What Makes Teachers Better? Evidence from a Long Panel of Classrooms

James Brand *

September 3, 2019

Abstract

In this paper I construct a long panel of classrooms linked to teachers, providing a more detailed picture of teacher experience than is available in most studies. I demonstrate that heterogeneity of past experience is predictive of current teacher effectiveness. First, I find that teachers who have taught more or larger classes (i.e. more students) tend to be more effective than their less experienced peers. Second, conditional on their years of experience, the number of classes they have taught, and their initial effectiveness, teachers who have previously been assigned more students in the past are more effective at improving students' math scores than those who have taught fewer. This result implies that teachers improve more by teaching larger classes, which is in conflict with the existing literature showing the importance of small class sizes for student achievement. I also find that teacher and student demographics affect the rates at which teachers improve. Even conditional on the number of students she has previously taught, the demographics of a teacher's past students predict her current effectiveness. Further, I show some evidence that teachers improve, and students benefit from their teachers' experience, at different rates according to a teacher's demographic similarity to the students she has previously taught.

*PhD student at the University of Texas at Austin. email: jamesbrand@utexas.edu. Thanks to Daniel Ackerberg, Esteban Aucejo, Jorge Balat, Sandy Black, Kevin DeLuca, Richard Murphy, and Steve Trejo for many helpful comments. This paper is a modified version of my second-year paper at the University of Texas at Austin.

1 Introduction

In this paper I study the determinants of teacher effectiveness using a long panel of classrooms linked to teachers in North Carolina. In particular, I focus on the extent to which the characteristics of the classrooms and students a teacher has been assigned in the past predict her current effectiveness. Although the rate at which teachers improve through experience has been widely studied, it has largely been treated as a secular trend, rather than an outcome of teachers' individual teaching histories. The same is true of the value-added literature, which has focused on teacher effectiveness as measured by average (within-teacher) effects (Aaronson et al., 2007; Rivkin et al., 2005) or as effects which trend exogenously over time (Chetty et al., 2014a,b). This modeling approach, motivated in large part by data limitations, has little to say about the causes of teacher growth, which are crucial for policymakers interested in producing better teachers and in the optimal assignment of teachers to particular (e.g. low-achieving) schools or students.

Many authors have shown the importance of class size (Angrist and Lavy, 1999; Clotfelter et al., 2007) and the characteristics of peers (Murphy and Weinhardt, 2014; Hoxby, 2000) in determining students' test scores. These document an important feature of the education production function: even conditional on a teacher-student pair, the number and composition of the other students in the classroom affect a student's outcome. There is also evidence that teachers are more productive when they share a demographic characteristic with their students. Dee (2004) finds positive effects of a match between student and teacher race; Ehrenberg et al. (1995) find that teachers' evaluations of students improve on average when the student and teacher are of the same race and gender. Gershenson et al. (2016) present evidence of racial bias in teachers' expectation about students' future educational achievement, and Papageorge et al. (2016) show that these kinds of biases have causal effects on students' college completion probability.[1] Thus, not only does teacher effectiveness depend on the characteristics of her classroom, it also appears to depend on her own characteristics and those of each student she is assigned.[2]

What has been understudied to date, and is examined in this paper, are the ways in which these same classroom characteristics may affect teachers' productivity in future classes. One notable exception is Ost (2014), who demonstrates that a teacher's effectiveness in a given grade depends on the grades she has taught in the past, and that this grade-specific experience depreciates over time. This is an important insight into the mechanisms for teacher growth; even conditional on the number of years a teacher has worked, her effectiveness in her current classes depends on the characteristics of the classes she has previously taught. In this paper I use detailed classroom-level data from North Carolina to study this relationship at a more granular level. Specifically, I construct a detailed history of the number of classes, students, and students of different races and genders previously taught by a large sample of teachers, and relate these measures of experience to the achievement of their current students. This allows me to study the rate at which teachers improve as they teach more students and classrooms, even among teachers with the same years of experience.

If teachers learn how to teach through their interactions with students, then it is natural to think of teachers who have taught more students or classes per year as having more experience than those who have taught relatively few. This simple framing of growth predicts that teachers with more student- or class-intensive careers will, on average, be more productive in the future. The results in section 4 are consistent with this story. My main estimates, which define teacher experience as the combination of years teaching and the total number of classes and students previously taught, imply

[1] Though not exactly a *demographic* match, Grönqvist and Vlachos (2016) also find some evidence high-achieving students benefit more from teachers' cognitive ability than low-achieving students.

[2] Recently, Aucejo et al. (2018) and Lavy (2015) show that the effects of classroom composition and student demographics (respectively) on a teacher's effectiveness also interact with the teaching practices she employs. I have no data on teaching practices, and thus cannot explore their interaction with my effects of interest.

that specifying teacher experience only as a function of years understates teachers' predicted effectiveness by nearly 50% for the median teacher with 15 years of experience.[3] This result is due in large part to the returns to teaching more students specifically. Because this effect is estimated conditional on the number of classes previously taught, it implies that teachers tend to improve most by teaching large classes. Most existing estimates find that students benefit from small classes, meaning that policy-makers face a dynamic trade-off: making a classroom bigger today hurts a teacher's current students' test scores but improves her future students' scores by making her more effective in following years.

To account for the endogenous sorting of students and teachers into classrooms and schools, I estimate all effects of interest conditional on teacher, school, and grade-year fixed effects, which permits some teachers and schools to have consistently larger classes than others. I also present estimates for all results which further condition on teachers' years of experience or on student fixed-effects. My main results are robust to each of these additions. Finally, still conditioning on years of experience, I instrument for the number of students taught with the number of students in the school-grade in the previous year. The validity of this instrument requires only that changes in average unobservable teacher quality at the school-grade-year level are uncorrelated with changes in student cohort size. Each of these specifications indicates significant growth as a function of students taught, implying substantial heterogeneity of teacher effectiveness even conditional on years of experience.

Given the aforementioned literature on demographic matching, we may expect teacher growth to also depend on demographic matches; for example, if female teachers are better able to teach female students, they may also learn faster from them. In section 5 I interact demographic-specific experience variables separately with the relevant

[3]This comes from a calculation of the effect of the total number of students taught from column 6 of table 5, which I compare to the main estimates from Ladd and Sorensen (2017) of the returns to years of teaching experience.

student and teacher demographics, and present p-values of the null hypotheses that (a) all interactions of experience variables with student demographics and (b) all interactions with teacher demographics are zero. I run these regressions and tests separately for race- and gender-based definitions of experience, and find consistent evidence for heterogeneity. When experience is defined by gender, I find that male teachers benefit significantly less from experience with female students than do female teachers (and symmetrically for female teachers and male students). I find no evidence that male (female) students benefit more or less from a teacher's experience with female (male) students than do female students. That is to say, my results indicate that teachers improve faster when exposed to students of the same gender, but future male and female students benefit from these gains equally.

When experience is instead defined by race, I find mixed evidence against both null hypotheses, indicating that even conditional on initial quality and years of experience, the rate at which a teacher improves, as well as her effectiveness with any given student, may depend on her race, the student's race, and the races of the students she has taught in the past. Because my estimates of individual interaction terms are imprecisely estimated, I can not confidently determine the signs of these differential effects, but I take this set of tests, as well as those for gender-specific experience, as evidence consistent with the importance of teachers' familiarity with students of different demographics in determining their effectiveness in a given classroom.

2 North Carolina Administrative Data

The data I use come from the North Carolina Education Research Data Center (NCERDC) and have been used in a plethora of existing economic research concerning education. For this paper, I use data from fiscal years 2007-2011, and observe all public school districts in North Carolina and test scores from the end-of-grade tests (for grades 4-8), which are state-wide standardized tests. In this section I describe the methods I use to match students to the appropriate teachers and my construction of all experience-related variables.[4]

2.1 Classroom Assignment Data

In order to construct the vector of experience which is of interest to this study, I use classroom composition files from 1995-2011, which detail the number and demographic characteristics of students in every classroom and official school activity, each of which is associated with a teacher.[5] I divide students into four race-based categories (white, black, Hispanic, and all other races),[6] and construct course length-weighted[7] total numbers of students for each race-gender bin, as well as the number of academically gifted students, for each teacher-year. I drop classes with more than 35 students in order to avoid counting non-academic subjects and to focus on standard classroom environments[8]. Only 0.3% of classrooms from 1995-2011 fall above this threshold. It is important to note that, because the classroom composition files are anonymous, any students which have a teacher for more than one class in a year and/or more than one year will be counted twice. There is no way to ameliorate this problem with the NCERDC data for such a long panel of teachers, so to the extent that this is a frequent issue in the data and teachers gain experience differently from students to whom they

[4]More information about my data construction, as well as my main results estimated under an alternative data construction method, can be found in the Appendix.

[5]I will use "classroom composition" and "classroom assignment" as synonyms for this set of files.

[6]I categorize students' races in this way because white and black students make up 80% of the students in my sample and the other two groups each make up approximately 10% of my sample.

[7]I drop the small subset of courses which last less than a year. I also note that this weighting approach means that I count a semester-long course with 30 students the same as a two-semester course with 15 students. My results change little if I ignore course length in my count of students, but weighting in this way permits easier interpretation of the counts I use as "number of student-years" of experience.

[8]For example, much of the discussion of the importance of class size in determining test scores considers differences between 10 and 20 students in a classroom. See, for example, Hanushek (1997) and Krueger (2003). Many of the classes containing more than 35 students have titles like "Self Study," "Health," "Band," or "Physical Education," so most would have been dropped later in my data construction, regardless.

have previously been exposed, I will over-count teachers' student-years of experience. Therefore, my estimates of the returns to teaching more students may understate the true returns.[9]

Although it comes at a considerable cost in terms of sample size, I only conduct analysis on teachers whose careers began in or after 1995, the year in which the classroom assignment data begins, to ensure that I can construct an accurate teaching history for each teacher. I also restrict my sample only to the set of teachers who never teach outside of North Carolina, and for the remainder of the paper all regressions and discussions of the data will refer to this restricted subset on which my definition of years of experience is most valid.[10] This sample restriction is similar to that in Ost (2014), who also uses the classroom assignment data to construct experience. Tables 1 and 2 present sample means of selected observable variables for the full and restricted samples. My sample of students is more likely to be black, economically disadvantaged, and lower achieving compared to the population, and my sample of teachers is more likely to have a lateral entry license and be male and/or black. Teachers in my sample are also younger on average mechanically, which also produces the difference in educational achievement relative to the population.

One benefit of this data construction is an alternative method to construct teachers' years of experience. Although most teachers can be linked to their experience through administrative pay codes, this is not the case for all teachers in my sample. For teachers who are not matched to the pay code data, I construct an alternate measure of experience by counting the number of years I observe a teacher in the classroom assignment files. This increases my sample size by nearly one hundred thousand student-year observations, and the two experience definitions are highly correlated. Whenever I report

[9] The files I use to match test score data to teachers could in principle be used to identify specific students and eliminate this double-counting issue, but those files are only available beginning in 2006 and would thus result in a much smaller sample of teachers.

[10] The classroom assignment data contain separate indicators for the first year a teacher has ever taught, her first year teaching in North Carolina, and her first year back after an absence. My sample drops all teachers who ever teach in another state but includes teachers with absences in their careers.

results from regressions controlling for years of experience, I report results for both definitions. I discuss the similarities and differences between the two measures themselves in the Appendix.

2.2 Matching Teachers and Classrooms to Students

In order to produce a working data set, I have followed the example set by Ladd and Sorensen (2017). The NCERDC is straightforward about the fact that the teachers which are associated with each student's test score are not necessarily the teacher of that student in the relevant subject. In fact, the teacher listed in that file is only the proctor of the state exam for that student, and may have no classroom relationship with the student. For example, according to the classroom assignment data, nearly 50% of the teachers designated as proctors for 8^{th} grade math exams with a year or more of experience have never been observed teaching a math course in my data.[11] Given the nature of the study, incorrectly assigning these teachers to students to this extent may amount to large biases on my estimates. Further, while previous researchers using the NCERDC data have argued that these mismatches are rare for students in earlier grades (Jackson and Bruegmann, 2009), Isenberg et al. (2015) show evidence from from Washington D.C. that indicates that significant numbers of teachers are improperly matched to students in administrative data even at the elementary school level.

With this in mind, I ignore all student-teacher matches in the administrative data and link students to the relevant subject teachers by the following simple procedure. Since 2006, the NCERDC has provided "Course Membership" files, which link every student to every course in which he or she was enrolled every year. These files also indicate the teacher of each course, and every course is given a numeric code which corresponds to the subject of the course (e.g. "Math", "Algebra"). In order to assign

[11] This is true for both my main sample of teachers as well as those for whom I have only incomplete teaching histories.

students to the teacher of the course most related to the state test scores, I use the following procedure: if a student took the most popular math course in her grade-year, that student is assigned the teacher for that course. If a student did not take the most popular math class but did take the second most popular, then they are assigned their teacher for that course, and so on. I drop from my sample the small fraction of students whom are left unmatched by this procedure.[12]

2.3 Heterogeneity of Experience

Figures 1 and 2 document the extensive variation in the number of classes and students taught, both in the cross section and over a teacher's career, and for middle and elementary school teachers separately.[13][14] The median teacher teaches six classes and 126 students per year, but the interquartile ranges show that some teachers have taught three times as many classes and students as others with the same years of experience.[15] On the low end, 5% of teachers with 10 years of experience have only taught 200 students. Figures 3 and 4 show similar variation in the numbers of black and white students teachers are assigned.

It has been well documented that school and teacher quality are largely unobservable (Hanushek, 1997). To the extent that (i) higher quality teachers are assigned more students than their lower-quality peers or (ii) unobservable teacher and school

[12] I perform this procedure at the level of state course code after restricting the sample to course codes beginning with "2001", "2023", "2020," and "2003" for my math score sample, and "1010,", "1001," and "1038" for reading. These cover the vast majority of math and reading courses in elementary and middle school. I include "Self-Contained" courses in both samples. See the Appendix for evidence that my constructed sample is similar to that in Ladd and Sorensen (2017).

[13] I will use "taught" and "assigned" as synonyms when describing data from the classroom assignment files. Except for an indicator for whether a class is a "teaching assignment," there is no way to know how frequently a teacher participated in actual classroom instruction.

[14] Note here that whenever I refer to regressions and tables including only "middle school teachers" or "elementary school teachers" I am restricting the sample to include only teachers who have only *ever* taught middle (grades 6-8) or elementary school (grades 3-5), respectively. Switching between the two is rare, but this exclusion aids interpretability.

[15] Some teachers have taught more than 8 classes per year on average. Hours in the school day are limited, so teachers in the far right tail of this distribution must either be in schools with shorter than average classes or represent some level of measurement error in the assignment data.

quality are correlated with school demographics, estimation of the effects of interest using all of the variation in figures 1-4 will be biased. In figure 5 I show the empirical densities of the proportions of white, black, and male students for teachers with one, five, and ten years of experience, which suggests significant non-random assignment of students of different races across teachers. Although the distribution of the fraction of male students is less apparently skewed, we may still be concerned that the number of male and female students that teacher is assigned is in part determined by her teaching abilities.

In order to isolate more plausibly exogenous variation, I regress the proportions in these figures on school and teacher fixed effects, and plot the densities of the residuals from those regressions in figure 6. Relative to the previous figures, these residualized demographic share distributions have much smaller variances, meaning a good deal of the raw variance of each share is explained by persistent teacher and school characteristics. Still, each of these plots demonstrate that there is still variation to inform the estimation procedure, even after controlling for both teacher and school fixed effects as well as years of experience.

To make clear what variation will be used in estimation, in table 3 I show the coefficient estimates and R^2 of regressions of male-experience on female-experience to demonstrate that the returns to teaching male and female students can be separately estimated in my sample. In the first column I show that when pooling all within- and across-teacher variation (i.e. including no other controls), male- and female-experience are essentially perfectly correlated. The remaining columns control for teacher and school fixed effects as well as years of experience and restrict to subsamples of teachers with less than two, five, and ten years of experience. The correlation between male-experience and female-experience is significantly lower after removing teacher and school fixed effects, which highlights the importance of including these as controls in all of my regressions; idiosyncratic variation in gender-specific experience, averaged out at the population level, plays a significant role within teacher, particularly for younger teachers. Almost no covariation between these demographic experience variables is added by teachers with more than five years of experience, as the inclusion of teachers with up to five years of experience (column 3) brings the R^2 of this regression up to 0.95. Thus, though I will include all teachers in my regressions, the majority of the identifying variation for my parameters of interest will come from teachers with fewer than five years of experience. This is also true for race-related regressions.

3 Estimation

I estimate a standard value-added education production function of the following form

$$y_{it} = f(y_{it-1}) + \beta^X X_i + \beta^T T_{jt} + \beta^C C_{ict} + \gamma E_{(i)jt} + \theta_j + \delta_k + \phi_{gt} + \epsilon_{ijt} \tag{1}$$

where y_{it} is the (math or reading) standardized test score of student i in year t and $f(\cdot)$ is a cubic function of lagged test score. The vectors X_i and T_{jt} are student and teacher characteristics including indicators for whether the teacher has an advanced degree and whether the student and teacher are the same race, the same gender, or both. I allow this match effect to differ for every combination of gender and the four racial categories described above. I also include a dummy variable indicating whether the student is eligible for free or reduced-price lunch. In C_{ict} I include indicators for groups of different class sizes, the fraction of classroom peers who are white, black, economically disadvantaged, and classroom peers' average lagged math and reading scores. To emphasize their role as the treatment of interest, I include all classroom assignment-related experience variables in $E_{(i)jt}$ in terms of the total number of students taught by a teacher before year t and/or a vector which decomposes this count into gender- or race-specific experience terms, sometimes interacted with the race or gender of student i (hence the potential dependence of this vector on i). To permit diminishing returns to experience, these variables enter either in log form or a cubic function in all regressions.[16]

In every regression including counts of students taught I also control for the number of classes taught by a teacher before the current year to differentiate between the effects of teaching more classes and teaching larger classes.[17][18] As for unobservables, θ_j , δ_k, and ϕ_{gt} are teacher, school, and grade-year fixed effects respectively, ϵ_{it} is an idiosyncratic error which I cluster at the teacher-year level.[19] I use the REGHDFE command in Stata (Correia, 2016) to estimate both the OLS and IV (discussed below) versions of the model. In this model, teachers and schools may differ persistently in unobservable ways, and these differences can be correlated with classroom assignment. Perhaps most importantly, some teachers may consistently be assigned larger classes than others. The variation identifying the parameters of interest γ come from changes to a given teacher's experience (measured in numbers of classes or students) over the course of her career, within a given school, after controlling for the size of each cohort of students each year.

The key identifying assumption in regressions like equation 1 is that there are no time-varying changes in a teacher's effectiveness which are correlated with the number of classes and students she is assigned. In particular, this excludes the possibility that schools observe growth in teacher effectiveness which is not captured by observables and use that growth to match teachers to students. Because this need only hold conditional on observables, this assumption is significantly weaker in regressions which include years of experience as a control. At the median there are one or fewer math

[16]I include a complete list of controls in the Appendix.

[17]Just as in my counting of teachers' history of students, I weight the number of classes taught by the length of the course. Full year classes receive a weight of one, and half-year classes receive a weight of 0.5. All other course lengths are omitted

[18]The unconditional correlation between the total number of students taught and the number of classes taught is approximately 0.91.

[19]In the Appendix I repeat the regressions for my main results without school fixed effects. My results change very little.

teachers with a given level of experience teaching each grade in a given school each year in my sample, meaning that the potential for administrators to sort teachers conditional on their years of experience is limited. As the following sections will show, my main results change little upon the addition of these controls.

4 Results: Students and Classrooms

4.1 Number of Classes and Average Class Size

Two mechanisms are consistent with the hypothesis that teachers with more interactions with students will be more effective, all else equal. If teachers benefit most from additional opportunities to practice delivering their lesson plans, then we should find that assigning a teacher an additional average-sized class of students should improve her future effectiveness. It may also be the case that, even conditional on the number of classes previously taught, teachers who have been assigned *larger* classrooms have had more opportunities to discuss the material with more students.

In figure 7(a) I present descriptive evidence of the main result of the paper: students tend to perform better on the state-wide standardized math exam when they are assigned teachers who have taught large classes in the past. In figures 7(b) and 7(c), which restrict the sample to teachers with ten years of experience, I show that this is not driven solely by some teachers having taught more students or classes than their peers; even when comparing only teachers with ten years of experience, these variables are each only weakly correlated with student achievement. Rather, it is the combination of the two (in average class size) which are most strongly correlated with achievement. To the best of my knowledge, the relationship in figure 7(a) has not been demonstrated in literature before, even as a non-causal relationship. Still, as stark as this correlation is, there are many potential explanations for this trend. If better (or more experienced) teachers, or better schools, have larger classes on average we would see this relationship even in the absence of a causal relationship between class size and teacher effectiveness. Further, if teachers who are assigned many classes tend to teach smaller classes, this would induce the negative correlation shown in figure 7(c). To address these potential issues jointly and isolate plausibly causal variation, I now move to the regression framework shown in equation 1.

In table 4 I show the results of six regressions of scaled math scores on the logs of teachers' average class size and total number of classes previously taught, as well as all controls discussed previously. In all columns, I include teacher, school, and grade-year fixed effects. Columns 2 and 3 estimate this regressions on the subsets of elementary and middle school students, respectively. Column 4 also includes student fixed effects. In column 5 I control for years of experience as constructed using the classroom assignment data and in column 6 I instead control for experience as defined by pay codes. In each column, the estimated effect of teaching additional classes is statistically significant at all standard levels, meaning that teachers who have taught more classes tend to be more effective than those who have taught fewer classes. Further, in columns 1, 5 and 6, I find that teachers who have taught larger classes are also on average more effective. Because the last two columns of the table control for years of experience, I take these results as suggestive evidence that both average class size and number of classes are important predictors of a teacher's future effectiveness.

Though an informative baseline, these regressions have a few drawbacks for our purposes. First, the log specification enforces a particular form of diminishing returns, which may not hold in practice. Second, this particular specification obscures my hypothesis that the number of students a teacher has been assigned represent a stock of experience above and beyond her years of teaching. It also reduces the available variation to explore this effect, because variation in average class size necessarily decreases with teacher tenure.[20] Finally, although it is important for our understanding

[20]This is because teachers' average class size in the later years of their careers approach their career average, by construction. Because I include teacher fixed effects in all regressions, only deviations from

of teachers' human capital accumulation that teachers improve as they teach more courses, it is unclear what policy levers available to a school could make use of this finding. The number of classes in a day is limited by the length of the school day, meaning assigning teachers additional classes may be infeasible. On the other hand, classes can be combined and rearranged, or teachers laid off, to assign some teachers larger classes. Therefore, it will be easier to operationalize these results in terms of total numbers of students, which is the focus of the following section.

4.2 Number of Students

In this section I move to treating the number of students taught in previous years as the variable of interest, to address the preceding concerns. I estimate all remaining results in terms of some function (logs, or a cubic in levels) of the total number of students a teacher has taught conditional on the number of classes she has taught. To see how this relates to the previous section, note a regression like those estimated in the preceding section

$$y_{ijt} = \beta log(AvgClassSize)_{jt} + \gamma log(\#Classes)_{jt} + \epsilon_{ijt} \tag{2}$$

is equivalent to the alternative regression

$$y_{ijt} = \beta log(\#Students)_{jt} + (\gamma - \beta) log(\#Classes)_{jt} + \epsilon_{ijt} \tag{3}$$

Thus, although including number of students instead of average class size changes the interpretation of the coefficient on $log(\#Classes)$[21], it does not change the interpretation of β as the effect of increasing the average class size. Therefore, although I estimate some results in levels and some in logs (meaning this equivalence is not always exact),

this average are used to identify the effects of interest.

[21]The difference is that γ in equation 2 represents the effect of assigning an additional course to a teacher, while $\gamma - \beta$ represents the division of a teacher's current students into more classes.

any estimated effects of total counts of students are identified by variation in average class size across teachers and over time.

I present the results of seven regressions in table 5, similar to table 4, where the dependent variable is students' math test score and the independent variable is a cubic function in terms of *hundreds* of students taught before the given year. For reference, recall that the median teacher in my sample teaches approximately 120 students per year, meaning that the reported (linear) coefficients can be loosely interpreted as the effect of increasing the number of students a teacher has taught by the median teacher's workload for one year. All regressions in table 5 include teacher, school, and grade-year fixed effects. In columns 1-4, I estimate that assigning a teacher 100 more students in one year will improve her average future student's test score by approximately 0.02 standard deviations (of state-wide math scores). I also find that having taught a given number of students in more separate classes (the last row of coefficients) negatively impacts teacher effectiveness, which indicates that the mechanism for the effect of teaching more students is indeed through larger classes.

Perhaps surprisingly (though this was also the case in table 4), estimated effect sizes for elementary teachers (column 2) are quite similar to those for middle school teachers (column 3). Given that teachers in elementary school teach far fewer students per year than middle school teachers (see figure 2), these estimates imply that average growth rates for middle school teachers are higher than those in elementary schools. This is not an argument that all teachers should spend their early careers in middle schools. Ost (2014) provides evidence that experience is at least in part grade-specific, so experience gained in middle school may be heavily discounted when applied to elementary. I do however take this as evidence that pooling elementary and middle school teachers and students is reasonable in the rest of table 5.

In columns 5 and 6 I add years of experience as an additional control, first as defined by the classroom assignment files and second as defined by pay codes. The significance and magnitude of my estimates drop significantly in these columns, though the linear effect is still statistically significant at the 5% and 0.1% levels, respectively. In order to demonstrate the size of these estimates, I plot teacher effectiveness, as implied by columns 5 and 6 of table 5, as a function of students previously taught in figures 8(a) and 8(b) respectively. These figures clearly demonstrate that, even conditional on years of experience and the number of classes taught, teachers who have taught more students in the past tend to be significantly more effective than their peers.

To address remaining endogeneity concerns, in column 7 I present the results of an IV regression which uses the previous year total enrollment in a given school-grade-year as an instrument for each teacher's total count of previous students taught. The reasoning here is that an increase in school size will tend to increase the number of students each teacher is assigned in a given year. Because the instrument is constant for teachers within a school-grade-year, identification of this parameter does not rely on exogenous sorting of students to teachers within a grade. Rather, the identifying assumption for these estimates is the exogeneity of student body growth with respect to unobservable teacher quality at the grade level. Practically, that requires that schools do not hire better than average teachers in response to enrollment growth, and that student enrollment does not respond within a year to school- or school-grade-level increases in teacher quality.

The coefficients estimated from this regression are much larger than those in other columns, implying that the first 100 assigned students increase teacher effectiveness by more than 0.04 standard deviations. Though each coefficient is insignificant alone, the full cubic function is jointly significant at the 5% level.[22] The size of this effect, which is estimated conditional on teachers' years of experience, is larger than the returns to years 11-20 of teaching experience as estimated by Ladd and Sorensen (2017). To highlight the cross-sectional variance of effectiveness implied by the estimates in table

[22]Further, if I estimate a single linear term, that coefficient is significant at the 5% level.

5, I calculate expected effectiveness for each teacher as a function of students taught as implied by columns 1,5 and 6 of table 5. Figure 9 plots the densities of these predicted values for teachers with one, five, and ten years of experience in 2011, all of which indicate substantial heterogeneity of effectiveness within years of experience.

In table 6 I present the same results for reading scores, and find a much weaker pattern of statistical and economic significance. Just as in the previous table, the top line coefficient remains approximately constant for columns 1-4, drops when I control for years of experience, and increases significantly in the IV specification. However, the magnitude of the OLS estimates are at most half of the effect estimates for math scores, and only a few coefficients are significant at more than the 5% level. This is consistent with recent studies of years of experience, which find that the effects of experience are much larger on math scores than on reading scores (Ladd and Sorensen, 2017; Wiswall, 2013; Papay and Kraft, 2015). Because this pattern of smaller and less significant results for reading persist through the rest of my estimates, in the following sections I focus on math scores only. Additional results for reading scores are included in the Appendix.

5 Results: Demographic Heterogeneity

In this section, I add new dimensions of heterogeneity to experience, allowing teacher effectiveness to depend not only on the number of students she has taught previously but also on their demographics. In the spirit of the literature on demographic matching, I also interact these experience terms with student and teacher demographics. Because the effects studied in this section are all conditional on total counts of students, the identifying assumptions are weaker here because they only require that the demographic proportions assigned to teachers are exogenous. Still, I estimate all effects conditional on teacher, school, and grade-year fixed effects. As in the previous section, standard errors are clustered at the teacher-year level.

In four of the seven sets of estimates presented in table 5, quadratic and cubic terms in the number of students are significant at the 1% level, making clear that choosing a specification which permits diminishing returns to teaching additional students is essential. On the other hand, in this section I estimate race- and gender- specific returns, which are then interacted with indicators for teacher or student demographics. This presents two challenges. First, this greatly increases the number of coefficients to estimate, which reduces the precision of the individual estimates. Second, tables of these results are harder to interpret, as each effect of interest becomes a function of multiple parameters. To increase precision and improve interpretability, in this section I substitute the cubic functions used in the previous section for a log specification. In the Appendix I include estimates from quadratic specifications corresponding to each of the results of this section as well as a discussion of the differences between the estimates.

5.1 Heterogeneity By Race

Teachers who have seen the same number of students over the course of their careers, but who have interacted with different racial distributions of students, may be differentially effective in the current classroom. This effect might be direct, e.g. through average racial biases and preferences, or indirect through the correlation between race and classroom instructional needs. The results in this section modify my definition of experience to instead be in terms of the total number of students in each of four race-based categories (now entering the regression in logs). If student race affects teacher growth, then the estimates of this model should indicate that the returns to teaching students of one race differ from those from teaching another.

I present the results from these regressions in table 7, all of which include school, teacher, and grade-year fixed effects. In general, the estimates from this regression confirm the magnitude of the estimates from the previous section, though there does appear to be some level of heterogeneity. For example, the estimated effect of teaching more white students is notably smaller than the same for black and other race students in most regressions. Taking column 1 as an example, these estimates imply that assigning the average teacher 1% more black students improves her students test scores by twice as much as if she were assigned 1% more white students instead. Because 85% of the teachers in my sample are white, this may indicate that teachers improve faster by teaching demographically dissimilar students. I investigate this potential mechanism in the next section.

For a more precise measure of heterogeneity, at the bottom of the table in the row titled "Homogeneity" I include the p-value corresponding to the null hypothesis that all of the four race-specific coefficients are equal, which amounts to a test of the hypothesis that the returns to teaching additional students is independent of the race of those students for the average teacher. Although I have no evidence against this hypothesis in the first three columns, in regressions including student fixed effects or years of experience I reject the null at the 10% level and nearly reject at the 5% level. Altogether, these results indicate that on average, teachers who have taught the same number, but different demographics, of students in the past are on average differentially effective in the current year.

5.1.1 Interactions

Next I estimate a fully interacted model, in which I interact the log of each race-specific total with dummy variables for the same four teacher and student race categories (white students and teachers are the omitted group). Where the previous section permitted the average teacher to experience growth which varies according to the races of her students, the specification I estimate in this section permits two additional dimensions of heterogeneity in this effect. First, because I interact the race of student i

with each effect, I permit that a teacher's prior experience teaching students of a given race may be more effective for students of some races than for others. This would be the case if, for example, black students benefit more than white students from teachers' prior experience with black students. Second, because I also interact all effects with teacher race, I allow teachers of different races to improve at different rates from students of a given race. If, for example, white teachers tend to learn slower from black students than black teachers do, I should estimate that these interactions with teacher race are statistically significant.

Although the log specification reduces the number of parameters to estimate significantly, the results of this model are still not conducive to a concise presentation. Instead, I offer a subset of regression coefficients in table 8 and include the full table at the end of the Appendix. The first takeaway from this table is that the baseline effects (e.g. the coefficient on $log(\#White)$) are similar to those in table 7, which is to be expected given that the modal teacher and student are white (and white is the excluded group for both sets of interactions). The second important feature of this table is that most of my estimates of interaction effects with indicators for black students and teachers, which are the largest minority group, are insignificant and are not a consistent sign. Interactions with other racial categories are similarly uninformative.

Because these individual interactions are estimated imprecisely, I focus instead on testing hypotheses that groups of coefficients are zero. In tables 9 and 10 I present p-values for two sets of tests. For each regression I estimate in table 8, for each racial category, I show the p-value of the test that interactions terms between the number of students of that race a teacher has previously taught with any (current) students' race variables are zero in table 9. In table 10 I present p-values from the analogous tests related to race-specific student experience totals interacted with teacher race. For example, the top left value in table 9 represents the p-value of the test that all interactions of a teacher's experience with white students with her current students' race are zero. The bottom row of the same table presents p-values for the joint test of the four hypotheses above it. Table 10 is organized analogously.

To be clear, these tables are testing whether the returns to *having previously taught* students of a particular race are equal for all students and teachers in a *current* classroom. That is, we wish to test the null hypotheses that (1) all students benefit equally from a teacher's experience with students in each racial category and (2) the rate at with a teacher improves as she is assigned students of one race is independent of her own race. Table 9 provides evidence against the former hypothesis. Although the patterns of heterogeneity vary by regression, in all but one regression I can reject the hypothesis that all interactions with a current student's race are zero (the final row), with $p < 0.001$. In table 10 I find a similar pattern; the significance of heterogeneity with respect to any given race varies by regression, but I can reject the null hypothesis of homogeneity with respect to teacher race at the 5% level in all but two specifications. Thus, although I cannot precisely determine the source of this heterogeneity, or the direction of these heterogeneous effects, these tables indicate that both the rate at which teachers improve and the returns to their experience in their current classrooms depend on their own races, those of the students they have taught in the past, and those of the students they are currently teaching.

I also produce figure 10, much like figure 9, which shows kernel density estimates of the total effect of a teacher's experience on her students in the classroom, produced using estimates from columns 1,5, and 6 of table 8.[23] As in figure 9, figure 10 indicates considerable variation in teacher effectiveness in each set of estimates, including after conditioning on years of experience, thought this variation is notable smaller than in the preceding figure (this is in part due to the diminishing returns enforced by the log specification). The most notable feature of this interacted model which shows up in

[23]Note that these densities are taken over all students, and that in this regression one teacher's experience may have different effects on two different students in the same classroom. This is in contrast to my main estimate shown in 9, which assumes homogeneous returns for each student within a teacher-year.

figure 10 is that, because some estimates of interaction terms are negative, the total effect of a teacher's experience on a student's test score may be negative. This would be the case if, for example, teachers who teach one race significantly more than others tend to become worse at teaching students of other races. Although I permit cases like this, I do not find significant evidence of biases of this magnitude.[24]

5.2 Heterogeneity by Gender

In table 11, I present my results from a set of regressions nearly identical to those in the previous subsection (in terms of the number of fixed effects and control variables included), differing only in that experience is modeled as the number of male and female students seen by a teacher over the course of her career. As table 11 shows, the number of male and female students previously taught are both significant predictors of teacher effectiveness. Similar to the pattern in table 8, estimates of individual interactions between gender-specific experience and current student gender are insignificant. Most of these coefficients are also two orders of magnitude smaller than the baseline effects. What is new in this table is that the interactions of gender-specific experience with *teacher* gender are consistently negative and large relative to the baseline effects, both for male- and female-experience. In each column, except when I restrict the regression to middle school teachers only, my estimates imply that male teachers improve significantly slower when they are assigned female students than when they are given male students (and symmetrically for female teachers).

To understand the magnitude of this effect, note that my smallest significant estimate of the coefficient on $(\#Female) * TeacherMale$, i.e. the difference in the returns to experience with female students between male and female teachers, is negative and nearly as large as the *largest* estimates of these returns for female teachers. This im-

[24]Although figure 10(b) appears to show a mass of negative effects, this arises largely due to the imprecision of estimates relating to Hispanic teachers, who make up less than one percent of the teachers in my sample. See column 5 of table A.18 for the full set of estimates used to construct figure 10(b).

plies that male teachers improve little to none when teaching female students, and the rest of the table implies the symmetric result for female teachers teaching male students. Given that these interactions remain large even conditional on years of experience, these coefficients clearly indicate that teachers improve substantially slower when teaching students of a different gender. I show in the Appendix that neither the general magnitude of these interactions nor their joint statistical significance are dependent on the log specification used here.

6 Policy Implications

The main result of this paper implies a tradeoff for policy-makers. Researchers consistently find that larger class sizes are detrimental to student achievement. If teachers benefit from teaching larger classes, then schools are presented with a dynamic problem: big classes mean bad test grades for current students, but good test grades for future students. A back-of-the-envelope calculation can inform the extent of the tradeoff. The median 8^{th} grade classroom contains 23 students, and the median middle school in my sample has four potential 8^{th} grade math teachers. Suppose that a hypothetical school initially employed four new teachers, each teaching a single median-sized classroom. Then, counterfactually, suppose that we remove two teachers from that school and sort students randomly into the remaining classrooms. In other words, let us double both the size of the average class and the number of students assigned to each teacher.[25]

According to my estimates from the regression in column 5 of table 5 (treating the estimated effects of current class size as causal), doubling class sizes in this way would reduce each student's test score by nearly 0.02 standard deviations on average[26]. On the

[25]One could alternatively consider increasing the number of separate classes taught by each teacher, but I ignore that potential momentarily (i) because it is unclear that my results hold when the length of class meeting times change, and (ii) to emphasize the short-run costs in cases where that is infeasible.

[26]I include five bins of class size in all regressions, and treat 21-30 students as the excluded bin. I estimate that classes with less than 6, 6-10, and 11-20 students improve students' test scores by 0.32, 0.29, and 0.18 standard deviations relative to the excluded bin, respectively. Classes with 31-40 and

other hand, it would improve each remaining teacher's future effectiveness by less than 0.001 standard deviations. This highlights one of the difficulties of putting the results of this paper to use: in the short run, the negative effects of larger classrooms may far outweigh the positive effects of better teachers. Note however that this is not true for all potential policies. Suppose instead we fired all but one teacher and assigned her the four pre-existing classrooms. In this case, all classrooms remain the same size, and the remaining teacher's future students will benefit from a 0.003 standard deviation increase in her effectiveness. As shown in figure 1, there appear to be some instructors who are assigned far fewer classes than average, so there may be room for more significant increases to some teachers' workloads.[27]

To see how these effects accumulate over time, consider two teachers in the data who have 15 years of experience. Suppose that the first has taught 70 students per year on average in her career, and the other has taught 140.[28] I use these numbers because they imply these two teachers have taught more students than 25% and 75% (respectively) of teachers with the same years of experience. My estimates in column 5 of table 5 imply that the more experienced of these teachers improves her students' test scores by 0.1 standard deviations more than her less experienced peer. This difference, which is between two teachers with the same years of experience and is the smallest of my estimates, is bigger than some estimates (Clotfelter et al., 2007) of the returns to more than 20 years of teaching. Therefore, although on the margin these effects appear small, over time they accumulate to imply substantial differences across teachers with the same years of experience.

My results indicating that this baseline effect depends broadly on the demographics

---

more than 40 students decrease scores by 0.005 and 0.18 standard deviations. Relative to to Ladd and Sorensen (2017), the most relevant comparison, my estimates imply larger benefits to small classes and slightly smaller losses from large classes.

[27]An important caveat here is that the effort costs of teaching additional classes, beyond current teaching assignments, may be large enough to offset some of the gains from additional experience. Without estimates of these costs of effort, I cannot address this directly here.

[28]This exercise assumes these teachers taught the same number of classes for simplicity.

of the students a teacher is assigned, as well as on her own demographics, are more complicated to apply to policy. My estimates in the preceding section indicate that gender matches between student and teacher increase the rate at which teachers learn. Thus, a simplistic and incomplete reading of the results in this paper could conclude that teachers and students should be segregated to maximize the number of demographic matches. Such an interpretation is deeply flawed and beyond the claims of the paper. Although I attribute *causal* interpretations to my estimates, in the sense that the addition of a student of a particular demographic causes a teacher to improve, I make no claim, and highly doubt, that demographics are the relevant underlying *structural* component explaining these differences. Differences across demographics may be explained by any number of aspects of family inputs, backgrounds, personalities, and classroom needs which differ on average along lines of race and gender. Teaching methods most frequently taught at universities may be better suited to students from some backgrounds than others, or the implementation of those methods may differ along some dimension related to race or gender. Regardless of the source of these differences, they motivate further study of the underlying structural reasons for teacher-student matches.

7 Concluding Remarks

In total, the results herein indicate that the process by which teachers accumulate human capital through experience is far more heterogeneous than has been previously documented. The most consistent finding in this paper is that, all else equal, the most effective teachers are those who have taught the most students, and I find this result even when I include teacher and school fixed effects and control for teachers' years of experience. Some teachers have taught far more students per year than others, and this variance (figures 2, 3, and 4) implies substantial heterogeneity in teacher effectiveness which cannot be captured by existing measures, such as years of teaching experience. I also show that these effects present a dynamic problem for policymakers, forcing a choice between small classes, to maximize current student achievement, and large classes, to maximize future teacher effectiveness.

There are a few broad patterns in the set of demographic-specific results. When experience is gender-specific, I find that teachers improve slower when assigned students of a different gender than when assigned students of the same gender. On the other hand, although in some regressions I can reject the null hypothesis that the returns to teaching students of different races are the same on average, the mechanisms for this effect are unclear. Bias, held by teachers against some demographic groups of students, could produce the results in table 10 (and those with respect to gender). If students of different backgrounds require different teaching methods, this mechanism could produce the results in table 9. Further, because I cannot reasonably estimate a fully interacted model (i.e. interacting (# White) with teacher *and* student race), either of these mechanisms may explain the statistical significance of the other through omitted variable bias.

Though this study does not have sufficient power to confidently determine the mechanisms for these differences, better data will soon exist. The same data used to match teachers to students and their test scores continues to be produced by the NCERDC each year, and can soon be used to differentiate teachers' experience on things like the socioeconomic status and lagged achievement of their students on a panel of similar length to this study. Unlike the classroom assignment data used here, this approach uniquely identifies students (avoiding any over-counting), and these additional student characteristics may be more informative than the coarse demographic bins I am able to use here.

References

Aaronson, Daniel, Lisa Barrow, and William Sander, "Teachers and student achievement in the Chicago public high schools," *Journal of labor Economics*, 2007, *25* (1), 95–135.

Angrist, Joshua D and Victor Lavy, "Using Maimonides' rule to estimate the effect of class size on scholastic achievement," *The Quarterly Journal of Economics*, 1999, *114* (2), 533–575.

Aucejo, Esteban M, Patrick Coate, Jane Fruehwirth, Sean Kelly, and Zachary Mozenter, "Teacher effectiveness and classroom composition," 2018.

Chetty, Raj, John Friedman, and Jonah Rockoff, "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, 2014, *104* (9), 2593–2679.

_ , _ , and _ , "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, 2014, *104* (9), 2633–2679.

Clotfelter, Charles T, Helen F Ladd, and Jacob L Vigdor, "Teacher credentials and student achievement: Longitudinal analysis with student fixed effects," *Economics of Education Review*, 2007, *26* (6), 673–682.

Correia, Sergio, "Linear Models with High-Dimensional Fixed Effects: An Efficient and Feasible Estimator," Technical Report 2016. Working Paper.

Dee, T.S., "Teachers, Race, and Student Achievement in a Randomized Experiment," *Review of Economics and Statistics*, 2004, *86* (1), 195–210.

Ehrenberg, Ronald, Daniel Goldhaber, and Dominic Brewer, "Do Teachers Race, Gender, and Ethnicity Matter? Evidence from the National Education Longitudinal Study of 1988," *Industrial and Labor Relations Review*, 1995, *48* (3), 547–561.

Gershenson, Seth, Stephen B Holt, and Nicholas W Papageorge, "Who believes in me? The effect of student–teacher demographic match on teacher expectations," *Economics of Education Review*, 2016, *52*, 209–224.

Grönqvist, Erik and Jonas Vlachos, "One size fits all? The effects of teachers' cognitive and social abilities on student achievement," *Labour Economics*, 2016, *42*, 138–150.

Hanushek, Eric A, "Assessing the effects of school resources on student performance: An update," *Educational evaluation and policy analysis*, 1997, *19* (2), 141–164.

Hoxby, Caroline, "Peer effects in the classroom: Learning from gender and race variation," 2000.

Isenberg, Eric, Bing ru Teh, and Elias Walsh, "Elementary school data issues for value-added models: Implications for research," *Journal of Research on Educational Effectiveness*, 2015, *8* (1), 120–129.

Jackson, C Kirabo and Elias Bruegmann, "Teaching students and teaching each other: The importance of peer learning for teachers," *American Economic Journal: Applied Economics*, 2009, *1* (4), 85–108.

Krueger, Alan B, "Economic considerations and class size," *The Economic Journal*, 2003, *113* (485), F34–F63.

Ladd, Helen F and Lucy C Sorensen, "Returns to teacher experience: Student achievement and motivation in middle school," *Education Finance and Policy*, 2017, *12* (2), 241–279.

Lavy, Victor, "What makes an effective teacher? Quasi-experimental evidence," *CESifo Economic Studies*, 2015, *62* (1), 88–125.

Murphy, Richard and Felix Weinhardt, "Top of the class: The importance of ordinal rank," 2014.

Ost, Ben, "How do teachers improve? The relative importance of specific and general human capital," *American Economic Journal: Applied Economics*, 2014, *6* (2), 127–51.

Papageorge, Nicholas W, Seth Gershenson, and Kyungmin Kang, "Teacher Expectations Matter," *IZA Discussion Paper No. 10165*, 2016.

Papay, John P and Matthew A Kraft, "Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement," *Journal of Public Economics*, 2015, *130*, 105–119.

Rivkin, Steven G, Eric A Hanushek, and John F Kain, "Teachers, schools, and academic achievement," *Econometrica*, 2005, *73* (2), 417–458.

Wiswall, Matthew, "The Dynamics of Teacher Quality," *Journal of Public Economics*, 2013, *100* (1), 61–78.

Table 1: Student Sample (Math)

	(1) All Students	(2) Math Sample	(3) t-test
Male	.502	.501	.305
White	.411	.394	.000
Black	.193	.201	.000
Disadvantaged	.534	.553	.000
Disabled	.107	.107	.061
Lag Math (5th Grade)	.008	-.027	.000
Lag Math (8th Grade)	.014	-.041	.000
Lag Reading (5th Grade)	.008	-.031	.000
Lag Reading (8th Grade)	.012	-.032	.000
Advanced Math	.152	.137	.000
Advanced Reading	.136	.121	.000
N	437595	211918	

t statistics in parentheses

Note: Comparison on observables between all students in 2011 matched to a math class in my sample ("All Students") and the students actually used in estimation, after making the sample restrictions described in the paper ("Math Sample"). The third column presents the t-test that the first two columns are equal.

Table 2: Teacher Sample (Math)

	(1) All Teachers	(2) Math Sample	(3) t-test
Experience (Pay Code)	11.971	7.719	.000
Experience (Classroom Count)		6.283	.000
Licensed	.994	.993	.269
Lateral Entry	.012	.018	.000
Advanced Degree	.343	.284	.000
Male	.139	.148	.004
White	.849	.856	.055
Black	.129	.123	.049
N	11497	5558	

t statistics in parentheses

Note: Comparison on observables between all teachers in 2011 matched to a math class in my sample ("All Teachers") and the teachers actually used in estimation, after making the sample restrictions described in the paper ("Math Sample"). The third column presents the t-test that the first two columns are equal.

Table 3: Identifying Variation: Gender-Experience

	(1)	(2) < 2	(3) < 5	(4) < 10	(5) All
# Female	1.001***	0.659***	0.638***	0.663***	0.705***
	(0.001)	(0.014)	(0.007)	(0.005)	(0.003)
N	63860	6276	20270	35586	55210
R^2	0.964	0.856	0.947	0.985	0.996

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Column 1 presents the correlation between experience teaching male and female students. Columns 2-5 present the coefficient of a regression of the number of male students previously taught on the number of female students, treating teacher-year as the level of observation. Each regression controls for teacher and school fixed effects as well as years of experience. Columns 2-4 restrict the sample of teachers to those with less than 2, 5, and 10 years of experience, respectively.

Table 4: Class Size and Number of Classes (Math)

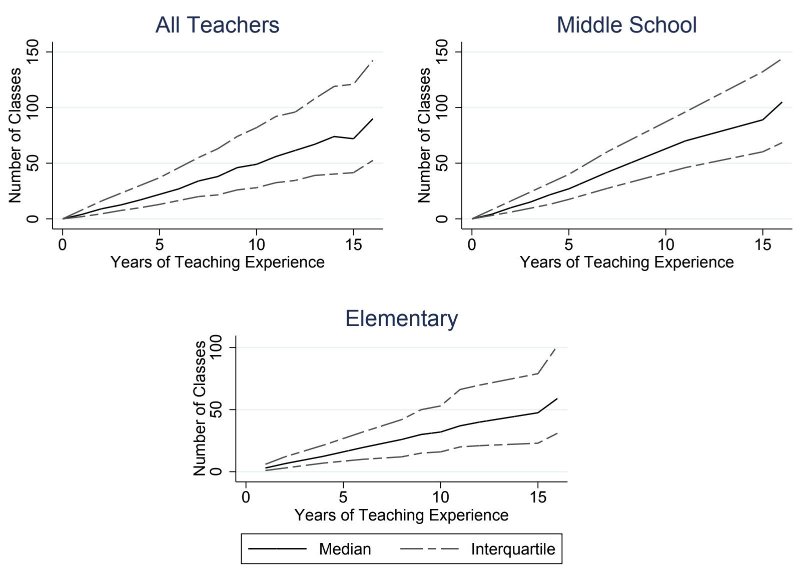
	(1)	(2) Elem	(3) Middle	(4)	(5)	(6)
log(AvgClassSize)	.0377*	.0213	.0253	.0122	.0328*	.051**
	(.0157)	(.0226)	(.0246)	(.0179)	(.0157)	(.0184)
log(#Classes)	.03***	.0242***	.042***	.0266***	.0125*	.0192***
	(.0036)	(.0045)	(.0072)	(.0049)	(.0051)	(.0049)
Student FE	No	No	No	Yes	No	No
Years Teaching	No	No	No	No	Class	Admin
R^2	.743	.74	.748	.916	.743	.743
N	856528	308998	434679	485244	856528	777435

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Estimated returns to the total number of students and classes previously taught by a teacher in any class with fewer than 35 students. The dependent variable is the state-wide math exam (end-of-grade) score, normalized to have mean zero and standard deviation one. Columns 1,2, and 3 contain my main estimates for all, elementary, and middle school teachers and students. Column 4 adds student fixed effects to the regression. Columns 5 and 6 control for years of experience as defined by the classroom assignment and pay code data, respectively. All regressions include teacher, school, and grade-year fixed effects. Standard errors are clustered at the teacher-year level.

Figure 1: Number of Classes



Note: Median and interquartile range of the total number of students taught, calculated for teachers in 2011 and shown separately for elementary, middle school, and all teachers in my sample.

Table 5: Homogeneous Coefficient (Math)

	(1)	(2) Elem	(3) Middle	(4)	(5)	(6)	(7) IV
Total Count	.0218***	.0277***	.0231***	.0206***	.0082*	.0149***	.0447
	(.0033)	(.0068)	(.0041)	(.0045)	(.0034)	(.0037)	(.0568)
(Total Count)^2	-5.8e-04***	-8.8e-04***	-5.6e-04***	-4.9e-04***	1.0e-04	-7.2e-05	-2.5e-04
	(7.9e-05)	(1.8e-04)	(9.0e-05)	(9.4e-05)	(9.1e-05)	(9.3e-05)	(.0059)
(Total Count)^3	5.6e-06***	1.4e-05***	4.4e-06**	4.5e-06**	-2.5e-06	-6.3e-07	4.2e-05
	(1.3e-06)	(3.5e-06)	(1.4e-06)	(1.5e-06)	(1.3e-06)	(1.4e-06)	(1.4e-04)
# Classes	-.0015*	-.0032*	-3.7e-04	-.0013	-.0013	-.0019**	-.0178**
	(6.8e-04)	(.0014)	(8.2e-04)	(9.4e-04)	(6.8e-04)	(7.3e-04)	(.0067)
Student FE	No	No	No	Yes	No	No	No
Years Teaching	No	No	No	No	Class	Admin	Class
R^2	.741	.739	.745	.914	.742	.742	.735
N	955876	351858	489987	583228	955876	847224	931715

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Estimated returns to the total number of students and classes previously taught by a teacher in any class with fewer than 35 students. The dependent variable is the state-wide math exam (end-of-grade) score, normalized to have mean zero and standard deviation one. Columns 1,2, and 3 contain my main estimates for all, elementary, and middle school teachers and students. Column 4 adds student fixed effects to the regression. Columns 5 and 6 control for years of experience as defined by the classroom assignment and pay code data, respectively. All regressions include teacher, school, and grade-year fixed effects. Standard errors are clustered at the teacher-year level.

Table 6: Homogeneous Coefficient (Reading)

	(1)	(2) Elem	(3) Middle	(4)	(5)	(6)	(7) IV
Total Count	.0063*	.0157**	.0065*	.0071	.0031	.0044	.138
	(.0025)	(.0054)	(.0032)	(.0036)	(.0027)	(.0028)	(.213)
$(\text{Total Count})^2$	-1.8e-04**	-1.2e-04	-2.7e-04***	-1.2e-04	-1.6e-05	-7.5e-05	-.023
	(5.5e-05)	(1.0e-04)	(7.5e-05)	(7.5e-05)	(7.0e-05)	(7.0e-05)	(.051)
$(\text{Total Count})^3$	2.1e-06*	1.6e-06	3.9e-06**	1.5e-06	2.5e-07	7.1e-07	5.3e-04
	(8.8e-07)	(1.4e-06)	(1.3e-06)	(1.3e-06)	(1.0e-06)	(1.0e-06)	(.0012)
# Classes	-5.8e-04	-.0034**	-5.1e-05	-4.9e-04	-5.6e-04	-3.9e-04	-.0041
	(5.2e-04)	(.0012)	(6.7e-04)	(7.5e-04)	(5.2e-04)	(5.5e-04)	(.0555)
Student FE	No	No	No	Yes	No	No	No
Years Teaching	No	No	No	No	Class	Admin	Class
R^2	.703	.697	.709	.902	.703	.701	.468
N	955400	347607	496314	589241	955400	850018	929662

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Estimated returns to the total number of students previously taught by a teacher in any class with fewer than 35 students. The dependent variable is the state-wide reading exam (end-of-grade) score, normalized to have mean zero and standard deviation one. Columns 1,2, and 3 contain my main estimates for all, elementary, and middle school teachers and students. Column 4 adds student fixed effects to the regression. Columns 5 and 6 control further for years of experience as defined by the classroom assignment and pay code data, respectively. All regressions include teacher, school, and grade-year fixed effects. Standard errors are clustered at the teacher-year level.

Table 7: Race-Specific Experience (Math, Logs)

	(1)	(2) Elem	(3) Middle	(4)	(5)	(6)
# White	.0048*	.008**	.0026	.0079**	-.0033	-4.2e-05
	(.002)	(.0025)	(.0033)	(.0025)	(.0027)	(.0027)
# Black	.0099***	.0113***	.0089*	.0011	.0047	.0064*
	(.0022)	(.0028)	(.0038)	(.0028)	(.0025)	(.0028)
# Hispanic	.0054*	.0029	.0063	.0029	.0036	.0046
	(.0025)	(.0031)	(.0045)	(.0031)	(.0025)	(.0029)
# Other	.0083***	.0047	.0109**	.0115***	.0069**	.0113***
	(.0022)	(.0028)	(.0041)	(.003)	(.0023)	(.0025)
Homogeneity	.452	.318	.555	.058	.054	.057
Student FE	No	No	No	Yes	No	No
Years Teaching	No	No	No	No	Class	Admin
R^2	.742	.739	.745	.914	.742	.742
N	955876	351858	489987	583228	955876	847224

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Estimated returns (in math scores) to the log of the total number of students previously taught, allowing the effect to differ according to the number of students in each of four racial categories. "Class" indicates the use of the definition of experience based on classroom assignment files, and "Admin" represents the use of the administrative pay code-based definition. Standard errors are clustered at the teacher-year level.

Table 8: Interactions with Student and Teacher Race (Math, Logs)

	(1)	(2) Elem	(3) Middle	(4)	(5)	(6)
# White	.0054*	.0072*	.0039	.0114***	-.0034	-.0016
	(.0022)	(.0028)	(.0037)	(.0028)	(.003)	(.0029)
(# White)* Black	.0016	.0012	.003	-.004	.0022	.003
	(.0016)	(.0025)	(.0023)	(.0026)	(.0016)	(.0017)
(# White)*TeacherBlack	-.0098	-.0028	-.0194*	-.0087	-.0047	.0022
	(.0064)	(.0085)	(.0095)	(.008)	(.0065)	(.0078)
# Black	.0071**	.012***	.0029	.0019	.0031	.0043
	(.0025)	(.0031)	(.0044)	(.0032)	(.0027)	(.003)
(#Black)*Black	5.4e-04	3.9e-04	-.0019	-.0051	-1.9e-04	8.1e-04
	(.0017)	(.0025)	(.0025)	(.0028)	(.0017)	(.0018)
(# Black)*TeacherBlack	.0078	-.0083	.0199*	.0069	.0029	.0014
	(.0063)	(.0094)	(.0094)	(.0072)	(.0064)	(.0085)
# Hispanic	.0059*	.0032	.0097	-.0016	.004	.0062*
	(.0027)	(.0034)	(.0051)	(.0034)	(.0027)	(.0031)
# Other Race	.0088***	.0055	.0121**	.0061	.0075**	.0121***
	(.0025)	(.0031)	(.0046)	(.0033)	(.0025)	(.0028)
Student FE	No	No	No	Yes	Yes	Yes
Years Teaching	No	No	No	No	Class	Admin
R^2	.742	.739	.745	.914	.742	.742
N	955876	351858	489987	583228	955876	847224

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Estimated returns to the log of the total number of students previously taught, allowing the effect to differ according to the number of students in each of four racial categories as well as the race of the teacher and student."Class" indicates the use of the definition of experience based on classroom assignment files, and "Admin" represents the use of the administrative pay code-based definition. Standard errors are clustered at the teacher-year level. Full table included in the Appendix.

Table 9: Tests for Homogeneous Effects Across Students (Math, Logs)

	(1)	(2) Elem	(3) Middle	(4)	(5)	(6)
# White	.273	.813	.27	.217	.188	.043
# Black	.324	.885	.034	.23	.351	.258
# Hispanic	.184	.539	.281	.001	.209	.103
# Other Race	.066	.366	.058	.000	.082	.039
All	.000	.379	.04	.000	.000	.000
Student FE	No	No	No	Yes	No	No
Years Teaching	No	No	No	No	Class	Admin
N	955876	351858	489987	583228	955876	847224

t statistics in parentheses

Note: Each column corresponds to the same regressions as in table 8. Each cell here contains the p-value of the test that *all* coefficients on terms interacting student race with the race-specific experience in that row are zero. These are treated as tests for effect heterogeneity, where the null hypothesis is homogeneity of the race-specific experience in that row-regression pair.

Table 10: Tests for Homogeneous Effects Across Teachers (Math, Logs)

	(1)	(2) Elem	(3) Middle	(4)	(5)	(6)
# White	.131	.211	.125	.218	.244	.618
# Black	.179	.133	.007	.198	.334	.036
# Hispanic	.034	.722	.002	.184	.04	.006
# Other Race	.169	.262	.036	.275	.124	.003
All	.031	.142	.000	.163	.035	.000
Student FE	No	No	No	Yes	No	No
Years Teaching	No	No	No	No	Class	Admin
N	955876	351858	489987	583228	955876	847224

t statistics in parentheses

Note: Each column corresponds to the same regressions as in table 8. Each cell here contains the p-value of the test that all coefficients on terms interacting *teacher* race with the race-specific experience in that row are zero. These are treated as tests for effect heterogeneity, where the null hypothesis is homogeneity of the race-specific experience in that row-regression pair.

Table 11: Heterogeneity Across Gender (Math, Logs)

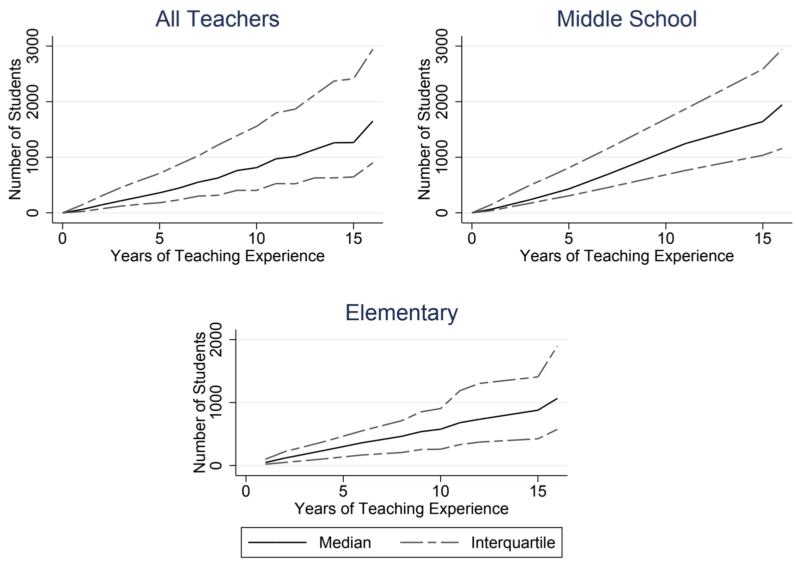
	(1)	(2)	(3) Elem	(4) Middle	(5)	(6)	(7)
# Male	.0089	.0014	.013	-.0175	.0155	-.003	-.0044
	(.0057)	(.0067)	(.0078)	(.0144)	(.0087)	(.0071)	(.0081)
# Female	.0111	.0187**	.0052	.0372*	.004	.0135*	.0209**
	(.0058)	(.0067)	(.0079)	(.0145)	(.0087)	(.0069)	(.008)
(# Male)*Male		1.9e-04	-.0025	.004	.0038	1.7e-04	.0014
		(.0043)	(.006)	(.0066)	(.0065)	(.0043)	(.0046)
(# Male)*TeacherMale		.0372*	.0605***	.0189	.034*	.0359*	.0523**
		(.0145)	(.0172)	(.0262)	(.0151)	(.0146)	(.0164)
(# Female)*Male		-2.6e-04	.0035	-.0025	-.0094	-2.3e-04	-.0013
		(.0043)	(.0061)	(.0066)	(.0065)	(.0043)	(.0046)
(# Female)*TeacherMale		-.0377**	-.0566***	-.0205	-.0374*	-.0363*	-.0509**
		(.0146)	(.017)	(.0264)	(.0152)	(.0147)	(.0166)
Student FE	No	No	No	No	Yes	No	No
Years Teaching	No	No	No	No	No	Class	Admin
R^2	.742	.742	.739	.745	.914	.742	.742
N	955876	955876	351858	489987	583228	955876	847224

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

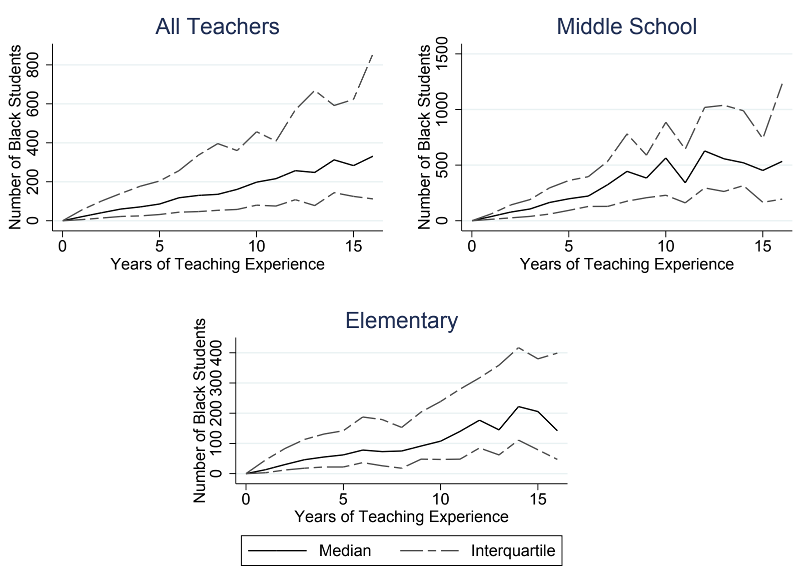
Note: Estimated returns (in math scores) to the log of the total number of students previously taught, allowing the effect to differ according to the genders of the students. Columns 1-4 present results from regressions with school and teacher fixed effects, including all teachers, elementary, and middle school teachers respectively. Column 5 replicates column 1, controlling for student fixed effects. Column 6 includes 15 indicators for years of teaching experience, and column 7 does the same using pay code data to calculate years of experience.

Figure 2: Number of Students



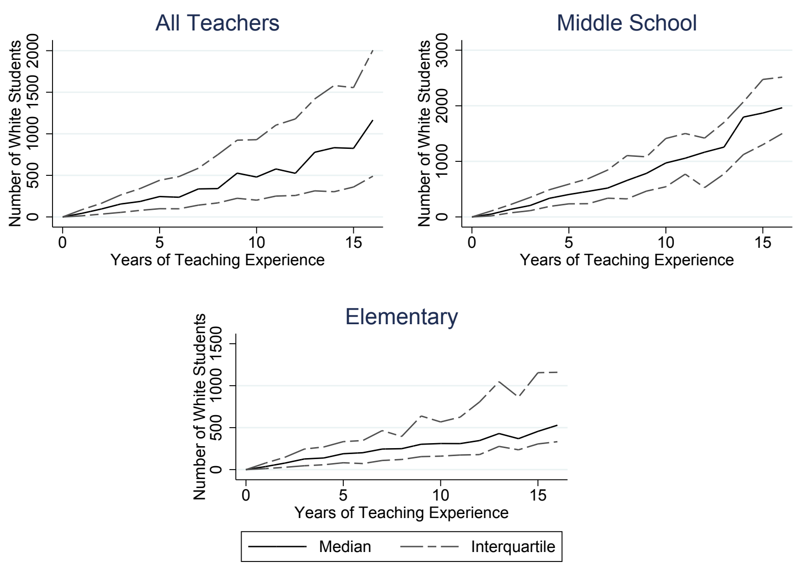
Note: Median and interquartile range of the total number of classes taught, calculated for teachers in 2011 and shown separately for elementary, middle school, and all teachers in my sample.

Figure 3: Distribution of Exposure to Black Students



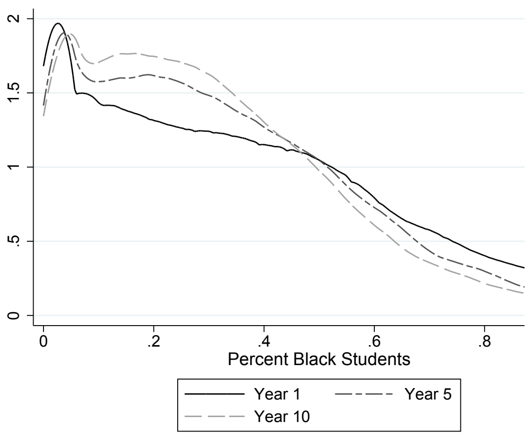
Note: Median, 25^{th}, and 75^{th} percentiles of the number of black students ever taught, by years of experience. Calculated for all teachers in 2011, and shown separately for elementary, middle school, and all teachers in my main sample.

Figure 4: Distribution of Exposure to White Students

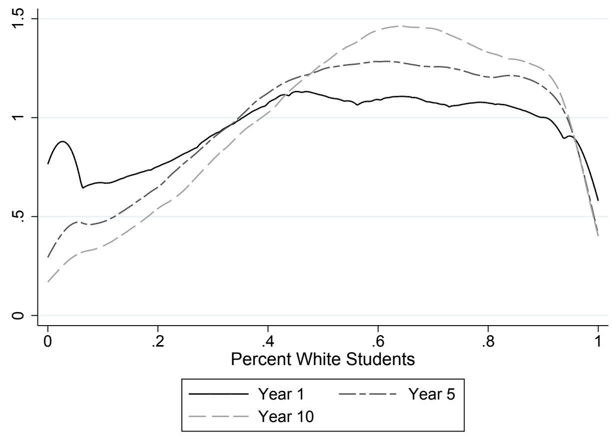


Note: Median, 25^{th}, and 75^{th} percentiles of the number of white students ever taught, by years of experience. Calculated for all teachers in 2011, and shown separately for elementary, middle school, and all teachers in my main sample.

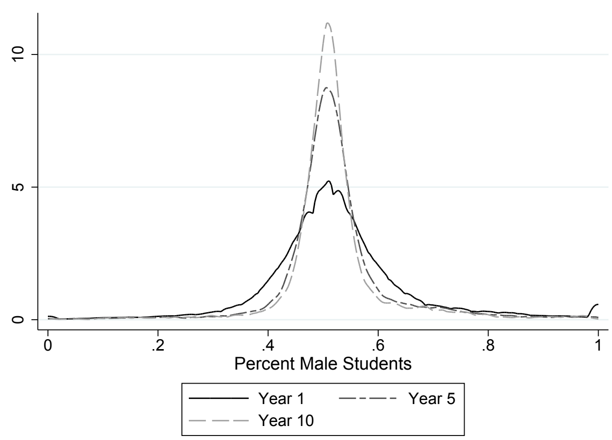
Figure 5: Raw Demographic Shares



(a)



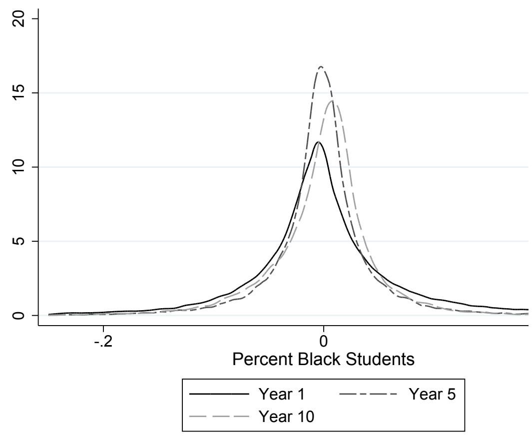
(b)



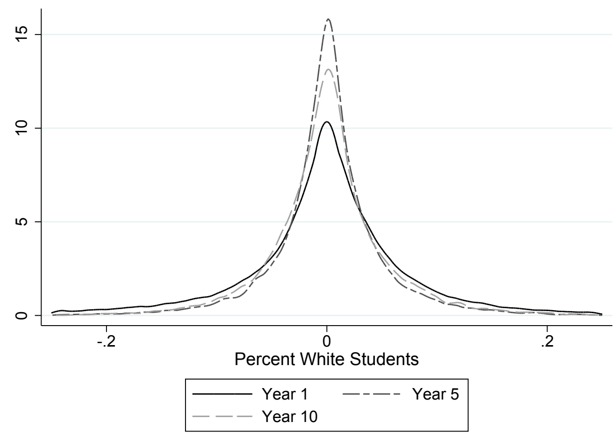
(c)

Note: Proportions of (a) black, (b) white, and (c) male students previously taught. Shown separately for teaches with 1, 5 and 10 years of experience. Calculated using all teachers in my main (math test score) sample.

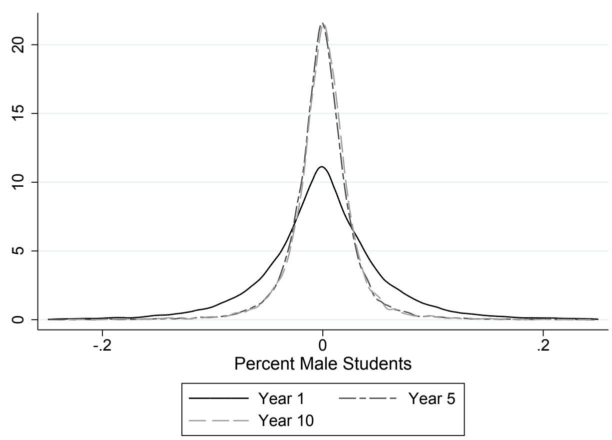
Figure 6: Residualized Demographic Shares



(a)



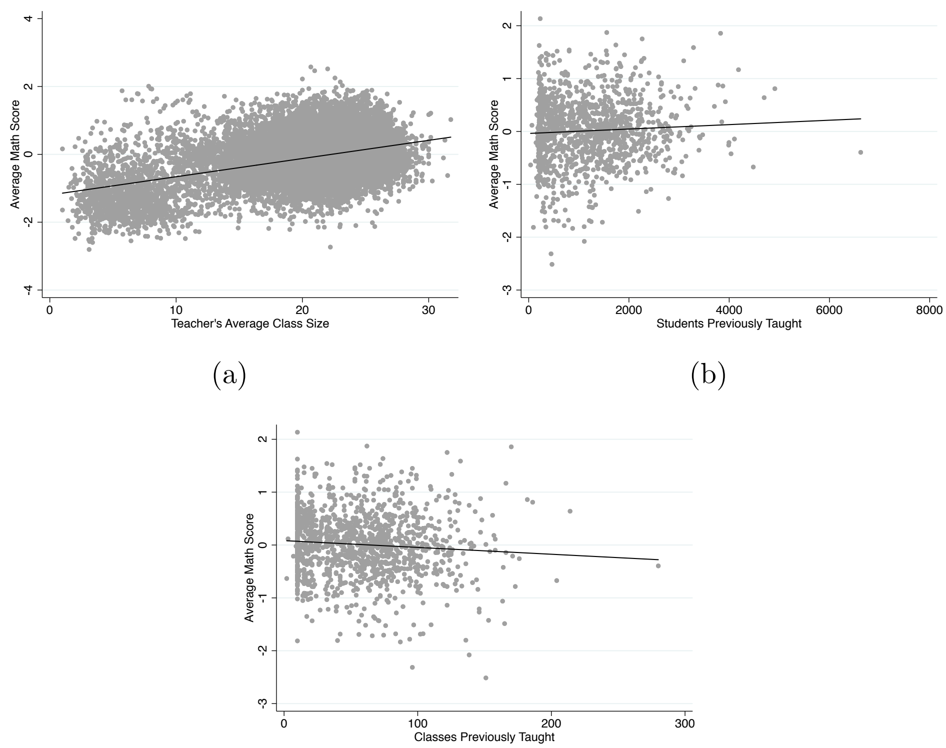
(b)



(c)

Note: Kernel density estimates of the fractions of (a) black, (b) white, and (c) male students previously taught. Shown separately for teaches with 1, 5 and 10 years of experience. Calculated using all teachers in my main (math test score) sample.

Figure 7: Class Size, Number of Classes, and Achievment

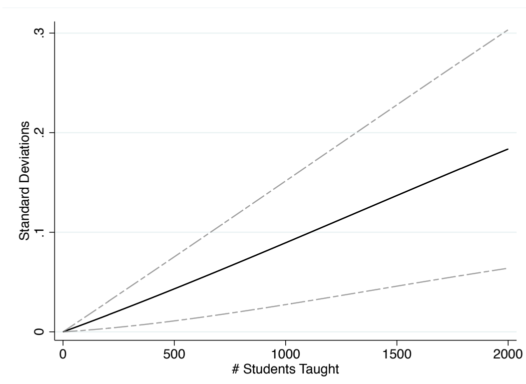


(a) (b)

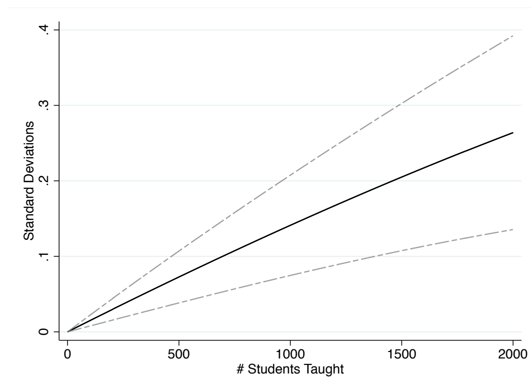
(c)

Note: Correlations and best fit lines between (a) average class size, (b) the total number of students previously taught, and (c) the number of classes previously taught, and the average scaled math score of a teacher's students in a given year. In (a) I include all teachers in my sample, and in (b) and (c) I restrict the sample to only teachers with ten years of experience, as a coarse control for average experience levels.

Figure 8: Within-Experience Effect Sizes



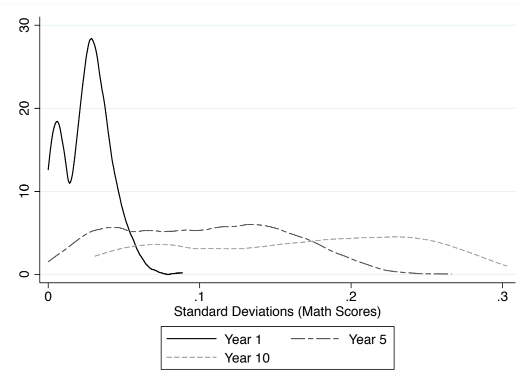
(a) Classroom Assignment



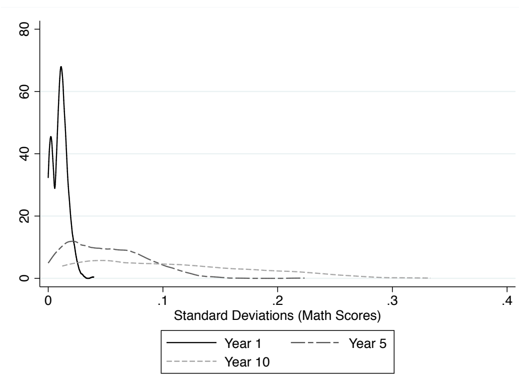
(b) Pay Code

Note: Estimated average effect (and 95% confidence intervals) of the total number of students previously taught by teacher on students' math test scores from columns 5 and 6 of table 5, shown in (a) and (b) respectively. Confidence intervals are calculated from standard errors clustered at the teacher-year level.

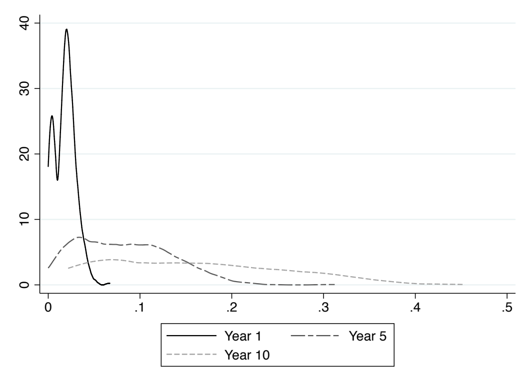
Figure 9: Distribution of Experience Effects, Homogeneous Coefficient



(a) Table 5, Column 1



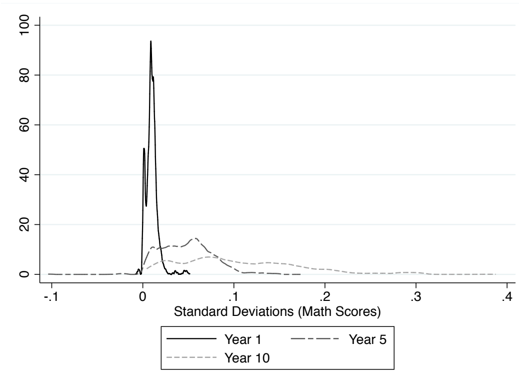
(b) Table 5, Column 5



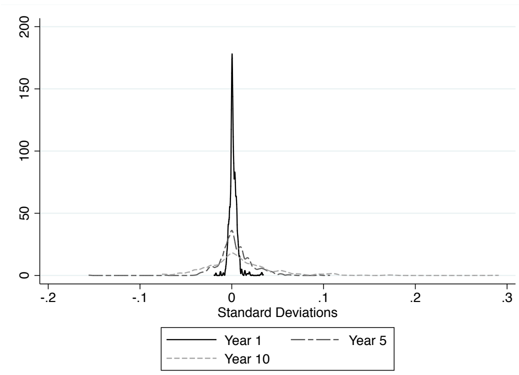
(c) Table 5, Column 6

Note: Kernel density estimates of expected teacher effectiveness resulting from the total number of students taught on math test scores. Calculated using columns 1, 5 and 6 of table 5. Figure (a) (column 1) controls for teacher, school, and grade-year fixed effects. Figures (b) and (c) (columns 5 and 6) control further for teachers' years of experience defined from the classroom assignment data and pay codes, respectively.

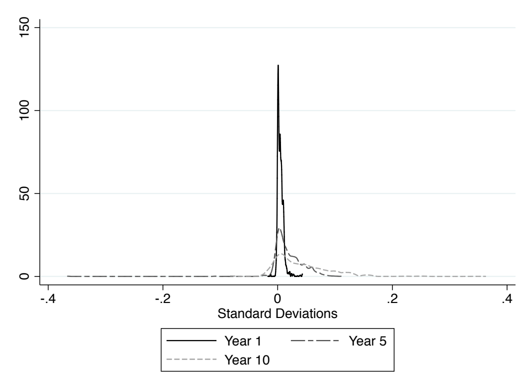
Figure 10: Heterogeneity In Teacher Effectiveness, by Race



(a) Table 8, Column 1



(b) Table 8, Column 5



(c) Table 8, Column 6

Note: Kernel density estimates of teacher effectiveness resulting from the log of the number and races of students taught (math scores). Calculated using columns 1, 5, and 6 of table 8. Figure (a) (Column 1) controls for teacher, school, and grade-year fixed effects. Figures (b) and (c) (Columns 5 and 6) control further for teachers' years of experience defined from the classroom assignment data and pay codes, respectively.

Appendix: Omitted Tables

Without School Fixed Effects

In my main results, I estimate a model with grade-by-year, teacher, and school fixed effects. Although I discuss the variation necessary for identification in the paper, it may still seem unclear where that variation can come from with so many fixed effects. To address this potential concern, I estimate my main OLS results for math scores a second time without school fixed effects. I present the results in table **??** below, which are quite similar to the results in the main text.

Table A.1: Homogeneous Coefficient, No School FE (Math)

	(1)	(2) Elem	(3) Middle	(4)	(5)	(6)
Total Count	.0217***	.0244***	.0232***	.0206***	.0085*	.0155***
	(.0033)	(.0066)	(.004)	(.0043)	(.0035)	(.0036)
$(\text{Total Count})^2$	-5.9e-04***	-8.2e-04***	-5.9e-04***	-5.4e-04***	8.0e-05	-9.5e-05
	(7.8e-05)	(1.7e-04)	(8.9e-05)	(9.3e-05)	(8.9e-05)	(9.1e-05)
$(\text{Total Count})^3$	5.7e-06***	1.4e-05***	4.8e-06***	5.3e-06***	-2.4e-06	-2.9e-07
	(1.3e-06)	(3.5e-06)	(1.4e-06)	(1.5e-06)	(1.3e-06)	(1.4e-06)
# Classes	-.0014*	-.0027*	-3.0e-04	-.0013	-.0012	-.002**
	(6.9e-04)	(.0014)	(8.3e-04)	(9.0e-04)	(6.9e-04)	(7.1e-04)
Student FE	No	No	No	Yes	No	No
Years Teaching	No	No	No	No	Class	Admin
R^2	.74	.737	.744	.913	.74	.741
N	955898	351882	490017	583262	955898	847249

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Estimated returns to the log of the total number of students previously taught by a teacher in any class with fewer than 35 students. The dependent variable is the state-wide math exam (end-of-grade) score, normalized to have mean zero and standard deviation one. Columns 1,2, and 3 contain my main estimates for all, elementary, and middle school teachers and students. Column 4 adds student fixed effects to the regression. Columns 5 and 6 control further for years of experience as defined by the classroom assignment and pay code data, respectively. All regressions include teacher and grade-year fixed effects. Standard errors are clustered at the teacher-year level.

Log/Level Specifications

In this appendix I include the estimates of my effects of interest on math scores which were excluded from the main text. In table A.2 and A.3 I show that, just as when estimated in levels, the returns to teaching more students are much larger for math than for reading, and that the returns to additional experience in reading scores is insignificant once I control for years of experience. In tables A.4, A.5, and A.6 I specify the returns from teaching additional students of each race as a quadratic function, which I then (for tables A.5 and A.6) interact with teacher and student races. These tables confirm the results from the log specifications. Specific evidence of heterogeneity for one specific racial category is mixed, but in all regressions except for those including only elementary school teachers I can reject the null that all effects are homogeneous.

In table A.7 we see one result which differs from the log specifications: although the returns (for female teachers) to teaching more male students is negligible and statistically insignificant in table 11, in some regressions here it is statistically significant and larger than the returns to female students. Because this result disappears with the inclusion of years of experience, I do not focus on it. Note that in this table, individual interaction coefficients are less precisely estimated than in table 11. Still, when I test the null hypothesis that interactions with student and teacher genders are all zero ("Student Homogeneity" and "Teacher Homogeneity" at the bottom of the table), I find evidence of heterogeneity in all but one regression. For completeness, I also include regressions with race- and gender-specific experience for reading scores in tables A.8 - A.13. I find negligible evidence of heterogeneity with respect to race or gender.

Table A.2: Homogeneous Coefficient (Math, Logs)

	(1)	(2) Elem	(3) Middle	(4)	(5)	(6)	(7) IV
Total Count	.0173***	.0167***	.017***	.0139***	.0106**	.0151***	.243*
	(9.1e-04)	(.0013)	(.0012)	(.0011)	(.0033)	(.0021)	(.0998)
Student FE	No	No	No	Yes	No	No	No
Years Teaching	No	No	No	No	Class	Admin	Class
R^2	.742	.739	.745	.914	.742	.742	.738
N	955876	351858	489987	583228	955876	847224	931715

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Estimated returns to the log of the total number of students previously taught by a teacher in any class with fewer than 35 students. The dependent variable is the state-wide math exam (end-of-grade) score, normalized to have mean zero and standard deviation one. Columns 1,2, and 3 contain my main estimates for all, elementary, and middle school teachers and students. Column 4 adds student fixed effects to the regression. Columns 5 and 6 control further for years of experience as defined by the classroom assignment and pay code data, respectively. All regressions include teacher, school, and grade-year fixed effects. Standard errors are clustered at the teacher-year level.

Table A.3: Homogeneous Coefficient (Reading, Logs)

	(1)	(2) Elem	(3) Middle	(4)	(5)	(6)	(7) IV
Total Count	.0041***	.0047***	.0036***	.0024**	.0013	.0027	.128
	(6.7e-04)	(.0011)	(8.7e-04)	(8.5e-04)	(.0024)	(.0019)	(.118)
Student FE	No	No	No	Yes	No	No	No
Years Teaching	No	No	No	No	Class	Admin	Class
R^2	.703	.697	.709	.902	.703	.701	.702
N	955400	347607	496314	589241	955400	850018	929662

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Estimated returns to the log of the total number of students previously taught by a teacher in any class with fewer than 35 students. The dependent variable is the state-wide reading exam (end-of-grade) score, normalized to have mean zero and standard deviation one. Columns 1,2, and 3 contain my main estimates for all, elementary, and middle school teachers and students. Column 4 adds student fixed effects to the regression. Columns 5 and 6 control further for years of experience as defined by the classroom assignment and pay code data, respectively. All regressions include teacher, school, and grade-year fixed effects. Standard errors are clustered at the teacher-year level.

Table A.4: Race-Specific Experience (Math, Levels)

	(1)	(2) Elem	(3) Middle	(4)	(5)	(6)
# White	.0105**	.0165*	.0101*	.0148**	.0069*	.0106**
	(.0034)	(.007)	(.0043)	(.0046)	(.0034)	(.0037)
(# White)^2	-2.1e-04***	-1.3e-04	-2.4e-04***	-3.4e-04***	8.1e-06	-3.6e-05
	(4.8e-05)	(1.1e-04)	(6.0e-05)	(6.3e-05)	(5.3e-05)	(5.4e-05)
# Black	.0228***	.0259**	.0231***	.0018	.0136**	.0176**
	(.005)	(.009)	(.0063)	(.0072)	(.005)	(.0054)
(# Black)^2	-7.3e-04**	-5.2e-04	-8.3e-04***	-5.6e-04*	-2.3e-05	-1.2e-04
	(2.3e-04)	(4.5e-04)	(2.5e-04)	(2.7e-04)	(2.3e-04)	(2.4e-04)
# Hispanic	.0197*	.0105	.0307*	.021	.0049	.0131
	(.0095)	(.0162)	(.0129)	(.0144)	(.0092)	(.0099)
(# Hispanic)^2	-.0013	-.004	-.0015	-.0015	.0012	8.3e-04
	(.0013)	(.0036)	(.0015)	(.0019)	(.0012)	(.0013)
# Other	.017***	.0224**	.0193**	.0471***	.0149***	.02***
	(.0043)	(.0077)	(.0071)	(.0089)	(.0043)	(.0047)
(# Other)^2	-5.5e-04**	-4.4e-04	-.0011**	-8.2e-04**	-3.6e-04*	-4.5e-04*
	(1.8e-04)	(2.8e-04)	(3.6e-04)	(3.2e-04)	(1.8e-04)	(1.8e-04)
Student FE	No	No	No	Yes	No	No
Years Teaching	No	No	No	No	Class	Admin
R^2	.741	.739	.745	.914	.742	.742
N	955876	351858	489987	583228	955876	847224
Homogeneity	.056	.188	.025	.000	.112	.126

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Estimated returns (in math scores) to the total number of students previously taught, allowing the effect to differ according to the number of students in each of four racial categories. "Class" indicates the use of the definition of experience based on classroom assignment files, and "Admin" represents the use of the administrative pay code-based definition. Standard errors are clustered at the teacher-year level.

Table A.5: Tests for Homogeneous Effects Across Students (Math, Levels)

	(1)	(2) Elem	(3) Middle	(4)	(5)	(6)
# White	.006	.897	.046	.187	.006	.006
# Black	.224	.408	.136	.415	.267	.196
# Hispanic	.881	.125	.955	.003	.915	.830
# Other Race	.02	.173	.000	.616	.014	.011
All	.000	.28	.001	.000	.000	.000
Student FE	No	No	No	Yes	No	No
Years Teaching	No	No	No	No	Class	Admin
N	955876	351858	489987	583228	955876	847224

t statistics in parentheses

Note: Each column corresponds to the same regressions as in table 8. Each cell here contains the p-value of the test that *all* coefficients on terms interacting student race with the race-specific experience in that row are zero. These are treated as tests for effect heterogeneity, where the null hypothesis is homogeneity of the race-specific experience in that row-regression pair.

Table A.6: Tests for Homogeneous Effects Across Teachers (Math, Levels)

	(1)	(2) Elem	(3) Middle	(4)	(5)	(6)
# White	.099	.149	.422	.018	.221	.1
# Black	.056	.29	.000	.145	.097	.104
# Hispanic	.002	.991	.000	.038	.007	.017
# Other Race	.535	.967	.138	.000	.524	.323
All	.018	.227	.000	.000	.065	.01
Student FE	No	No	No	Yes	No	No
Years Teaching	No	No	No	No	Class	Admin
N	955876	351858	489987	583228	955876	847224

t statistics in parentheses

Note: Each column corresponds to the same regressions as in table 8. Each cell here contains the p-value of the test that all coefficients on terms interacting *teacher* race with the race-specific experience in that row are zero. These are treated as tests for effect heterogeneity, where the null hypothesis is homogeneity of the race-specific experience in that row-regression pair.

Table A.7: Heterogeneity Across Gender (Math, Levels)

	(1)	(2)	(3) Elem	(4) Middle	(5)	(6)	(7)
# Male	.0218***	.0195***	.0304**	.0123	.0331***	.0106	.0126
	(.0056)	(.0059)	(.0098)	(.0107)	(.0085)	(.006)	(.0064)
$(\#Male)^2$	-1.0e-03***	-9.6e-04**	-9.6e-04	-8.8e-04	-.0013**	-3.9e-04	-3.4e-04
	(3.0e-04)	(3.2e-04)	(5.6e-04)	(5.3e-04)	(3.9e-04)	(3.2e-04)	(3.5e-04)
# Female	.0128*	.017**	.0144	.0261*	.0074	.0128*	.0209***
	(.0053)	(.0058)	(.0088)	(.0107)	(.0081)	(.0058)	(.0062)
$(\#Female)^2$	-7.1e-06	-2.7e-05	2.9e-04	-2.5e-04	2.6e-04	1.1e-04	-1.4e-04
	(2.9e-04)	(3.2e-04)	(5.9e-04)	(5.3e-04)	(3.8e-04)	(3.2e-04)	(3.5e-04)
(# Male)*StudentMale		-6.0e-04	-1.1e-04	5.7e-04	6.0e-04	-5.8e-04	-6.8e-04
		(.0026)	(.004)	(.0039)	(.0039)	(.0026)	(.0027)
(# Male)*TeacherMale		.0186	.0265	.0154	.0183	.0127	.0164
		(.0119)	(.0213)	(.0206)	(.0135)	(.0117)	(.0127)
$(\#Male)^2$*TeacherMale		-6.2e-04	-.0019	-3.2e-04	-.0012	-6.5e-05	-4.0e-04
		(8.8e-04)	(.0026)	(.0012)	(.001)	(8.6e-04)	(9.2e-04)
$(\#Male)^2$*StudentMale		1.4e-04	4.1e-05	1.2e-04	2.2e-04	1.3e-04	1.4e-04
		(1.2e-04)	(2.4e-04)	(1.8e-04)	(1.7e-04)	(1.2e-04)	(1.3e-04)
(# Female)*StudentMale		-9.3e-04	7.6e-04	-3.3e-04	-.0074	-9.6e-04	-.001
		(.0026)	(.0042)	(.004)	(.004)	(.0026)	(.0028)
(# Female)*TeacherMale		-.0177	-.0226	-.0154	-.0223	-.0154	-.0184
		(.0116)	(.0219)	(.0204)	(.0134)	(.0114)	(.0123)
$(\#Female)^2$*StudentMale		-3.3e-05	-5.7e-05	-6.4e-05	3.0e-05	-3.0e-05	-2.9e-05
		(1.3e-04)	(2.6e-04)	(1.8e-04)	(1.8e-04)	(1.3e-04)	(1.3e-04)
$(\#Female^2)$*TeacherMale		-7.0e-05	9.4e-04	-3.3e-04	8.1e-04	-3.0e-04	-9.4e-05
		(8.3e-04)	(.0028)	(.0011)	(9.8e-04)	(8.2e-04)	(8.7e-04)
Student Homogeneity		.03	.951	.004	.000	.029	.03
Teacher Homogeneity		.002	.336	.019	.031	.01	.005
Student FE	No	No	No	No	Yes	No	No
Years Teaching	No	No	No	No	No	Class	Admin
R^2	.741	.741	.739	.745	.914	.742	.742
N	955876	955876	351858	489987	583228	955876	847224

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Estimated returns (in math scores) to the total number of students previously taught, allowing the effect to differ according to the genders of the students. Columns 1-4 present results from regressions with school and teacher fixed effects, including all teachers, elementary, and middle school teachers respectively. Column 5 replicates column 1, controlling for student fixed effects. Column 6 includes 15 indicators for years of teaching experience, and column 7 does the same using pay code data to calculate years of experience.

Additional Reading Results

Estimated in Levels

Table A.8: Race-Specific Experience (Reading, Levels)

	(1)	(2) Elem	(3) Mid	(4)	(5)	(6)
# White	.0073**	.0173**	.0064	.0042	.0059*	.0067*
	(.0027)	(.0058)	(.0033)	(.0039)	(.0027)	(.0028)
(# White)^2	-1.2e-04**	-9.8e-05	-1.2e-04*	-5.9e-05	-2.9e-05	-7.7e-05
	(4.1e-05)	(8.7e-05)	(5.3e-05)	(6.4e-05)	(4.7e-05)	(4.7e-05)
# Black	.0018	.0182*	-.0015	.0013	-6.4e-04	-.0019
	(.0034)	(.0074)	(.0043)	(.0055)	(.0034)	(.0037)
(# Black)^2	1.3e-05	-1.9e-04	1.4e-04	7.5e-06	2.2e-04	2.1e-04
	(1.2e-04)	(4.7e-04)	(1.4e-04)	(1.7e-04)	(1.3e-04)	(1.3e-04)
# Hispanic	-.0045	.0019	-.0082	.0191	-.0081	-.0087
	(.0069)	(.0132)	(.0096)	(.0118)	(.0069)	(.0074)
(# Hispanic)^2	-1.6e-04	-.001	2.8e-04	9.9e-04	5.7e-04	.0012
	(9.6e-04)	(.003)	(.0014)	(.0015)	(9.6e-04)	(9.9e-04)
# Other	-.0011	.0106	.008	.0125*	-.0016	-3.5e-04
	(.0034)	(.0063)	(.0053)	(.0063)	(.0034)	(.0036)
(# Other)^2	-7.8e-06	1.4e-04	-.0011**	-2.3e-04	3.2e-05	-1.6e-05
	(9.4e-05)	(1.0e-04)	(4.0e-04)	(1.4e-04)	(9.2e-05)	(9.7e-05)
Student FE	No	No	No	Yes	No	No
Years Teaching	No	No	No	No	Class	Admin
N	.703	.697	.709	.902	.703	.701
R^2	955400	347607	496314	589241	955400	850018
Wald Test	.012	.493	.049	.359	.008	.004

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Estimated returns (in reading scores) to the total number of students previously taught, allowing the effect to differ according to the number of students in each of four racial categories. "Class" indicates the use of the definition of experience based on classroom assignment files, and "Admin" represents the use of the administrative pay code-based definition. Standard errors are clustered at the teacher-year level.

Table A.9: Model with Interactions with Student and Teacher Race (Reading, Levels)

	(1)	(2) Elem	(3) Middle	(4)	(5)	(6)
# White	.0076**	.0198***	.0064	.004	.0062*	.0073*
	(.0027)	(.0058)	(.0034)	(.004)	(.0027)	(.0029)
(# White)^2	-1.1e-04**	-9.9e-05	-9.4e-05	-4.0e-05	-2.7e-05	-8.2e-05
	(4.2e-05)	(9.0e-05)	(5.5e-05)	(6.7e-05)	(4.7e-05)	(4.8e-05)
# Black	3.5e-04	.0207*	-.0038	2.3e-04	-.0026	-.0046
	(.0039)	(.0082)	(.0051)	(.0064)	(.004)	(.0043)
(# Black)^2	-4.3e-05	-4.1e-04	1.1e-04	-4.1e-05	2.3e-04	2.6e-04
	(2.0e-04)	(6.4e-04)	(2.3e-04)	(2.7e-04)	(2.0e-04)	(2.2e-04)
# Hispanic	-.0076	6.9e-04	-.0122	.0184	-.011	-.0107
	(.0077)	(.0145)	(.0109)	(.0131)	(.0077)	(.0083)
(# Hispanic)^2	3.0e-04	-2.8e-04	5.8e-04	.0028	9.5e-04	.0015
	(.0011)	(.0036)	(.0015)	(.0017)	(.0011)	(.0011)
# Other Race	-9.8e-05	.006	.0235*	.0138	-8.2e-04	7.1e-04
	(.0044)	(.0097)	(.0092)	(.0081)	(.0044)	(.0047)
(# Other Race)^2	-2.3e-04	-1.6e-04	-.0042*	-8.4e-04	-1.4e-04	-2.6e-04
	(3.6e-04)	(.0017)	(.0018)	(6.1e-04)	(3.7e-04)	(3.7e-04)
Student FE	No	No	No	Yes	Yes	Yes
Years Teaching	No	No	No	No	Class	Admin
R^2	.703	.697	.709	.902	.703	.701
N	955400	347607	496314	589241	955400	850018

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Estimated returns (in reading scores) to the total number of students previously taught, allowing the effect to differ according to the number of students in each of four racial categories as well as the race of the teacher and student. Returns are estimated as a cubic function under three specifications: columns 1-3 present results from regressions with school and teacher fixed effects, including all teachers, elementary, and middle school teachers respectively. Column 4 replicates column 1, controlling for student-level averages of all independent variables. Column 5 includes 15 indicators for years of teaching experience, and column 6 does the same using pay code data to calculate years of experience.

Table A.10: Heterogeneity Across Gender (Reading, Levels)

	(1)	(2)	(3) Elem	(4) Middle	(5)	(6)	(7)
# Male	.0092*	.0051	.0068	.0077	.0035	.0021	.007
	(.0043)	(.0047)	(.0079)	(.0072)	(.007)	(.0047)	(.0051)
$(\#Male)^2$	-5.7e-04*	-4.0e-04	-1.7e-04	-4.1e-04	-7.8e-05	-2.1e-04	-3.6e-04
	(2.3e-04)	(2.5e-04)	(4.9e-04)	(3.6e-04)	(3.4e-04)	(2.5e-04)	(2.6e-04)
# Female	7.3e-04	.0042	.0207**	2.2e-04	.0075	.0036	9.5e-04
	(.004)	(.0043)	(.0072)	(.0068)	(.0068)	(.0044)	(.0047)
$(\#Female)^2$	3.3e-04	1.8e-04	1.2e-04	1.9e-04	-4.0e-05	2.1e-04	2.4e-04
	(2.3e-04)	(2.4e-04)	(4.7e-04)	(3.6e-04)	(3.3e-04)	(2.4e-04)	(2.6e-04)
(# Male)*StudentMale		.0045	.0085*	1.9e-04	.0036	.0044	.0051
		(.0025)	(.0035)	(.0042)	(.0038)	(.0025)	(.0026)
(# Male)*TeacherMale		.0131	.0213	.0057	-.0193	.0124	.0034
		(.0103)	(.017)	(.0189)	(.0136)	(.0102)	(.0116)
$(\#Male)^2$*TeacherMale		-5.3e-04	-8.0e-04	-2.1e-04	.0026*	-5.2e-04	-1.2e-04
		(8.7e-04)	(.002)	(.0011)	(.001)	(8.7e-04)	(.0011)
$(\#Male)^2$*StudentMale		-1.5e-04	-3.0e-04	1.1e-04	-1.3e-04	-1.5e-04	-1.7e-04
		(1.2e-04)	(1.7e-04)	(2.1e-04)	(1.6e-04)	(1.2e-04)	(1.2e-04)
(# Female)*StudentMale		-.0034	-.0077*	-2.9e-04	-.0021	-.0033	-.0039
		(.0024)	(.0034)	(.0041)	(.0037)	(.0024)	(.0025)
(# Female)*TeacherMale		-.0099	-.0157	-.0017	.0265	-.0104	-.0059
		(.01)	(.0188)	(.0183)	(.0136)	(.0098)	(.0111)
$(\#Female)^2$*StudentMale		1.4e-04	2.7e-04	-4.3e-05	6.9e-05	1.4e-04	1.6e-04
		(1.1e-04)	(1.5e-04)	(1.9e-04)	(1.5e-04)	(1.1e-04)	(1.1e-04)
$(\#Female^2)$*TeacherMale		2.6e-04	-2.3e-04	-1.4e-04	-.0029**	3.1e-04	1.2e-04
		(8.8e-04)	(.0024)	(.0011)	(.0011)	(8.7e-04)	(.001)
Student FE	No	No	No	No	Yes	No	No
Years Teaching	No	No	No	No	No	Class	Admin
R^2	.703	.703	.697	.709	.902	.703	.701
N	955400	955400	347607	496314	589241	955400	850018

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Estimated returns (in reading scores) to the total number of students previously taught, allowing the effect to differ according to the genders of the students. Returns are estimated as a cubic function under three specifications: columns 1-4 present results from regressions with school and teacher fixed effects, including all teachers, elementary, and middle school teachers respectively. Column 5 replicates column 1, controlling for student-level averages of all independent variables. Column 6 includes 15 indicators for years of teaching experience, and column 7 does the same using pay code data to calculate years of experience.

Estimated in Logs

Table A.11: Race-Specific Experience (Reading, Logs)

	(1)	(2) Elem	(3) Mid	(4)	(5)	(6)
# White	.0038*	.0034	.0031	.0036	.002	.0041
	(.0016)	(.0021)	(.0025)	(.002)	(.002)	(.0021)
# Black	.0032	.0083***	-2.2e-05	-.0028	.0022	6.1e-04
	(.0017)	(.0024)	(.0026)	(.0022)	(.0019)	(.0021)
# Hispanic	-.0017	-.0028	-.0019	-.0013	-.0024	-.0037
	(.002)	(.0026)	(.0033)	(.0027)	(.002)	(.0022)
# Other	-8.2e-04	-.0046*	.0039	.004	-.0013	4.9e-04
	(.0017)	(.0023)	(.0027)	(.0023)	(.0017)	(.002)
Student FE	No	No	No	Yes	No	No
Years Teaching	No	No	No	No	Class	Admin
N	.703	.697	.709	.902	.703	.701
R^2	955400	347607	496314	589241	955400	850018
Homogeneity	.174	.005	.439	.101	.435	.147

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Estimated returns (in reading scores) to the log of the total number of students previously taught, allowing the effect to differ according to the number of students in each of four racial categories. "Class" indicates the use of the definition of experience based on classroom assignment files, and "Admin" represents the use of the administrative pay code-based definition. Standard errors are clustered at the teacher-year level.

Table A.12: Interactions with Student and Teacher Race (Reading, Logs)

	(1)	(2) Elem	(3) Middle	(4)	(5)	(6)
# White	.0033	.0039	.0012	.0029	.002	.0036
	(.0017)	(.0024)	(.0028)	(.0022)	(.0023)	(.0024)
(# White)*Black	.0034*	.0021	.0035	9.2e-04	.0033*	.0033*
	(.0015)	(.0024)	(.0022)	(.0026)	(.0015)	(.0016)
(# White)*TeacherBlack	-.0024	-.0024	.0045	.0051	-.0018	-.0029
	(.0049)	(.0085)	(.0065)	(.0064)	(.005)	(.0054)
# Black	.0076	.023**	-1.7e-04	-.0084	.0061	.0058
	(.0041)	(.0071)	(.0054)	(.0053)	(.0044)	(.0048)
(#Black)*White	.0019	.0016	.0023	.0011	.0019	.0027
	(.0014)	(.0022)	(.0022)	(.0028)	(.0015)	(.0015)
(# Black)*TeacherWhite	-.008	-.0196**	-.0026	.0067	-.0071	-.0092
	(.0045)	(.0075)	(.0062)	(.0058)	(.0046)	(.0052)
# Hispanic	.0355	.0264	.076	.0414	.0345	.0845**
	(.0186)	(.0191)	(.147)	(.0275)	(.0189)	(.0297)
# Other Race	.025*	.036**	.0065	.0086	.0241*	.0297*
	(.0108)	(.0124)	(.0241)	(.0155)	(.0109)	(.0128)
Student FE	No	No	No	Yes	Yes	Yes
Years Teaching	No	No	No	No	Class	Admin
R^2	.703	.697	.709	.902	.703	.701
N	955400	347607	496314	589241	955400	850018

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Estimated returns (in reading scores) to the log of the total number of students previously taught, allowing the effect to differ according to the number of students in each of four racial categories as well as the race of the teacher and student. Returns are estimated as a cubic function under three specifications: columns 1-3 present results from regressions with school and teacher fixed effects, including all teachers, elementary, and middle school teachers respectively. Column 4 replicates column 1, controlling for student-level averages of all independent variables. Column 5 includes 15 indicators for years of teaching experience, and column 6 does the same using pay code data to calculate years of experience.

Table A.13: Heterogeneity Across Gender (Reading, Logs)

	(1)	(2)	(3) Elem	(4) Middle	(5)	(6)	(7)
# Male	2.2e-03	-2.1e-03	-1.0e-02	2.4e-03	-3.9e-03	-4.3e-03	1.3e-03
	(4.5e-03)	(5.1e-03)	(6.8e-03)	(8.4e-03)	(7.2e-03)	(5.4e-03)	(6.6e-03)
# Female	2.6e-03	5.7e-03	1.5e-02*	4.5e-04	5.1e-03	4.2e-03	5.7e-04
	(4.6e-03)	(5.2e-03)	(6.9e-03)	(8.5e-03)	(7.3e-03)	(5.2e-03)	(6.5e-03)
(# Male)*Male		7.9e-03*	8.5e-03	1.1e-02	9.9e-03	7.9e-03*	9.6e-03*
		(4.0e-03)	(5.2e-03)	(6.7e-03)	(6.4e-03)	(4.0e-03)	(4.3e-03)
(# Male)*TeacherMale		9.2e-04	2.1e-02	-3.1e-02	-2.7e-02	8.2e-04	-8.1e-03
		(1.3e-02)	(2.1e-02)	(2.0e-02)	(1.4e-02)	(1.3e-02)	(1.4e-02)
(# Female)*Male		-6.3e-03	-8.2e-03	-9.1e-03	-8.1e-03	-6.3e-03	-7.9e-03
		(4.0e-03)	(5.2e-03)	(6.6e-03)	(6.4e-03)	(4.0e-03)	(4.3e-03)
(# Female)*TeacherMale		2.3e-03	-1.7e-02	3.4e-02	3.2e-02*	2.3e-03	9.9e-03
		(1.3e-02)	(2.2e-02)	(2.0e-02)	(1.5e-02)	(1.3e-02)	(1.4e-02)
Student FE	No	No	No	No	Yes	No	No
Years Teaching	No	No	No	No	No	Class	Admin
R^2	.703	.703	.697	.709	.902	.703	.701
N	955400	955400	347607	496314	589241	955400	850018

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Estimated returns (in reading scores) to the log of the total number of students previously taught, allowing the effect to differ according to the genders of the students. Returns are estimated as a cubic function under three specifications: columns 1-4 present results from regressions with school and teacher fixed effects, including all teachers, elementary, and middle school teachers respectively. Column 5 replicates column 1, controlling for student-level averages of all independent variables. Column 6 includes 15 indicators for years of teaching experience, and column 7 does the same using pay code data to calculate years of experience.

Appendix: Data

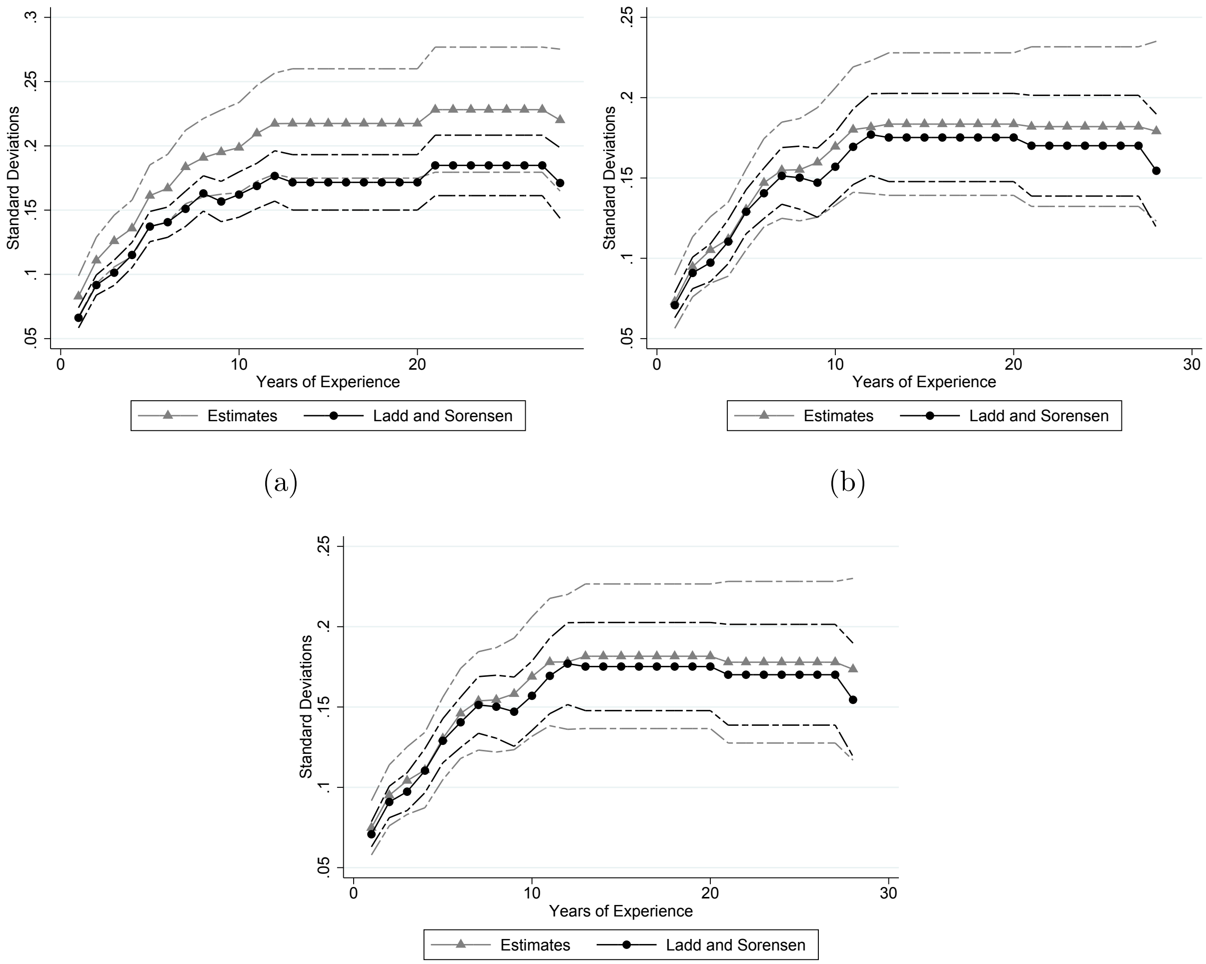
Full List of Controls

- *Teacher:* Race, gender, lateral entry, licensed, advanced degree
- *Student:* Race, gender, indicators for being academically in math and reading, cubic in lagged score
- *Student-Teacher:* Same race, same gender, same race and gender (interacted with each combination of student and teacher race and gender
- *Classroom:* Class size, percent white, percent black, percent free or reduced price lunch, average lagged math score, average lagged reading score. All but class size are calculated among other peers within the classroom.
- *Experience:* Number of classrooms ever assigned which contained 35 or fewer students
  - Pay code: Dummy variables for 1-15 years of experience and bins for 16-20, 21-27, and more than 27 years.
  - Classroom data: Dummy variables for 1-16 years of experience

Replication

The sample construction procedure I follow for all students is the procedure used by Ladd and Sorensen (2017) for middle school students and teachers. I also use the same years of North Carolina data. They impute a few controls for some years (e.g. those related to parental education), but in principle my and their estimates of the returns to years of experience should be comparable. In figure 11 I show the results of estimating my approximation of the regressions from columns 1 and 2 of their table 2. Surrounding their estimates I plot the robust standard errors they report, and I report teacher-year clustered errors for my estimates, consistent with the rest of my paper. In figure 11(a) I estimate a regression much like column 1 of that table and report the estimated returns to years of experience from my sample and from Ladd and Sorensen (2017). Although the point estimates differ, our confidence intervals overlap significantly. When I add student fixed effects to this regression (figure 11(b)) or run a regression like their column 2 (which includes student fixed effects, figure 11(c)), the point estimates are nearly identical. I take this as evidence that my sample construction procedure is reasonable.

Figure 11: Comparison with Ladd and Sorensen (2017)



(a)

(b)

(c)

Note: In (a) and (b) I estimate regressions most comparable to columns 1 and 2 of table 2 in Ladd and Sorensen (2017). In (c) I add student fixed effects to the regression in (a). In (a) and (c) I also show the estimates reported in column 1 therein, and in (b) I show the estimates in their column 2. As in the rest of this paper, confidence intervals for my estimates are clustered at the teacher-year level. Only robust standard errors are reported in Ladd and Sorensen (2017), so I use those to construct confidence intervals for their comparison estimates.

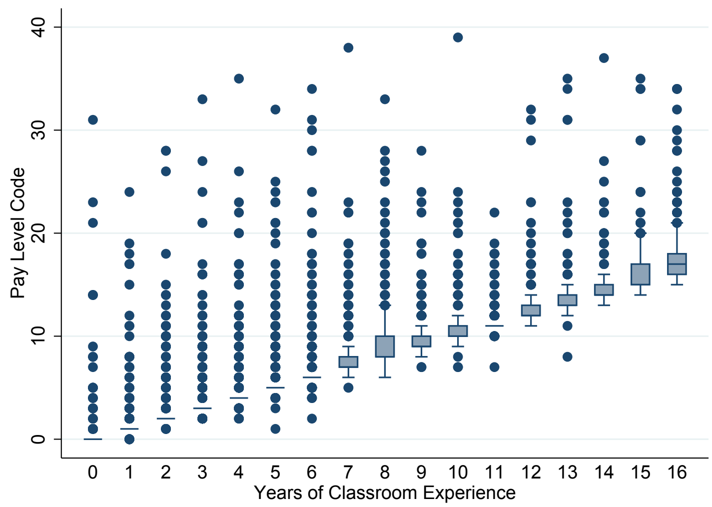
Measuring Years of Experience

As discussed in the paper, the same files used to determine classroom composition also provide a measure of years of experience. This allows me to construct years of experience even for teachers who are not matched to pay code data. In this appendix I discuss the validity of this alternative measure of experience and the ways in which it differs from pay code data when both are available.

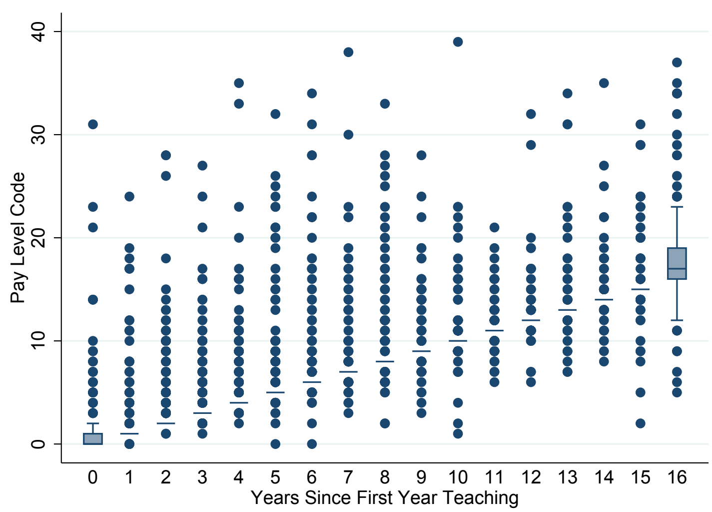
The classroom data I use to construct experience is assembled from two sets of files. The first, containing personnel information, contains information on all "those employed by the public school system who have direct student contact at a public school in a classroom or non-classroom activity for which a state course code or personnel assignment type exists," according to the data description provided by the NCERDC. The second, called the School Activity Report (SAR), contains "a snapshot of all activities occurring in a school on a given day. Activities include traditional academic classes as well as non-class events (e.g., study hall, lunch period). The file includes activities that meet all year as well as those that meet for only part of the year." Another documentation file describes the SAR as " a collection of data that shows an individual school's full-year academic schedule, courses offered, enrollment in classes, length of classes, and staffing of classes." The classroom assignment data should therefore be a full count of teachers in any regular contact with students. This gives the measure of years experience implied by this data some credibility.

I mention in the main text that counting a teacher's years of experience does not always match her years of experience as indicated by the salary data. For example, if a teacher entered the profession in my earliest year of data (1995), then by 2010 she would have 15 years of experience. However, in 2010, approximately 5% of teachers whose entire careers should be observed have more than 15 years of experience in the file containing pay codes. In figure 12 I show box plots of experience as defined by the pay code data over (a) the actual number of years I observe teachers in a classroom and (b) the number of years since a teacher's first year teaching. Both show, quite surprisingly, that even teachers who appear to have just begun teaching may be marked as having ten or more years of experience in the pay code data. In fact, even in my main sample, where I have restricted teachers to have begun teaching in or after 1995, I can estimate the coefficients on dummy variables for 16-20, 21-27, and greater than 28 years of experience. I include these observations (those with more experience than the classroom data indicates is possible) when I control for pay code experience levels, because (i) the teachers for whom the pay code and classroom assignment measures differ may be nonrandom (meaning dropping them would bias my results) and (ii) the pay code variables may accurately capture the sum of teachers' years teaching and working as administrators, which may be a valid definition of experience.

Figure 12: Comparing Measures of Experience



(a)



(b)

Note: The y-axis in (a) is teacher experience as defined by the administrative pay code data, and in (b) is the maximum number years a teacher could possibly have, given the year in which the classroom data indicates she began teaching. The x-axis is the number of years a teacher has been observed teaching a standard classroom assignment. The sample I use here is the set of all teachers in 2010 for whom I observe the beginning of their career and who do not exit the teaching profession at any point.

Appendix: Additional Estimates

In this Appendix I present three additional tables exploring the returns to teaching more students. First, I address the potential concern that my main results are only a result of teachers on the far tails of the experience distribution. To address this concern, I calculate the average number of students and classes taught by each teacher, and drop any teacher-year observation for which the teacher is in the upper quartile of the distribution of either of these variables. More precisely, I drop any teachers who have taught more than 220 students, or more than 8 classes, per year on average. A related concern is that the marginal returns are only large very early (with respect to the number of students she has taught) in a teacher's career and are quickly diminishing. In this way, teachers with very *few* students could be driving my results. To alleviate this concern, I also drop teachers who have taught fewer than 80 students per year, which is approximately the 25^{th} percentile of that distribution. I report estimates in table A.14 from all regressions reported previously (for the main sample) in table 5. Although my estimates are statistically insignificant in columns 5 and 6 (when I control for years of experience), I take the fact that the point estimates do not change significantly in these columns, as well as the significance of the estimate in column 7 (which uses the IV described in the paper) that these potential outliers are not the force driving my main estimates.

Because the more common approach to data construction in studies using the NCERDC is to assume that the teachers who are matched to students in the test score files (i.e. the proctor of the exam) are also the student's teacher, one may also worry that my construction of the data is driving my results. With this in mind, for any interested readers, I construct my sample in this more standard way, and report estimates from the regressions in columns 1,4,5, and 6 of table 5 using only fourth and fifth graders. These new estimates are shown in table A.15. The sample sizes in this set of regressions (especially when student fixed effects are included) are small relative to my main estimates, but in columns 1 and 2 I can easily reject the null hypothesis that all coefficients are zero. The linear coefficient in column 1 is significant at the 10% level and is of the same magnitude as in table 5. In columns 3 and 4, when I control for years of experience, the estimates shrink significantly just as in the main sample. In these last two columns, I cannot reject the joint test that all coefficients are zero, though the estimates of column 1 are contained in the confidence intervals of columns 3 and 4.

In the main text, I take the similarity of the estimates using middle and elementary school students separately as support for pooling both groups for many regressions. If columns 3 and 4 in table A.15 represent true zeros conditional on years of experience, then it is possible that the same is true for middle school teachers, and that the estimates in columns 5 and 6 of table 5 arise only because the two are pooled. Switching back to my main sample, I show in table A.16 estimates of column 3 of table 5 which control further for years of experience. In both of these regressions the effect sizes are comparable to those using the full sample (i.e. pooling elementary and middle school teachers), and the null hypothesis that all coefficients are zero is easily rejected ($p < 0.01$) in both regressions.

Table A.14: Homogeneous Coefficient (Math, Levels)

	(1)	(2) Elem	(3) Middle	(4)	(5)	(6)	(7) IV
Total Count	.0248**	.017	.0283**	.0375*	.013	.0135	.319*
	(.0076)	(.0167)	(.009)	(.0172)	(.0109)	(.0092)	(.132)
(Total Count)^2	-8.1e-04**	-.0012	-7.5e-04*	-.0016**	-1.5e-06	-1.7e-04	
	(2.8e-04)	(6.9e-04)	(3.4e-04)	(5.1e-04)	(5.6e-04)	(4.2e-04)	
(Total Count)^3	1.6e-05*	4.1e-05*	1.3e-05	3.7e-05**	7.7e-07	3.2e-06	
	(7.0e-06)	(1.9e-05)	(8.1e-06)	(1.3e-05)	(1.1e-05)	(9.1e-06)	
Student FE	No	No	No	Yes	No	No	No
Years Teaching	No	No	No	No	Class	Admin	Class
R^2	.739	.739	.739	.923	.739	.739	.733
N	361615	87645	219317	123404	361615	325315	272235

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Estimates from regressions identical to those in table 5, where I exclude teachers in the the upper and lower quartiles of students-per-year distribution, and those in the upper quartile of the classes-per-year distribution.

Table A.15: Homogeneous Coefficient (Math, Levels, Elementary School)

	(1)	(2)	(3)	(4)
Total Count	.0117	-.0056	8.1e-04	.0013
	(.0063)	(.0183)	(.0064)	(.0067)
(Total Count)^2	-7.4e-04***	-7.5e-04**	-1.6e-05	-1.5e-04
	(1.2e-04)	(2.7e-04)	(1.4e-04)	(1.4e-04)
(Total Count)^3	1.0e-05***	1.3e-05**	1.1e-06	2.9e-06
	(1.9e-06)	(4.6e-06)	(1.8e-06)	(1.8e-06)
Student FE	No	Yes	No	No
Years Teaching	No	No	Class	Admin
R^2	.737	.947	.737	.737
N	347269	116609	347269	313000

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Estimates from the same regressions as columns 1,4,5, and 6 (respectively) of table 5. Data consists of fourth and fifth grade students only, where the proctor of the test is assumed to be the student's classroom teacher.

Table A.16: Homogeneous Coefficient (Math, Levels, Middle School)

	(1)	(2)
Total Count	.0058	.0163***
	(.005)	(.0049)
(Total Count)^2	1.9e-04	-8.5e-05
	(1.5e-04)	(1.3e-04)
(Total Count)^3	-4.0e-06*	-1.3e-06
	(1.8e-06)	(1.8e-06)
Years Teaching	Class	Admin
R^2	.745	.747
N	489987	422690

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Estimates from regressions nearly identical to column 3 of table 5, differing only in that they are also conditional on teachers' years of experience (defined by the two measures discussed in the main text).

Appendix: Full Interacted Tables

Logs

Table A.17: Interactions with Student and Teacher Race (Math, Logs)

	(1)	(2) Elem	(3) Middle	(4)	(5)	(6)
(# White)	.0054*	.0072*	.0039	.0114***	-.0034	-.0016
	(.0022)	(.0028)	(.0037)	(.0028)	(.003)	(.0029)
(# White)*StudentBlack	.0016	.0012	.003	-.004	.0022	.003
	(.0016)	(.0025)	(.0023)	(.0026)	(.0016)	(.0017)
(# White)*StudentHispanic	9.0e-04	2.1e-04	.0029	-.0034	.0012	.002
	(.0019)	(.0028)	(.0029)	(.0033)	(.0019)	(.002)
(# White)*StudentOther	.0039	.0029	.0053	-.0065	.0041*	.0058**
	(.0021)	(.0031)	(.003)	(.0036)	(.0021)	(.0022)
(# White)*TeacherBlack	-.0098	-.0028	-.0194*	-.0087	-.0047	.0022
	(.0064)	(.0085)	(.0095)	(.008)	(.0065)	(.0078)
(# White)*TeacherHispanic	.0355	.0408*	.0554	.0468	.0388	.0236
	(.0204)	(.0203)	(.0475)	(.0286)	(.0208)	(.0182)
(# White)*TeacherOther	-.002	.0133	-.0069	-.0124	.0022	-.0044
	(.0163)	(.0236)	(.0217)	(.0177)	(.0162)	(.0211)
(# Black)	.0071**	.012***	.0029	.0019	.0031	.0043
	(.0025)	(.0031)	(.0044)	(.0032)	(.0027)	(.003)
(#Black)*StudentBlack	5.4e-04	3.9e-04	-.0019	-.0051	-1.9e-04	8.1e-04
	(.0017)	(.0025)	(.0025)	(.0028)	(.0017)	(.0018)
(# Black)*StudentHispanic	.0013	.0017	-.0026	-.0054	9.8e-04	.0016
	(.0019)	(.0028)	(.003)	(.0035)	(.0019)	(.002)
(# Black)*StudentOther	-.0032	.0019	-.0087**	-.0037	-.0034	-.0034
	(.0021)	(.0032)	(.003)	(.0038)	(.0021)	(.0022)
(# Black)*TeacherBlack	.0078	-.0083	.0199*	.0069	.0029	.0014
	(.0063)	(.0094)	(.0094)	(.0072)	(.0064)	(.0085)
(# Black)*TeacherHispanic	-.0273	-.0224	-.0403	-.0445	-.0286	-.0551*
	(.0274)	(.0295)	(.134)	(.029)	(.0281)	(.0226)
(# Black)*TeacherOther	.033	-.0521*	.0814**	.0251	.0308	.0372

	(.0211)	(.0247)	(.0282)	(.0217)	(.0209)	(.0236)
(# Hispanic)	.0059*	.0032	.0097	-.0016	.004	.0062*
	(.0027)	(.0034)	(.0051)	(.0034)	(.0027)	(.0031)
(# Hispanic)*StudentBlack	.0033	7.9e-04	.0036	.0096***	.0032	.0039*
	(.0017)	(.0027)	(.0026)	(.0028)	(.0017)	(.0018)
(# Hispanic)*StudentHispanic	.0029	-.0016	.0064	.0081*	.0027	.0029
	(.0024)	(.0034)	(.0037)	(.0038)	(.0024)	(.0025)
(# Hispanic)*StudentOther	-8.0e-04	-.005	.0025	.0112**	-8.2e-04	-.0016
	(.0026)	(.0039)	(.0039)	(.0042)	(.0026)	(.0027)
(# Hispanic)*TeacherBlack	-.0024	.0043	-.0077	6.8e-04	-.001	-.0101
	(.0076)	(.0111)	(.0112)	(.0086)	(.0076)	(.0093)
(# Hispanic)*TeacherHispanic	-.0018	.0324	-.0219	.0506	-.0025	.0501*
	(.0269)	(.0297)	(.145)	(.0304)	(.0275)	(.0242)
(# Hispanic)*TeacherOther	-.0482**	.0023	-.0856***	-.0279	-.0469**	-.0478**
	(.0164)	(.0151)	(.0225)	(.0196)	(.0163)	(.0181)
(# Other Race)	.0088***	.0055	.0121**	.0061	.0075**	.0121***
	(.0025)	(.0031)	(.0046)	(.0033)	(.0025)	(.0028)
(# Other Race)*StudentBlack	-.0013	3.4e-04	-.0046	.0105***	-.0012	-.0026
	(.0017)	(.0025)	(.0025)	(.0028)	(.0017)	(.0017)
(# Other Race)*StudentHispanic	-.0057*	-.0051	-.0065	.0112**	-.0056*	-.0062**
	(.0023)	(.0032)	(.0036)	(.0038)	(.0023)	(.0024)
(# Other Race)*StudentOther	.0013	-7.6e-04	.0036	.01*	.001	.0014
	(.0024)	(.0036)	(.0035)	(.0041)	(.0024)	(.0025)
(# Other Race)*TeacherBlack	.0069	-.0034	.0179	-.0014	.0061	.0084
	(.0066)	(.0091)	(.0109)	(.0078)	(.0066)	(.007)
(# Other Race)*TeacherHispanic	-.029	-.0306	-.0396	-.0479*	-.0318*	-.0483***
	(.0156)	(.0234)	(.0223)	(.0244)	(.0158)	(.0143)
(# Other Race)*TeacherOther	-.0061	.0232	-.0207	-.0012	-.0103	-.0133
	(.0125)	(.0162)	(.0151)	(.0134)	(.0125)	(.0138)
Student FE	No	No	No	Yes	Yes	Yes
Years Teaching	No	No	No	No	Class	Admin
R^2	.742	.739	.745	.914	.742	.742
N	955876	351858	489987	583228	955876	847224

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Levels

Table A.18: Interactions with Student and Teacher Race (Math, Levels)

	(1)	(2) Elem	(3) Middle	(4)	(5)	(6)
(# White)	.0095**	.0158*	.0086	.0151**	.0063	.0097**
	(.0036)	(.007)	(.0045)	(.0047)	(.0035)	(.0038)
(# White)*StudentBlack	.0026***	9.8e-04	.0022*	.0011	.0024**	.0023**
	(7.8e-04)	(.002)	(9.3e-04)	(.0012)	(7.7e-04)	(7.9e-04)
(# White)*StudentHispanic	3.6e-04	-.0014	8.1e-04	.001	3.3e-04	3.5e-04
	(9.8e-04)	(.0024)	(.0012)	(.0015)	(9.8e-04)	(.001)
(# White)*StudentOther	.0027**	.0012	.0033**	.0024	.0027**	.0031**
	(9.7e-04)	(.0025)	(.0012)	(.0015)	(9.7e-04)	(9.9e-04)
(# White)*TeacherBlack	-.0073	.0147	-.0097	-.0228*	-.0097	-.0082
	(.0077)	(.0195)	(.0094)	(.0094)	(.0075)	(.0084)
(# White)*TeacherHispanic	.0416	.0177	.0669	.0148	.0449	.0341
	(.0411)	(.138)	(.0636)	(.0575)	(.0408)	(.0403)
(# White)*TeacherOther	.0143	.0803	.0169	-.0034	.0082	.0224
	(.0126)	(.0701)	(.0138)	(.021)	(.0131)	(.0141)
(# White)^2	-1.8e-04***	-9.1e-05	-2.2e-04***	-3.6e-04***	1.9e-05	-2.0e-05
	(5.0e-05)	(1.1e-04)	(6.3e-05)	(6.5e-05)	(5.5e-05)	(5.5e-05)
(# White)^2*StudentBlack	-6.6e-05*	1.2e-05	-4.3e-05	-1.6e-05	-5.5e-05*	-5.0e-05
	(2.8e-05)	(7.7e-05)	(3.1e-05)	(4.1e-05)	(2.7e-05)	(2.8e-05)
(# White)^2*StudentHispanic	-1.2e-05	5.3e-05	-1.4e-05	-6.3e-05	-7.3e-06	-5.5e-06
	(3.5e-05)	(9.9e-05)	(4.2e-05)	(4.6e-05)	(3.5e-05)	(3.6e-05)
(# White)^2*StudentOther	-6.9e-05*	-4.3e-05	-8.2e-05*	-9.8e-05*	-6.4e-05*	-7.2e-05*
	(3.2e-05)	(9.6e-05)	(3.8e-05)	(4.4e-05)	(3.2e-05)	(3.3e-05)
(# White)^2*TeacherBlack	-1.1e-04	-.0024	3.0e-05	9.2e-04**	9.3e-05	-3.1e-05
	(2.7e-04)	(.0014)	(2.9e-04)	(3.2e-04)	(2.6e-04)	(3.0e-04)
(# White)^2*TeacherHispanic	.0021	.0164	-.0036	-4.4e-04	.0021	.0025
	(.0027)	(.0371)	(.0045)	(.0037)	(.0026)	(.0028)
(# White)^2*TeacherOther	-2.4e-04	-.018	-2.0e-04	5.9e-04	-1.7e-04	-5.0e-04
	(2.7e-04)	(.01)	(2.9e-04)	(3.8e-04)	(2.7e-04)	(2.9e-04)
(# Black)	.0239***	.0332***	.0205**	1.3e-04	.0136*	.0158**

	(.0055)	(.0098)	(.0073)	(.0076)	(.0054)	(.0059)
(# Black)*StudentBlack	.0018	-3.4e-04	-5.0e-04	.0022	.0016	.0023
	(.0012)	(.0023)	(.0022)	(.0025)	(.0012)	(.0013)
(# Black)*StudentHispanic	9.6e-04	-5.6e-04	8.6e-04	.0036	9.2e-04	.0014
	(.0017)	(.0035)	(.0028)	(.0032)	(.0017)	(.0018)
(# Black)*StudentOther	-.0033	.0059	-.0064**	-.0021	-.0031	-.0031
	(.0019)	(.0039)	(.0024)	(.0029)	(.0019)	(.0019)
(# Black)*TeacherBlack	-.0092	-.0347*	-.0011	-.0028	-.0052	-2.2e-04
	(.0087)	(.0163)	(.0106)	(.0103)	(.0085)	(.0093)
(# Black)*TeacherHispanic	-.0171	.0187	-.404	-.231	-.0288	.004
	(.0767)	(.0824)	(.442)	(.161)	(.0774)	(.0707)
(# Black)*TeacherOther	.0206	-.0608	.0513**	.0589*	.0176	.0164
	(.0144)	(.0444)	(.0188)	(.0271)	(.0151)	(.0149)
(# Black)^2	-.0011***	-8.9e-04	-.001**	-4.0e-04	-2.0e-04	-2.6e-04
	(2.9e-04)	(6.1e-04)	(3.6e-04)	(3.2e-04)	(2.7e-04)	(2.9e-04)
(# Black)^2*StudentBlack	-1.4e-04*	-2.8e-05	-5.0e-05	-2.4e-04	-1.3e-04*	-1.6e-04*
	(6.4e-05)	(5.9e-05)	(1.4e-04)	(1.6e-04)	(6.3e-05)	(6.5e-05)
(# Black)^2*StudenHispanic	-8.0e-05	-6.4e-05	-3.9e-05	-3.6e-04	-7.9e-05	-9.6e-05
	(9.8e-05)	(1.2e-04)	(1.8e-04)	(2.0e-04)	(9.7e-05)	(9.9e-05)
(# Black)^2*StudentOther	1.7e-04	-3.2e-04	3.1e-04*	2.3e-05	1.6e-04	1.6e-04
	(1.1e-04)	(1.9e-04)	(1.4e-04)	(1.5e-04)	(1.1e-04)	(1.1e-04)
(# Black)^2*TeacherBlack	9.0e-04	.0018	5.6e-04	1.7e-04	5.1e-04	4.2e-04
	(4.8e-04)	(.0011)	(5.2e-04)	(5.7e-04)	(4.5e-04)	(4.8e-04)
(# Black)^2*TeacherHispanic	.0052	-6.6e-04	.176	.0248	.0066	.0069
	(.0088)	(.0097)	(.1)	(.0297)	(.0089)	(.0087)
(# Black)^2*TeacherOther	4.9e-04	.0019	-3.1e-04	-.0038	6.9e-04	6.7e-04
	(7.4e-04)	(.0042)	(8.8e-04)	(.0019)	(7.6e-04)	(7.5e-04)
(# Hispanic)	.0248*	.009	.0466**	.013	.0082	.0193
	(.0102)	(.0173)	(.0143)	(.015)	(.0099)	(.0106)
(# Hispanic)*StudentBlack	.005	.0014	.0012	.024**	.0046	.0066
	(.0053)	(.0113)	(.0067)	(.0073)	(.0052)	(.0053)
(# Hispanic)*StudentHispanic	.0047	.0021	.0062	.0203*	.0046	.007
	(.0059)	(.0118)	(.008)	(.0085)	(.0058)	(.0061)
(# Hispanic)*StudentOther	-5.1e-04	-.0198	.0043	.0356***	-.001	-9.2e-04

	(.0073)	(.0156)	(.009)	(.0106)	(.0072)	(.0075)
(# Hispanic)*TeacherBlack	-.0185	.0023	-.0516	-.0438	-.0042	-.0257
	(.026)	(.0541)	(.0329)	(.0304)	(.0255)	(.0285)
(# Hispanic)*TeacherHispanic	-.0091	.133	.631	.34	-.0753	-.186
	(.172)	(.453)	(.758)	(.283)	(.171)	(.169)
(# Hispanic)*TeacherOther	-.197**	.0313	-.347**	-.151	-.187**	-.22**
	(.0726)	(.114)	(.106)	(.102)	(.0715)	(.074)
(# Hispanic)^2	-.0023	-.0036	-.003	-3.5e-04	4.4e-04	5.9e-05
	(.0014)	(.0039)	(.0017)	(.002)	(.0013)	(.0014)
(# Hispanic)^2*StudentBlack	-7.6e-04	-.0013	-1.9e-04	-.0027*	-8.4e-04	-.0011
	(.0011)	(.003)	(.0012)	(.0012)	(.001)	(.001)
(# Hispanic)^2*StudentHispanic	-3.2e-04	3.8e-04	-6.2e-04	-.0027*	-4.8e-04	-9.5e-04
	(.0011)	(.0028)	(.0014)	(.0013)	(.0011)	(.0011)
(# Hispanic)^2*StudentOther	-1.5e-04	-.0013	-2.0e-04	-.0042*	-2.7e-04	-3.7e-04
	(.0014)	(.0035)	(.0016)	(.0017)	(.0013)	(.0014)
(# Hispanic)^2*TeacherBlack	.0061	.002	.0068	.0089**	.0046	.006
	(.0032)	(.013)	(.0037)	(.0033)	(.0031)	(.0033)
(# Hispanic)^2*TeacherHispanic	-.111	-.129	-.676	-.124	-.0673	.0218
	(.0792)	(.594)	(.392)	(.132)	(.0788)	(.0981)
(# Hispanic)^2*TeacherOther	.0333*	.0106	.0562**	.037	.0378*	.0415**
	(.0151)	(.042)	(.0209)	(.0198)	(.0151)	(.0157)
(# Other)	.0209***	.0195	.028***	.0551***	.0186***	.0248***
	(.0052)	(.0108)	(.0083)	(.0093)	(.0051)	(.0055)
(# Other)*StudentBlack	-.0017	.0086	-.0082*	-2.7e-04	-.0027	-.0041
	(.0025)	(.0045)	(.0038)	(.0041)	(.0025)	(.0026)
(# Other)*StudentHispanic	-.0071*	-3.4e-04	-.0192***	6.3e-04	-.008*	-.0088*
	(.0034)	(.0061)	(.0054)	(.0053)	(.0034)	(.0036)
(# Other)*StudentOther	-1.8e-04	.0061	-6.4e-04	.0017	-.0013	-.0014
	(.003)	(.0057)	(.0043)	(.0047)	(.003)	(.0031)
(# Other)*TeacherBlack	.0227	-.0179	.0972**	.0694***	.0192	.0251
	(.0185)	(.0288)	(.0364)	(.0196)	(.0183)	(.0194)
(# Other)*TeacherHispanic	-.0018	.0499	.0288	.0348	-.0497	-.142
	(.0666)	(.131)	(.135)	(.0863)	(.0642)	(.0847)
(# Other)*TeacherOther	-.007	-.0078	-.0123	-.0419*	-.0088	-.0072

	(.0125)	(.0179)	(.0219)	(.0195)	(.0122)	(.013)
(# Other)^2	-9.3e-04*	-6.2e-04	-.0015**	-.0013**	-5.6e-04	-8.1e-04
	(4.3e-04)	(.0016)	(5.6e-04)	(4.4e-04)	(4.0e-04)	(4.3e-04)
(# Other)^2*StudentBlack	-1.5e-04	-4.9e-04*	4.1e-04	-1.1e-04	-1.0e-04	-4.9e-05
	(1.4e-04)	(2.4e-04)	(3.1e-04)	(2.1e-04)	(1.4e-04)	(1.4e-04)
(# Other)^2*StudentHispanic	2.8e-04	2.6e-04	.0014***	-6.2e-05	3.2e-04	3.5e-04
	(1.9e-04)	(4.3e-04)	(3.9e-04)	(2.2e-04)	(1.9e-04)	(2.0e-04)
(# Other)^2*StudentOther	-1.4e-04	-4.5e-04	3.2e-05	-2.5e-04	-9.4e-05	-6.9e-05
	(1.3e-04)	(2.3e-04)	(2.6e-04)	(2.1e-04)	(1.3e-04)	(1.3e-04)
(# Other)^2*TeacherBlack	-.0034	.0053	-.0216*	-.0181***	-.0015	-.0024
	(.0042)	(.0064)	(.0086)	(.0042)	(.0042)	(.0044)
(# Other)^2*TeacherHispanic	-.0092	-.0209	-.0321	-.003	.0036	.0376
	(.0207)	(.0324)	(.0485)	(.0289)	(.0196)	(.0238)
(# Other)^2*TeacherOther	6.6e-04	7.3e-04	7.4e-04	.0013*	4.3e-04	5.4e-04
	(5.4e-04)	(.0017)	(8.6e-04)	(6.5e-04)	(5.2e-04)	(5.5e-04)
Student FE	No	No	No	Yes	Yes	Yes
Years Teaching	No	No	No	No	Class	Admin
R^2	.741	.739	.745	.914	.742	.742
N	955876	351858	489987	583228	955876	847224

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$