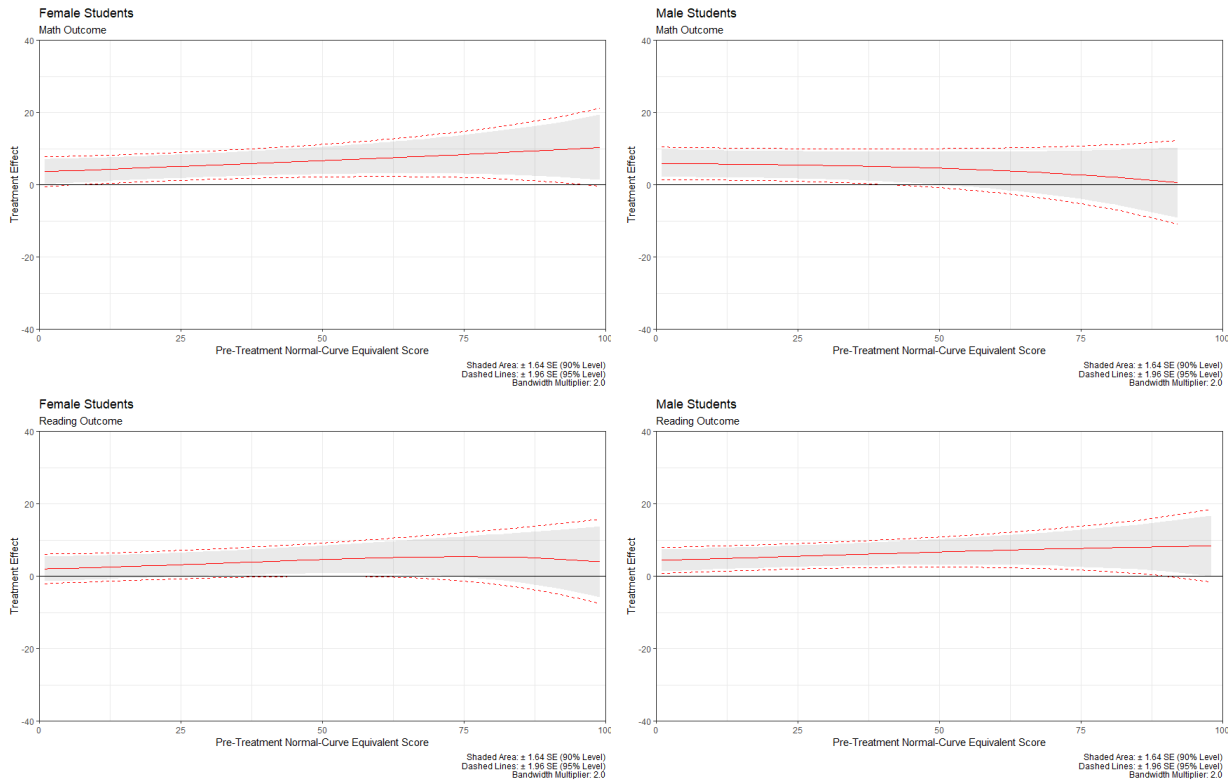
The effect of reducing the bandwidth multiplier is nearly identical for reading outcomes. The qualitative story of the estimated CATE function is largely unchanged - significant effects are observed in roughly the same places, and the general shape of the function is similar. Again, there appears to 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 over-smoothing. 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:



8.2 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:

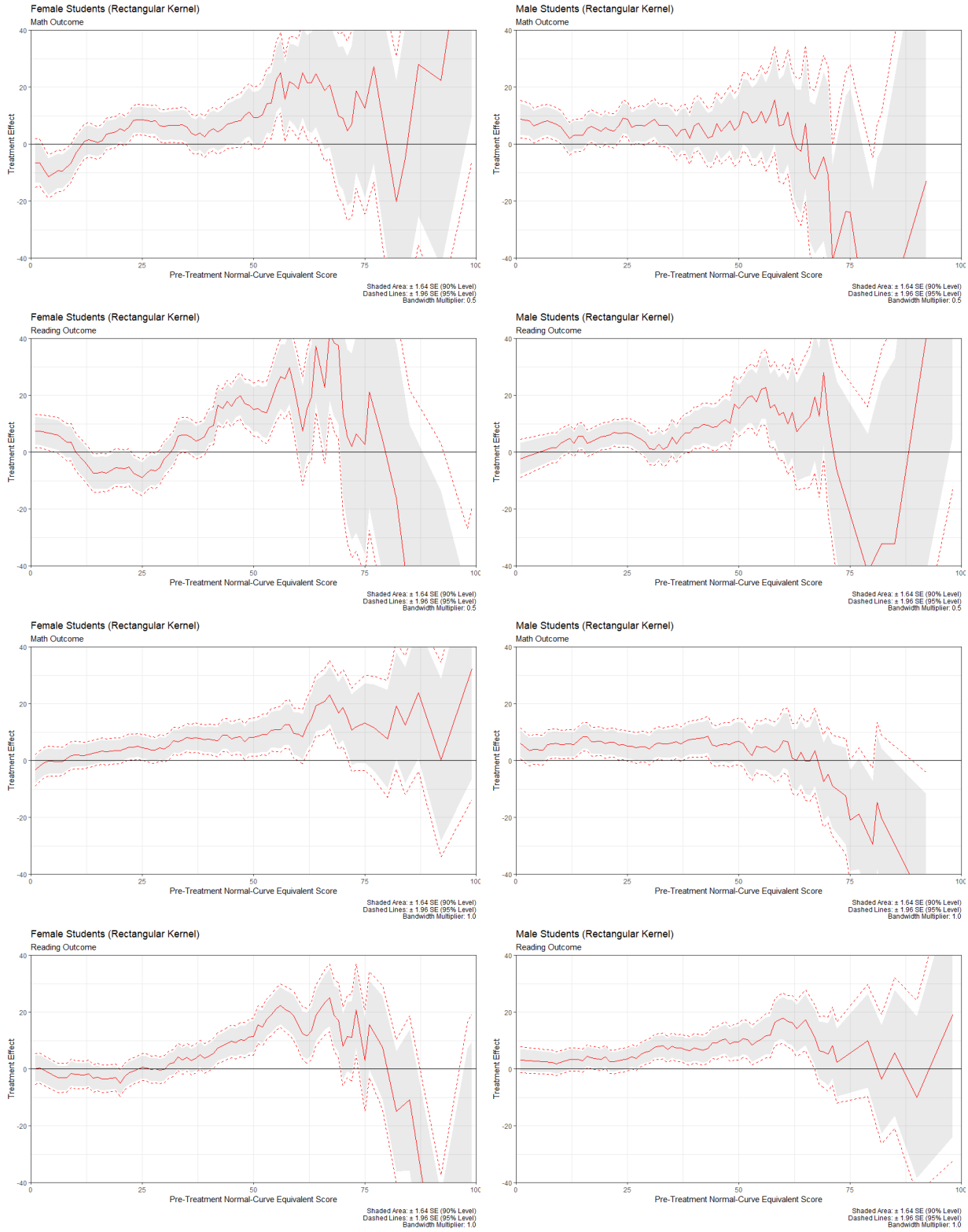
$$K_r(u) = \begin{cases} \frac{1}{2} \text{ if } |u| \leq 1 \\ 0 \text{ otherwise} \end{cases}$$

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

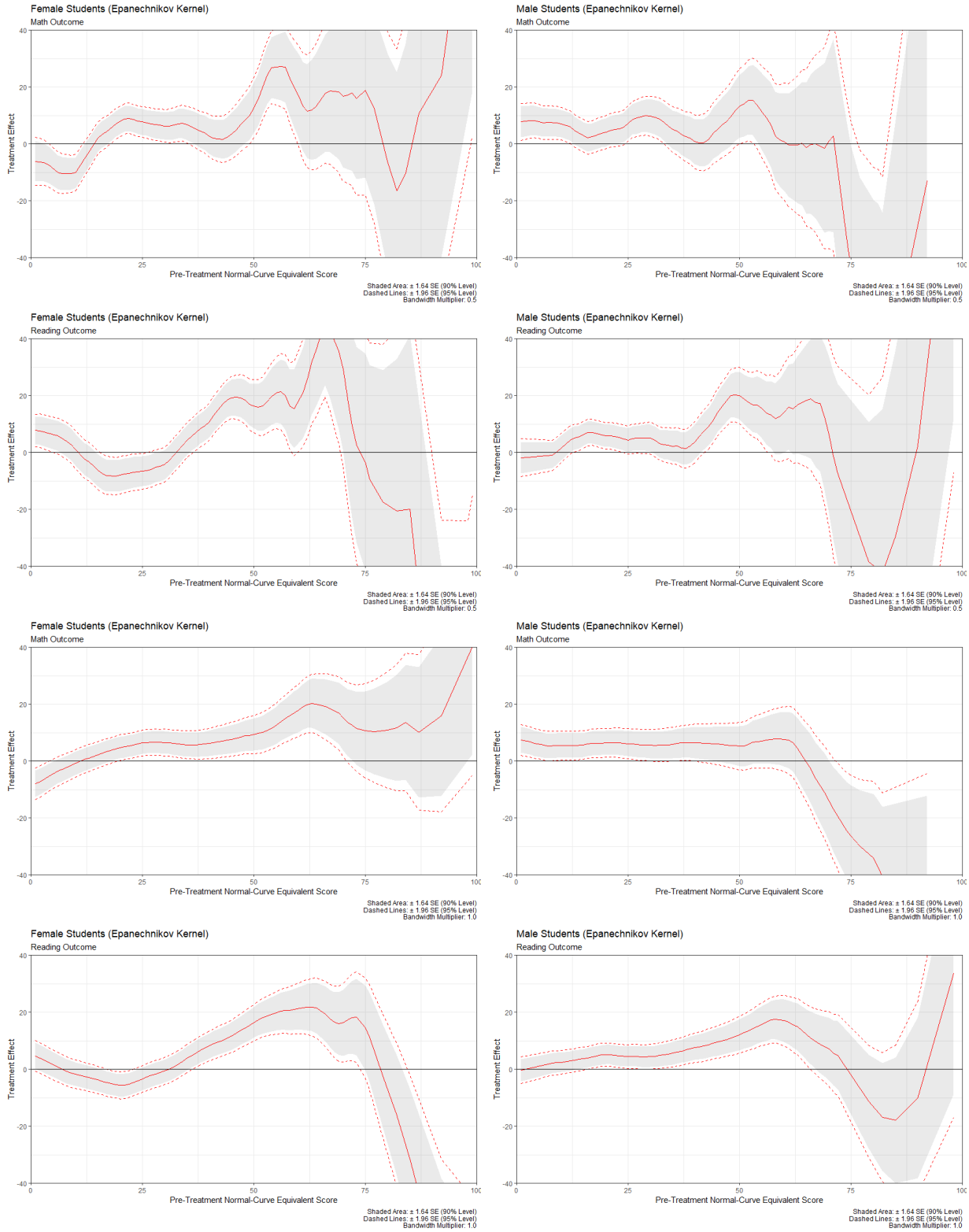
The primary difference between these kernels and the Gaussian kernel is that weights decrease towards zero more rapidly, particularly with the rectangular kernel. This results in less smooth estimates of the CATE function, but the qualitative story is largely unchanged. The effect of bandwidth choice 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:



8.3 Propensity Score Estimation

8.3.1 Details of the main specification

Recall the main specification:

$$\ln \frac{P(FTEACH_i = 1)}{1 - P(FTEACH_i = 1)} = \beta_0 + \beta_1 SC_i^{'} + \beta_2 TC_i^{'} + \beta_3 R_i^{'} + \beta_4 TFA_i + \beta_5 CS_i + u_i$$

$SC_i^{'}$ is a vector of student characteristics. It includes indicators for a student being black or Hispanic, the relevant pre-treatment test score in math or reading as measured on the normal curve equivalent scale, and an indicator for whether the student's class contained a disruptive student.

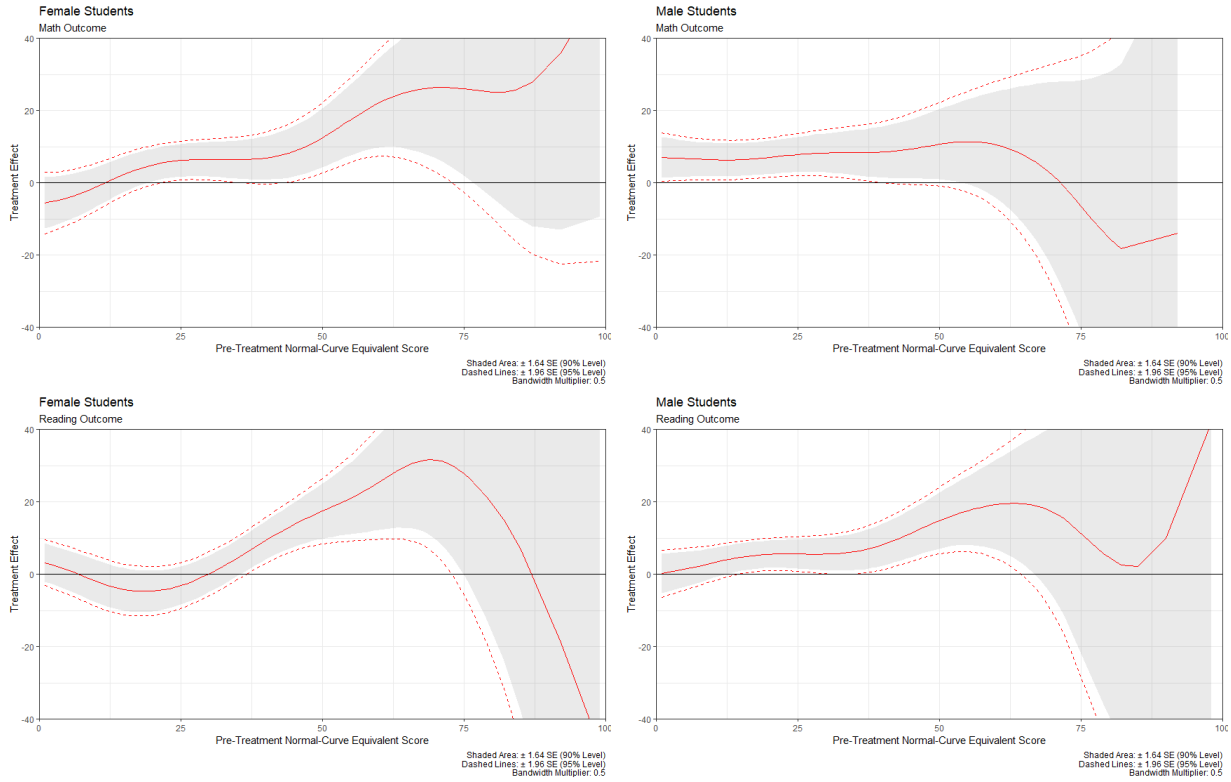
$TC_i^{'}$ contains the teacher's experience measured in years as well as indicators for whether the teacher was black or Hispanic. In some of the following alternative specifications, it also includes an indicator for possession of a regular teacher certification.

$R^{'}$ is a vector of region indicators. There were 6 regions in the experiment, containing 7 school districts because the Mississippi Delta contributed two school districts. TFA_i is an indicator for whether the teacher was a TFA teacher or not. CS_i is the class size, measured as the number of students in the class at the end of the year[21].

8.3.2 Alternative specifications

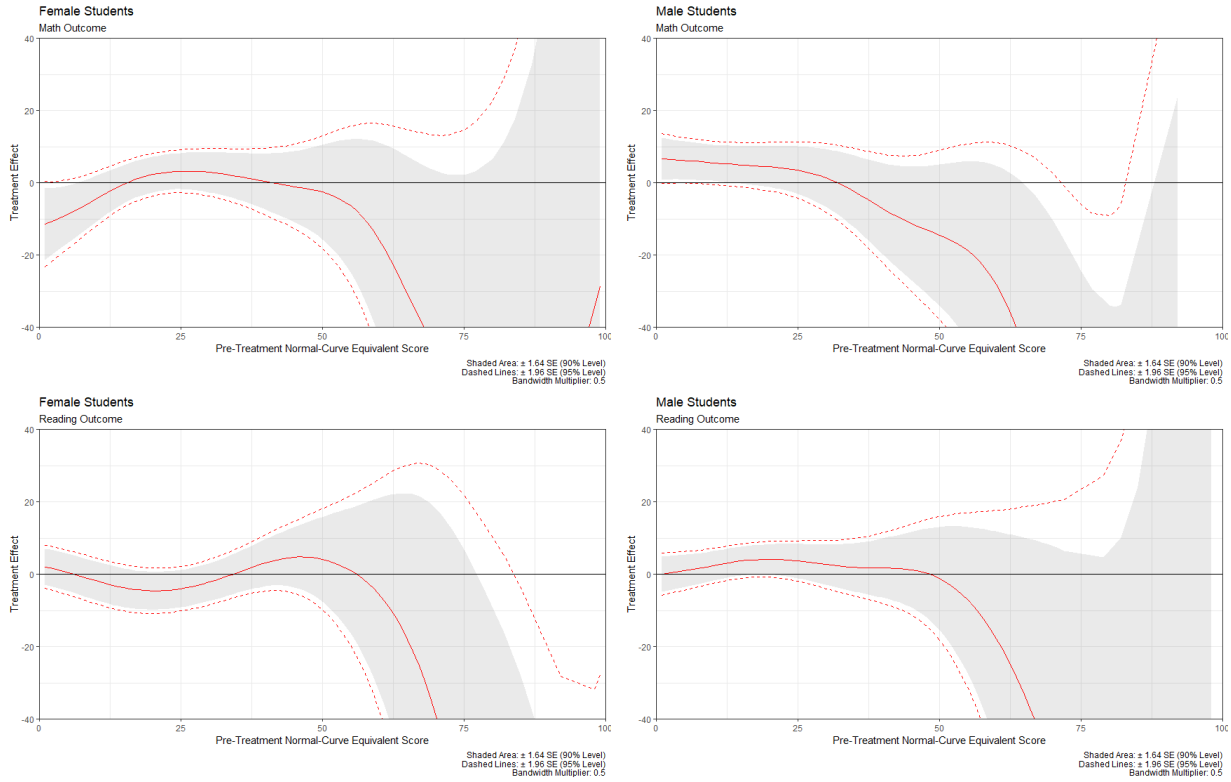
First, I consider the addition of the indicator for a traditional teacher certification. The main specification excludes this variable because previous research (e.g. Staiger and Rockoff (2010)) suggests that teacher certifications are not good predictors of teacher quality, and thus balancing of samples 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.

[21] This is the 'true' class size in that it counts students that are not part of the research sample.



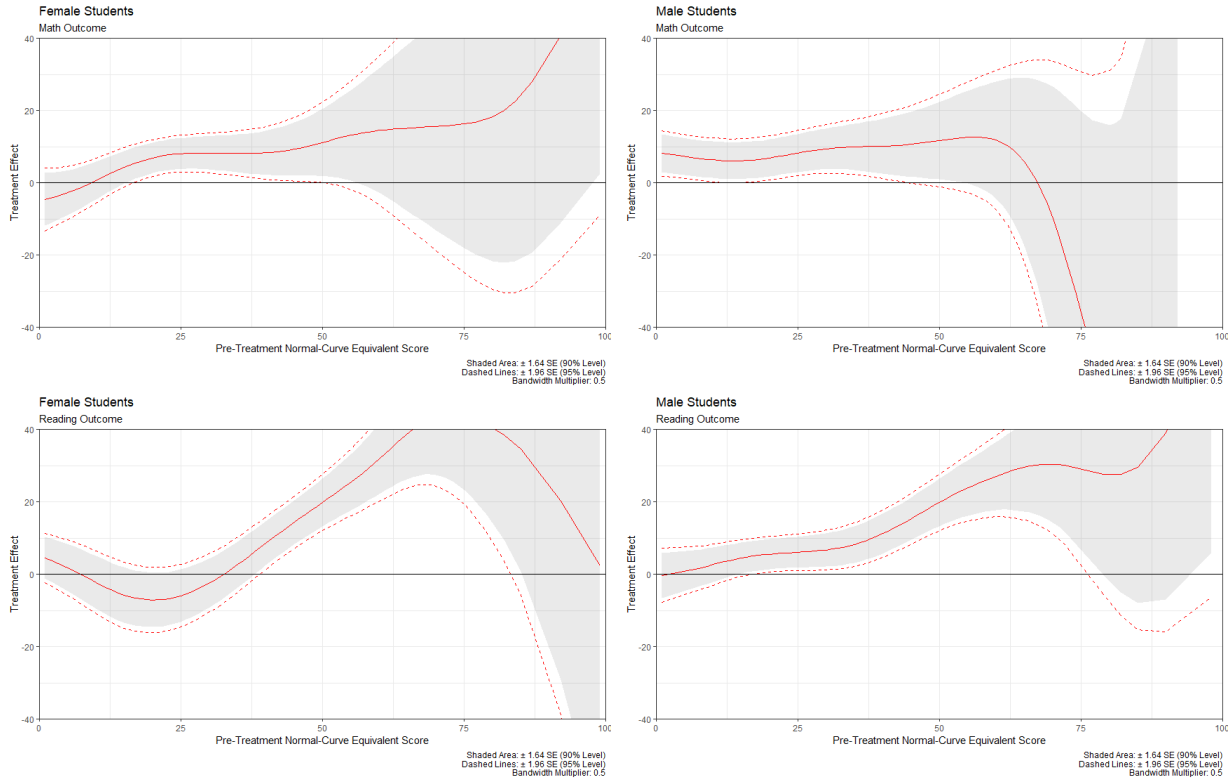
I omit reports for other bandwidth multipliers because the results of that exercise are identical - larger confidence bands, with no significant change to the underlying function.

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 trade-off. 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.

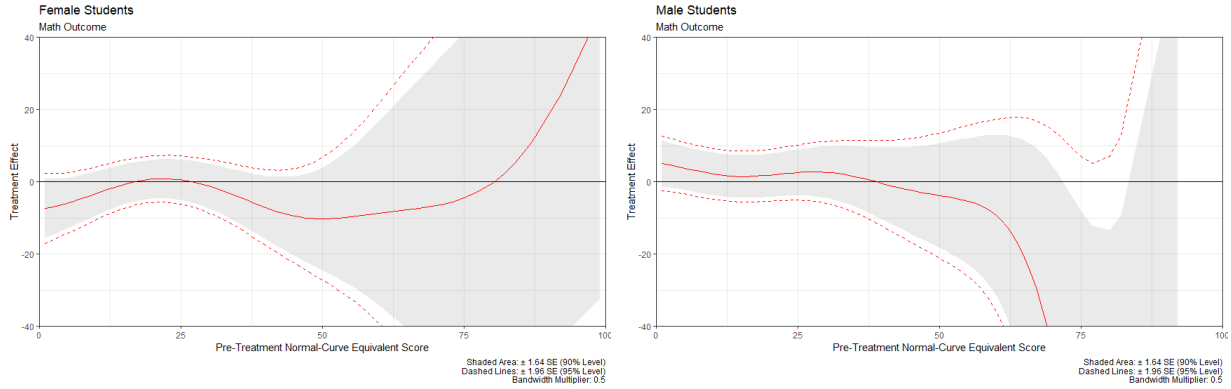


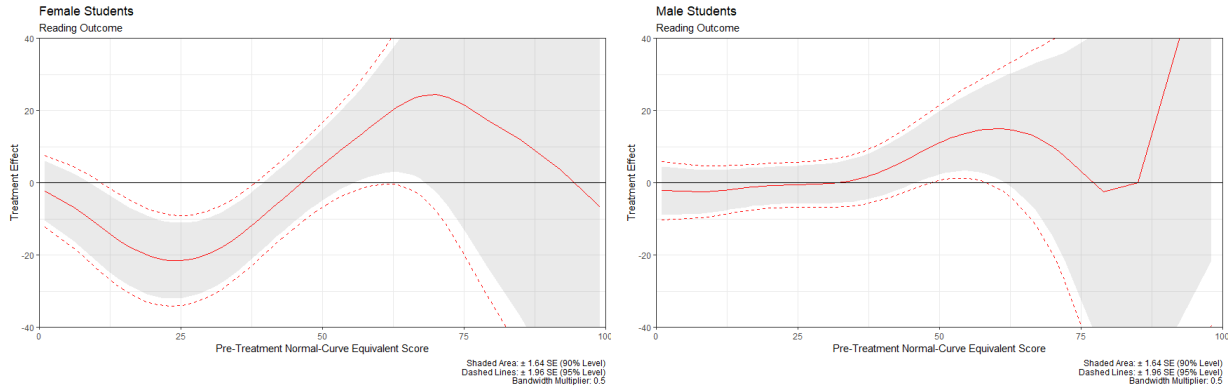
While the results for male students are very marginally consistent with the results from my main specification, particularly in math, it is clear that (as prior research would suggest) the demographic variables excluded in this specification are relevant. If they were irrelevant or had a sufficiently 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.



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.