RACIAL INTERACTION EFFECTS AND STUDENT ACHIEVEMENT

Jeffrey Penney
Department of Economics
Pontificia Universidad Javeriana
S.J., Bogotá D.C., Colombia
dr.jeffrey.penney@gmail.com

Abstract

Previous research has found that students who are of the same race as their teacher tend to perform better academically. This paper examines the possibility that both dosage and timing matter for these racial complementarities. Using a model of education production that explicitly accounts for past observable inputs, a conditional differences-in-differences estimation procedure is used to nonparametrically identify dynamic treatment effects of various sequences of interventions. Applying the methodology to Tennessee's Project STAR class size experiment, I find that racial complementarities may vary considerably according to the treatment path. Early exposures to same-race teachers yield benefits that persist in the medium run. This same-race matching effect may explain a nontrivial portion of the black–white test score gap.

doi:10.1162/EDFP_a_00202

1. INTRODUCTION

There is a conventional wisdom among educators that when a minority student is paired with a teacher of the same racial or ethnic background, he or she is more likely to excel educationally (Dee 2004). There are a number of theories that have been put forth as the reason for these racial complementarities, perhaps the most popular of which is that minority teachers can serve as important role models for minority students; another is the idea of "cultural synchronicity" that is hypothesized to occur between minority students and teachers who share the same cultural background (Ingersoll and May 2011). These arguments, among others, have been advanced to encourage the recruitment of minority teachers.

Racial complementarities in educational achievement have much empirical support. There is a positive correlation between academic achievement and racial match for both black and white students on concrete measures of performance such as test scores (Hanushek et al. 2005; Egalite, Kisida, and Winters 2015). Using data from Project STAR (Student/Teacher Achievement Ratio) and examining each gender–race combination separately, Dee (2004) finds that the contemporaneous test score benefits are of the order of approximately 4 percentile points in all subjects—with the exception of white girls in reading, where the gain is smaller but not statistically significant. On average, the gains are almost as large as those obtained through small class size interventions that are found with the same data reported in Krueger (1999). The possibility of favoritism in grading for students who share the teacher's racial or ethnic background as the source of these test score gains is a concern, because there is evidence that teachers evaluate students of the same race more favorably using subjective assessments of academic performance (Ouazad 2014). Nevertheless, the complementarities are also present when the evaluations are externally administered, which effectively rules out this possibility. Clotfelter, Ladd, and Vigdor (2010) demonstrate the importance of accounting for racial interactions in education production by illustrating their effect on race coefficients when same-race dummies are excluded from the regression equation.

Cost–benefit exercises of any policy should consider not just the immediate effects but also those in the medium- and long-run time horizons. Moreover, the interventions themselves can exhibit substantial differences in terms of timing and dosage, both of which may matter for the outcomes. These considerations suggest that ignoring the cumulative and dynamic nature of education production may lead to incorrect inferences. Investigating the effect of small class sizes using data from Project STAR, Krueger (1999) finds that attending a small class yields contemporaneous test score benefits in kindergarten through third grade. On the other hand, estimating a dynamic model of education production that takes into account the full history of observable inputs, Ding and Lehrer (2010) conclude that statistically significant achievement gains from the small class intervention in Project STAR are present only in kindergarten and first grade; they further find that the intervention may actually *reduce* achievement in third grade for those who were not in small classes from kindergarten through second grade.[1]

1. Ding and Lehrer (2010) also pay important attention to the selective attrition and noncompliance issues of Project STAR.

These results are consistent with that of Hanushek (1999), who finds an erosion of the small class effect in later grades.[2]

In this paper, I extend the work of Dee (2004) by conducting an analysis using an approach similar to that of the one developed in Ding and Lehrer (2010). I investigate the effect of racial interactions on student achievement using a dynamic model of education production that takes into account the full history of observable inputs. The specification allows for both the timing and the dosage to matter for the outcome at each grade—for example, being instructed by a teacher of the same race in kindergarten and first grade may have a different effect on third-grade math test scores than having the same-race teacher in the first and second grades. To estimate the model, I use a conditional differences-in-differences procedure to nonparametrically identify dynamic treatment effects of exposure(s) to a same-race teacher in the short and medium run. The effect of any particular treatment path can be estimated. The regression model can be thought of as a value-added model that allows for nonuniform decay of past inputs.

This study makes use of data from Project STAR, a highly influential education experiment that took place in Tennessee in the 1980s that sought to determine the effect of class size on student achievement. I use data from a cohort that participated in the experiment from kindergarten until the end of third grade.

The main findings of the empirical analysis are as follows. I find that both the timing and the dosage of being assigned to a teacher of the same race can matter for test score gains. The contemporaneous benefits are strongest in the early grades. The estimated dynamic treatment effects show that the benefits persist in the medium run, with early grade exposure to same-race teachers having statistically significant benefits to scores in later grades. I examine whether the findings are merely an artifact of across-school sorting of teachers by race. Repeating the analysis using classroom fixed effects in order to expunge any possible bias arising from within-school quality differences between black and white teachers, I find no substantive changes in the results. The main results were qualitatively, and in many cases quantitatively, similar when the analysis was repeated using various subgroups of students by (1) those who complied with their treatment assignment, (2) size of school, (3) race, (4) gender, and (5) socioeconomic status.

To conclude the article, I discuss the policy implications of the empirical findings. There are economically significant gains in achievement which are moderate in magnitude when students are taught by teachers of the same race, ranging from approximately 4 to 10 percentile points on third grade test scores for continuous treatment from kindergarten through third grade. The existence of effects that persist past the short run and the economic significance of the effects indicate that future research should investigate the channels through which they occur. Once the channels are identified, policy prescriptions relating to within-school racial sorting may be found to be desirable. In kindergarten, the own-race teacher effect on achievement explains approximately 14 percent and 22 percent of the black-white reading and math test score gaps,

2. Chetty et al. (2011), however, find that improvements in other outcomes, such as higher rates of college attendance, do occur as a result of smaller class sizes, even though their test score benefits fade over time.

respectively, because minorities are far less likely to be matched with a teacher of the same race.

This paper is organized as follows. Section 2 details the theory and estimation of the econometric model. I outline the data and perform some initial analyses in section 3. The empirical exercise and robustness checks are performed in section 4. I conclude the main body of the paper with a discussion of policy implications in section 5. In the online Appendix, which is available on the *Education Finance and Policy*'s Web site at www.mitpressjournals.org/doi/suppl/10.1162/EDFP_a_00202, I examine issues relating to the validity of the experiment as well as other technical concerns.

2. MODEL

The primary purpose of this paper is to derive estimates of the effects of racial matches between students and teachers on academic achievement for both the short and medium term. To this end, I use an approach similar to that of Ding and Lehrer (2010), wherein the estimated parameters from a system of equations (one equation for each school grade) are used to calculate the dynamic effects of various sequences of interventions. In order for it to be possible to obtain these estimates, the usual analysis of education production is augmented by explicitly including past observable inputs and same-race dummies into the model. I begin this section by detailing the system of equations, then continue by describing the procedure through which the dynamic effects of own-race teachers are obtained and how they are interpreted.

Theory and Estimation

Define A_{ig} as the achievement of student i in grade g, and let grade $g = k$ denote kindergarten. Let X be a matrix of control variables and a constant term, and define d to be a dummy for the same-race intervention where $d = 1$ if the student has a same-race teacher and $d = 0$ otherwise. Student fixed effects are given by υ. The α vectors denote the estimated effects of the controls, and the β coefficients the effects of the treatments. For a given coefficient γ_{lm} where $\gamma = \{\alpha, \beta\}$, l denotes the level of achievement that is affected by the input, and m is the time period of the input; for example, β_{31} is the estimated effect of the same-race treatment in first grade on third grade academic achievement. The system of equations to be estimated is as follows:

$$A_{ik} = \upsilon_i + X_{ik}\alpha_{kk} + \beta_{kk}d_{ik} + \varepsilon_{ik}, \tag{1}$$

$$A_{i1} = \upsilon_i + X_{i1}\alpha_{11} + X_{ik}\alpha_{1k} + \beta_{11}d_{i1} + \beta_{1k}d_{ik} + \varepsilon_{i1}, \tag{2}$$

$$A_{i2} = \upsilon_i + X_{i2}\alpha_{22} + X_{i1}\alpha_{21} + X_{ik}\alpha_{2k} + \beta_{22}d_{i2} + \beta_{21}d_{i1} + \beta_{2k}d_{ik} + \varepsilon_{i2}, \tag{3}$$

$$A_{i3} = \upsilon_i + X_{i3}\alpha_{33} + X_{i2}\alpha_{32} + X_{i1}\alpha_{31} + X_{ik}\alpha_{3k} + \beta_{33}d_{i3} + \beta_{32}d_{i2} + \beta_{31}d_{i1} + \beta_{3k}d_{ik} + \varepsilon_{i3}, \tag{4}$$

where ε is the usual error term. This formulation allows for the effect of inputs to vary over time: For example, the effect of having a same-race teacher in kindergarten on contemporaneous achievement can be different in kindergarten compared to first grade (that is, I allow the possibility that $\beta_{kk} \neq \beta_{1k}$). The inclusion of past inputs also serves as a prophylactic against omitted variable bias. To illustrate, consider the equation for

A_{i1}: If past assignment to a same-race teacher d_{ik} is correlated with current assignment to a same-race teacher d_{i1}, and past assignment affects contemporaneous achievement $\beta_{1k} \neq 0$, then omitting past assignment d_{ik} as a regressor in the equation for A_{i1} will cause the estimate of the effect of current assignment on current achievement β_{11} to be biased and inconsistent.

I make the following assumptions to permit inference on the regression results.[3] The fixed effect v_i can be correlated with the observed and unobserved determinants of achievement; it contains the effect of not only student ability but also other time-invariant inputs and characteristics. Other unobservable inputs into the education production function are assumed to be either fixed over the course of the sample (and are thus absorbed by the fixed effects) or uncorrelated with the included inputs.[4] I assume no pretreatment effects—that is, treatment assignment in future periods does not affect current achievement. The matrix of controls X contains the following: teacher characteristics (teacher's race, years of experience, and whether the teacher has a graduate degree), the type of class the student attends (a small class, a regular class, or a regular class with a full-time teacher's aide), school fixed effects that I allow to vary by grade, and free lunch status. Given these assumptions, any differential effect of changes in treatment assignment will reveal themselves in the same-race coefficient vectors $\{\beta_k, \beta_1, \beta_2, \beta_3\}$ where $\beta_1 = [\beta_{11} \quad \beta_{1k}]$, and so forth.[5]

There remains the issue of the fixed effect v_i, which is unobservable. If it is correlated with the included inputs but excluded from the regression equation, estimation of the system of equations 1–4 is biased and inconsistent. Under the assumptions outlined above, the regression coefficients can be consistently estimated by first-differencing the system of equations—such a transformation will eliminate the individual fixed effects. Miquel (2003) and Lechner and Miquel (2010) demonstrate that this conditional differences-in-differences approach can be used to nonparametrically identify the causal effects of sequences of interventions. The equation for achievement in first grade is:

$$A_{i1} - A_{ik} = X_{i1}\alpha_{11} + X_{ik}(\alpha_{1k} - \alpha_{kk}) + \beta_{11}d_{i1} + (\beta_{1k} - \beta_{kk})d_{ik} + \varepsilon^*_{i1}, \tag{5}$$

where $\varepsilon^*_{i1} = \varepsilon_{i1} - \varepsilon_{ik}$. Note that the kindergarten equation remains unchanged in this transformation: Although the fixed effect is still present, random assignment in this grade ensures that the fixed effect is not correlated with the included covariates. Because of potentially nonrandom attrition in the following grades, however, we require that the fixed effect be differenced out in the other achievement equations.

Note that the differencing procedure, under the assumptions above, is an identification strategy to obtain unbiased and consistent estimates of the system of equations 1–4. This can be thought of as analogous to a fixed effects procedure—although some

3. A discussion about the particulars regarding identification can be found in the online appendix.
4. Using a restricted version of the method developed in Ding and Lehrer (2014), Ding and Lehrer (2010) find evidence that the effect of unobserved inputs on achievement in the STAR data for second and third grades is relatively constant.
5. Some scholars believe that dynamic complementarities in inputs ought to be modeled in education production functions. Such an approach is not required here because random assignment (see section 3) guarantees that such effects, if they exist, would not bias the coefficients of interest. For the same reason, peer effects are not explicitly modeled (although they are taken into account in classroom fixed effect models; see section 4).

of the variables are transformed in order to perform the estimation, the interpretation of the coefficients does not change as a result.

Because the differenced system of equations is triangular, it can be estimated using equation-by-equation ordinary least squares to obtain the coefficient estimates. Moreover, no assumptions are necessary as to the distribution of the error terms. As the parameters enter recursively into the equations, one is required to estimate them in a sequential fashion (starting with kindergarten) because the desire is to separately identify the coefficients of interest. For example, we require the estimates of α_{kk} and β_{kk} from equation 1 to enter into equation 5 in order to obtain the estimates of α_{11} and β_{11}.[6]

This specification can be thought of as a value-added model.[7] Using the language of Rothstein (2010), the model estimated here is most similar to the VAM2 specification (value-added model with a lagged achievement variable as a regressor), which implicitly includes the effect of past inputs by including a lagged term in achievement. Including lagged achievement as a regressor imposes an assumption of constant decay—that is, past inputs, both observed and unobserved, are all assumed to decay at a constant rate. The model used here relaxes the constant decay assumption for the observed inputs but at the cost of assuming that past *time-variant* unobservables are uncorrelated with future observables.

Dynamic Treatment Effects

I now describe the procedure to produce the estimates of the dynamic effects of own-race teachers on student achievement. In this paper, they are dynamic average treatment on the treated (DATT) estimates—that is, the net effect of the treatment *sequence* compared with some other sequence for those who have experienced that treatment path.[8] Denote $t(a, b)$ to be the treatment sequence of an individual where a is the treatment experienced in the first period and b is the treatment received in the second. For $i = \{a, b\}$, let $i = 1$ if treatment was received and $i = 0$ otherwise. Then, a person experiencing treatment in both periods would be denoted as receiving the treatment sequence $t(1, 1)$, a person being treated only in the second period but not in the first is denoted as experiencing the sequence $t(0, 1)$, and so forth. Using this notation, I can define the dynamic treatment effects of interest.

For purposes of exposition, I consider the case of two periods. Let $\tau(a, b)(w, x)$ be the DATT of the treatment sequence $t(a, b)$ with the counterfactual sequence $t(w, x)$—put simply, the net benefit (or cost, if the estimate is negative) of experiencing $t(a, b)$ rather than $t(w, x)$, but the estimate of this effect is only for those who have experienced

6. Even if simultaneous estimation of the system were possible, estimates of the fixed effect would still be inconsistent because the number of observations for each fixed effect is limited to four by construction. In a first-differenced approach, the treatment effects are consistent and unbiased.
7. It is important to note that although this specification can be thought of as a value-added model, this is only an observation and not an indication that value-added assumptions in this paper are being used to identify the causal effects of interest for the sequences of interventions.
8. Ding and Lehrer (2010) instead use the terminology dynamic treatment effects for treated (DTET) instead of DATT used here. These both refer to the same thing; the latter terminology is used because I feel that the acronym solidly connects it to the frequently used average treatment effects on the treated, which is commonly shortened as ATT.

$t(a, b)$.[9] For example, $\tau(1, 1)(0, 0)$ refers to the DATT of having an own-race teacher in kindergarten and first grade compared with not having same-race teachers in both grades for those who had teachers of the same race in both grades, and $\tau(1, 0)(0, 0)$ describes the effect on achievement of an exposure to a teacher of the same race in kindergarten compared with never having had a same-race teacher for those who have only had a same-race teacher in kindergarten. Using the estimated parameters from equation 5, the DATT for the two examples would be calculated as follows:

$$\tau\,(1,\,1)\,(0,\,0) = \hat{\beta}_{11} + \hat{\beta}_{1k},$$

$$\tau\,(1,\,0)\,(0,\,0) = \hat{\beta}_{1k}.$$

The standard errors of these effects are calculated using the standard formula for sums of random variables.[10] The same logic extends to more than two periods.

3. DATA

Description

The data used in this study come from a cohort of students who participated in Project STAR, an experiment that took place in Tennessee and ran from 1985 until 1989. The experiment was legislated into existence and funded by the state government[11] at a cost of approximately $12 million over five years—this figure includes the data analysis and reporting that took place in the fifth year. The primary goal of the STAR experiment, as its acronym implies, was to determine the effect of class size on student achievement in primary education (Finn et al. 2007). Across the state, 79 schools signed up for the experiment and had to commit to participation for four years. Data were also gathered from nonparticipating schools to use as a benchmark. To qualify for participation in Project STAR, schools required enough students to support at least three different classes per grade. Students and teachers were randomly assigned within schools to one of three class types: a small class (13 to 17 students), a regular class (22 to 25 students), or a regular class with a full-time teacher's aide. Regular classes in first through third grade still had a part-time teacher's aide available to assist the class for approximately 25 percent to 33 percent of the time, on average. It was initially intended that students stay in their assigned class type from kindergarten through third grade, although after kindergarten students in regular or regular with aide classes were randomly permanently reassigned between these two class types. An examination of 1,581 students enrolled in kindergarten found that compliance was almost perfect (Krueger 1999). In first grade and beyond, however, there were some problems with noncompliance, with a number of students switching in or out of small classes. Noncompliance was primarily due to parental complaints or discipline problems (Krueger 1999). At the end of each year, all participating students were given a battery of academic and nonacademic tests. More detailed overviews of Project STAR can be found in Krueger (1999) and Finn et al. (2007).

9. This condition of the estimate only referring to those who experienced the sequence $t(a, b)$ is necessary because treatment effect on the treated estimates is being obtained. Additional assumptions are required in order to interpret them simply as dynamic average treatment effects; I do not make these assumptions here.

10. For example, the standard error of $\tau(1, 1)(0, 0) = \hat{\beta}_{11} + \hat{\beta}_{1k}$ is equal to $\sqrt{\text{var}(\hat{\beta}_{11}) + \text{var}(\hat{\beta}_{1k}) + 2\text{cov}(\hat{\beta}_{11}, \hat{\beta}_{1k})}$.

11. See Word et al. (1990).

Table 1. Proportion of Students with a Teacher of the Same Race

	Kindergarten	First Grade	Second Grade	Third Grade
White students	0.9414	0.9588	0.9213	0.9501
Black students	0.4023	0.4454	0.4480	0.5036

Notes: Numbers calculated from sample data. The table shows the proportion of students of a given race who had a teacher of the same race for the listed grade in the Project STAR cohort.

In this paper, the measures of student achievement examined are obtained from the seventh edition Stanford Achievement Test scores in mathematics, reading, and word recognition. The tests were designed so that the scores were comparable across grades (Finn et al. 2007)—that is, students effectively took the same tests in each subject.[12] I elect to use the natural scaled scores in this analysis in order to avoid potential pitfalls associated with some transformations of the test score data. Cascio and Staiger (2012) show that the use of normalized scores mechanically cause the estimated impacts of interventions to appear to fade over time.[13] Percentile scores are typically used when the scaled test scores across several grades are not directly comparable, which is not the case here. Nonetheless, the results of this paper are qualitatively, and in most cases quantitatively, similar (in terms of precision of the estimates and relative magnitude of the coefficients) when examined in percentile and normalized form, which should assuage concerns raised in Bond and Lang (2013) regarding the ordinality of test score variables and its effects on inference.

I follow the STAR cohort of students who entered the program in 1985, excluding students who joined after kindergarten. This is done to more credibly estimate the full sequence of dynamic effects (Ding and Lehrer 2010). I only keep students whose race is either black or white, which results in a loss of 33 students from the sample (under 1 percent). Dropping these students does not affect the results.[14]

Summary Statistics

The Project STAR cohort of students is highly segregated according to school—only about one in five of any particular Project STAR grade has a racial balance that lies between 20 percent and 80 percent of students being of a single race.[15] Moreover, most teachers these cohorts encounter that have predominantly white student bodies are themselves white, whereas teachers who teach Project STAR cohorts with majority black student bodies have a more even racial distribution. The proportions of students for each grade who are taught by a teacher of the same race are displayed in table 1.

12. There is considerable overlap in the test scores across grades. For example, the top kindergarten students performed similarly to the median third grade students in mathematics.
13. Cascio and Staiger (2012) show that these results stem from the increasing variance in accumulated knowledge as students move through school. There is no such pattern of increasing variance in the scaled test scores in the data used here.
14. There are twelve teachers in the sample (less than 1 percent of the teacher pool) who are neither black nor white, and they all teach third grade. Excluding them from the analysis does not change the results.
15. Unfortunately, information is not available on the racial composition at the level of the entire school.

Table 2. Transition Tree

Kindergarten	First Grade	Second Grade	Third Grade
			$t(1, 1, 1, 1) = 1{,}946$
			$t(1, 1, 1, 0) = 117$
		$t(1, 1, 1) = 2, 290$	$t(1, 1, 1, \cdot) = 227$
		$t(1, 1, 0) = 252$	
		$t(1, 1, \cdot) = 613$	$t(1, 1, 0, 1) = 176$
			$t(1, 1, 0, 0) = 44$
	$t(1, 1) = 3, 155$		$t(1, 1, 0, \cdot) = 32$
	$t(1, 0) = 358$		
	$t(1, \cdot) = 1, 301$		$t(1, 0, 1, 1) = 101$
			$t(1, 0, 1, 0) = 43$
		$t(1, 0, 1) = 182$	$t(1, 0, 1, \cdot) = 38$
		$t(1, 0, 0) = 64$	
		$t(1, 0, \cdot) = 112$	$t(1, 0, 0, 1) = 18$
			$t(1, 0, 0, 0) = 31$
			$t(1, 0, 0, \cdot) = 15$
$t(1) = 4, 814$			
$t(0) = 1, 435$			
			$t(0, 1, 1, 1) = 101$
			$t(0, 1, 1, 0) = 34$
		$t(0, 1, 1) = 164$	$t(0, 1, 1, \cdot) = 29$
		$t(0, 1, 0) = 164$	
		$t(0, 1, \cdot) = 108$	$t(0, 1, 0, 1) = 76$
			$t(0, 1, 0, 0) = 44$
	$t(0, 1) = 436$		$t(0, 1, 0, \cdot) = 44$
	$t(0, 0) = 502$		
	$t(0, \cdot) = 497$		$t(0, 0, 1, 1) = 48$
			$t(0, 0, 1, 0) = 32$
		$t(0, 0, 1) = 115$	$t(0, 0, 1, \cdot) = 35$
		$t(0, 0, 0) = 233$	
		$t(0, 0, \cdot) = 154$	$t(0, 0, 0, 1) = 52$
			$t(0, 0, 0, 0) = 136$
			$t(0, 0, 0, \cdot) = 45$

Notes: The number of students that experience each treatment path are given after the equal sign. A downward move corresponds to $d_{ig} = 0$ in the previous period, and an upward move signifies $d_{ig} = 1$ in the previous period. A floating dot symbol $\cdot$ denotes attrition in period g. For example, 108 children had the sequence $t(0, 1)$ then attrited in the second grade, and 43 children have undergone the treatment sequence $t(1, 0, 1, 0)$.

The transitions that students experience are displayed in table 2. We see that the vast majority of students have a teacher of the same race throughout the grades, and that other treatment paths have less support. This means that the standard errors produced in the estimation process will be conservative in terms of inference by favoring the null hypothesis, ceteris paribus.

An initial look at the relationship between having an own-race teacher and test score performance is presented in table 3. The average test score is never higher for white students with black teachers in any of the twelve grade–subject pairs; for black students, it is higher in two of the twelve categories if they have a white teacher (in the third grade). To see whether these findings may hold across the distribution, I perform

Table 3. Average Test Scores by Race and Racial Match

	White Students		Black Students	
	Teacher is Same Race	Different Race	Teacher is Same Race	Different Race
Kindergarten				
Mathematics	491.79	487.99	481.51	468.49
Reading	440.77	437.36	432.74	426.49
Word recognition	439.07	433.77	429.21	422.89
Observations	3,655	183	725	1,131
First grade				
Mathematics	545.52	525.20	521.39	511.32
Reading	540.56	518.95	503.81	499.43
Word recognition	529.34	513.80	502.61	498.70
Observations	2,446	105	505	608
Second grade				
Mathematics	596.64	593.09	569.64	565.89
Reading	602.56	595.43	570.85	566.93
Word recognition	601.72	599.65	576.50	567.31
Observations	2,277	199	430	519
Third grade				
Mathematics	633.04	620.39	607.63	609.35
Reading	630.23	624.09	608.54	608.09
Word recognition	627.80	620.17	602.88	604.01
Observations	2,096	110	404	362

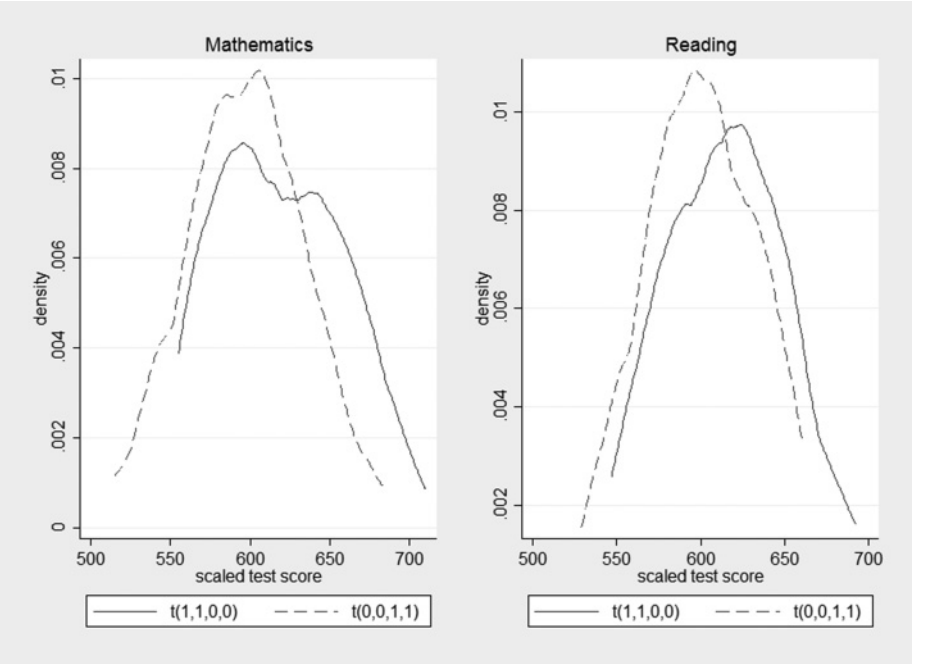
Notes: Numbers calculated from sample data. The table displays the average scaled Stanford Achievement Test scores by subject, race, racial match, and grade.

Kolmogorov-Smirnov tests on the kindergarten test scores. I find that students in classes with a teacher of the same race have higher mathematics, reading, and word recognition test scores compared with those who do not ($p = 0.000$ for all tests).

One of the primary purposes of this paper is to determine whether the timing of the same-race teacher treatment matters. To illustrate the potential importance of this, in figure 1 I plot the density of the third-grade math and reading test scores for two different treatment paths: one in which the student experienced two early treatments (the treatment path $t(1, 1, 0, 0)$), and the other in which the treatments arrived later (the treatment path $t(0, 0, 1, 1)$). In both subjects, students who were exposed to teachers of the same race earlier have a distribution of test scores that lies to the right of the distribution that corresponds to those who were exposed later, despite both groups of students experiencing the same number of same-race treatments.[16]

This preliminary analysis allows us to come to several substantive conclusions: There may be reason to believe that own-race teachers increase student achievement and, should this be true, it may be that this can explain part of the black–white student test score gap because white students are far more likely to be paired with a teacher of the same race compared with black students. Moreover, the timing of the treatments may matter for academic outcomes.

16. The results are also similar for word recognition test scores.



Notes: Kernel density estimates of the probability density functions for two different treatment paths calculated from sample data using the Epanechnikov kernel with optimal bandwidth. The kernel densities are evaluated at 50 points. See section 2 for notational details.

Figure 1. Distribution of Third Grade Test Scores by Treatment Path.

4. EMPIRICAL ANALYSIS

Results

Table 4 presents the estimates on the coefficients of the d_{ig} variables obtained by estimating the system of equations described in section 2.[17] I denote the estimated coefficients from the table as structural because they (and their covariance matrix) are required to calculate the DATT estimates.

Taken in isolation, the estimated parameters from the system of equations correspond to dynamic average treatment on the treated estimates for *single* exposures. For example, the estimate of the coefficient on d_{ik} in grade 3 is the estimate of $\tau(1,0,0,0)(0,0,0,0)$, which is the estimated DATT for a student who has an own-race teacher in kindergarten but never again for those who have only had an own-race teacher in kindergarten. Examining these results we see that, for a single intervention, early exposure generally benefits children more than late exposure. There appear to be precisely estimated positive effects up until second grade for the case of mathematics. The effect of the same-race teacher treatment can be persistent in the medium run: The benefit from kindergarten for a single exposure is statistically significant in all grade–subject combinations.

17. Note that these estimates assume the effect of racial interactions is the same across races and school-grade racial compositions; I examine these considerations in the next section.

Table 4. Structural Coefficient Estimates

	Mathematics	Reading	Word Recognition
Kindergarten			
β_{kk}	11.40**	5.08**	5.10**
	(2.52)	(1.63)	(1.87)
First grade			
β_{1k}	4.39	8.78**	9.09**
	(2.58)	(2.57)	(3.18)
β_{11}	12.00**	3.79	3.22
	(2.68)	(2.63)	(3.18)
Second grade			
β_{2k}	6.33**	4.94	4.55
	(2.42)	(2.65)	(3.57)
β_{21}	−0.17	−1.39	1.71
	(2.95)	(3.04)	(4.03)
β_{22}	6.24**	3.98	1.31
	(2.77)	(2.74)	(3.19)
Third grade			
β_{3k}	5.65	11.26**	12.76**
	(2.99)	(2.36)	(3.97)
β_{31}	1.53	2.05	7.03
	(2.81)	(2.45)	(3.91)
β_{32}	5.14	1.79	−0.19
	(2.99)	(2.74)	(3.29)
β_{33}	−4.53	−2.38	0.81
	(2.45)	(2.51)	(3.60)

Notes: The table contains the structural coefficient estimates of an own-race teacher on the d_{ig} variables in the system of equations described in section 3 that are to be used in the calculation of the DATT; see the text for details. Standard errors clustered at the level of the classroom are given in parentheses. Scaled test scores are used as the response variable. Observations are weighted using inverse probability weights; see section A.3 of the online appendix.

**Statistical significance at the 1% level.

The case of having an own-race teacher in multiple grades is displayed in table 5. Multiple exposures are shown to be beneficial in many cases. Nonetheless, although the number of doses matters, so does their timing: Examining $\tau(1, 1, 0, 0)(0, 0, 0, 0)$ and $\tau(0, 0, 1, 1)(0, 0, 0, 0)$ in third grade, we see that the former sequence of treatments gives far more of a benefit than the latter in all subjects, even though both sequences give two exposures to a teacher of the same race. The difference between these treatment paths is statistically significant at the 5 percent level for mathematics and at the 1 percent level for reading and word recognition. Differences in timing do not always result in differences in outcomes—in second grade, there are no statistically significant differences in the dynamic average treatment on the treated estimates between $\tau(0, 1, 1)(0, 0, 0)$ and $\tau(1, 1, 0)(0, 0, 0)$ in any of the subjects. Another insight is that additional doses of treatment on a treatment path may not always yield additional tangible benefits. Comparing $\tau(1, 1, 1, 1)(0, 0, 0, 0)$ to $\tau(1, 1, 1, 0)(0, 0, 0, 0)$, the former sequence does not appear to be that much more beneficial for all subjects because the estimated DATTs are well within each other's confidence intervals (that is, there is no statistically significant difference at the 5 percent level). Hence, the benefit of a teacher

Table 5. Dynamic Average Treatment on the Treated Estimates

	Mathematics	Reading	Word Recognition
Kindergarten			
$\tau(1)(0)$	11.40**	5.08**	5.10**
	(2.52)	(1.63)	(1.87)
Observations	5,782	5,701	5,762
First grade			
$\tau(1, 1)(0, 0)$	16.38**	12.57**	12.31**
	(2.74)	(2.76)	(3.15)
$\tau(1, 0)(0, 0)$	4.39	8.78**	9.09**
	(2.58)	(2.57)	(3.18)
Observations	3,958	3,865	3,359
Second grade			
$\tau(1,1,1)(0,0,0)$	12.40**	7.53*	7.57
	(3.20)	(2.95)	(4.00)
$\tau(1,1,0)(0,0,0)$	6.16*	3.55	6.26
	(3.09)	(3.29)	(4.29)
$\tau(0,1,1)(0,0,0)$	6.07	2.59	3.02
	(3.64)	(3.35)	(4.45)
Observations	2,336	2,338	2,348
Third grade			
$\tau(1,1,1,1)(0,0,0,0)$	7.80*	12.73**	20.41**
	(3.41)	(3.03)	(5.10)
$\tau(1,1,1,0)(0,0,0,0)$	12.32**	15.10**	19.60**
	(3.45)	(3.42)	(5.44)
$\tau(1,1,0,0)(0,0,0,0)$	7.18*	13.31**	19.79**
	(3.34)	(2.93)	(4.53)
$\tau(0,0,1,1)(0,0,0,0)$	0.61	−0.58	0.62
	(3.77)	(3.19)	(4.46)
Observations	1,840	1,852	1,877

Notes: The table displays the dynamic average treatment on the treated estimates for exposure to a teacher of the same race for a given treatment path $\tau(\cdot)$. Standard errors clustered at the level of the classroom are given in parentheses. Scaled test scores are used as the response variable. Regressions as controls include class type, free lunch status, teacher years of experience and its square, whether the teacher has a graduate degree, and whether the teacher is black. Student fixed effects are included in the specification. Observations are weighted using inverse probability weights; see section A.3 of the online appendix.

*Statistical significance at the 5% level; **statistical significance at the 1% level.

of the same race for mathematics, reading, and word comprehension in third grade may potentially be limited.

This paper has heretofore demonstrated that there exist statistically significant benefits to academic achievement by sorting students and teachers along the dimension of race. The question remains as to the policy relevance, which depends on the *economic* significance of these gains. Dividing the coefficients of the DATT from table 5 by the standard deviations of the test scores in the respective grades, I find that the gains from having a same-race teacher to be about the level of the benefit of being assigned to a small class in Project STAR (see, e.g., Mueller 2013). For example, the benefits of treatment in kindergarten range from 0.14 standard deviations for word recognition test scores to 0.24 standard deviations for mathematics scores. Continuous treatment in reading from kindergarten through third grade yields a test score increase of 0.34 standard deviations. The magnitude of the effects is roughly comparable with those found

in Dee (2004).[18] Overall, these represent moderate gains in academic achievement and are therefore policy relevant.

Robustness

Teacher Sorting

There is a concern that the results of the analysis are driven by selection due to teacher sorting *across* schools since teachers were randomized only *within* schools. This is an important consideration, as it has been shown that teacher–school matching is a relevant factor in education production (Jackson 2013). If schools whose students were primarily white attracted high-quality white teachers and poor-quality black teachers, and predominantly black schools attracted high-quality black teachers and low-quality white teachers, then the estimates of the racial complementarities would be biased upwards. Because approximately 85 percent of the total variation in teacher quality occurs within schools (Chetty, Friedman, and Rockoff 2014; Rothstein 2014), this pattern of sorting is a possibility that must be taken seriously. Note that controlling for teacher observables does not solve the selection problem in this case because the significance of teacher unobservable heterogeneity in the determination of student achievement is quite high, and is responsible for far more of its variation than observable characteristics, such as the teacher's qualifications or experience (Rivkin, Hanushek, and Kain 2005).

The effects of teacher quality on the robustness of the racial interaction effects can be assessed by using classroom fixed effects in place of school fixed effects in the regression (Dee 2004). This will result in the racial interaction effects being identified using within-classroom variation; therefore, any potential teacher sorting across schools by quality and race will no longer be conflated with the racial interaction effect. This is because the estimate of the effect is no longer also capturing any potential within-school quality differences of a low-quality teacher whose students are mostly of the opposite race with a high-quality teacher whose students are largely of the same race (which is a potential danger if the racial interaction effects are identified using within-school variation—i.e., by using a school fixed effect). An additional benefit of including classroom fixed effects is that it also controls for other unobservable teacher inputs and classroom effects (such as peer effects). Estimating the system of equations using classroom fixed effects, I find no substantive differences in the results, which are displayed in table 6.[19]

Subsample Analysis

There is a moderate level of noncompliance with classroom type assignment in the Project STAR data. Although noncompliance was estimated to be only about 0.3 percent of the sample in kindergarten (Krueger 1999), a significant number of students moved between regular, regular with aide, and small classes in first grade and beyond. In the sample, approximately 5 percent do not comply in first grade, about 13 percent

18. Dee (2004) uses percentile scores in his analysis—an indirect comparison can be made by examining what the percentile scores correspond to on average in terms of standard deviations.
19. These regressions contain far fewer control variables. In particular, there is no teacher race variable, because this would be perfectly collinear with the "same race as teacher" variable. For example, if the teacher is black, all white students would have the same-race dummy equal to zero.

Table 6. Own-race Teacher Effect on Achievement, Using Classroom Fixed Effects

	Mathematics	Reading	Word Recognition
Kindergarten			
$\tau(1)(0)$	15.39**	6.00**	5.28*
	(2.63)	(1.76)	(2.05)
Observations	5,782	5,701	5,762
First grade			
$\tau(1, 1)(0, 0)$	17.12**	13.01**	13.15**
	(2.69)	(2.86)	(3.40)
$\tau(1, 0)(0, 0)$	5.57*	10.94**	10.28**
	(2.55)	(2.93)	(3.79)
Observations	3,958	3,865	3,359
Second grade			
$\tau(1,1,1)(0,0,0)$	13.76**	8.46*	9.22*
	(3.36)	(3.40)	(4.11)
$\tau(1,1,0)(0,0,0)$	7.89**	3.28	4.01
	(3.00)	(2.95)	(3.81)
$\tau(0,1,1)(0,0,0)$	6.68*	3.55	5.26
	(3.31)	(3.27)	(4.42)
Observations	2,336	2,338	2,348
Third grade			
$\tau(1,1,1,1)(0,0,0,0)$	13.68**	17.85**	26.94**
	(3.25)	(3.26)	(5.13)
$\tau(1,1,1,0)(0,0,0,0)$	16.90**	18.52**	25.26**
	(3.39)	(3.77)	(5.63)
$\tau(1,1,0,0)(0,0,0,0)$	11.32**	14.50**	20.01**
	(3.18)	(3.05)	(4.43)
$\tau(0,0,1,1)(0,0,0,0)$	2.36	3.35	6.93
	(3.38)	(3.22)	(4.30)
Observations	1,840	1,852	1,877

Notes: The table displays the dynamic average treatment on the treated estimates for exposure to a teacher of the same race for a given treatment path $\tau(\cdot)$. Standard errors clustered at the level of the classroom are given in parentheses. Scaled test scores are used as the response variable. Regressions only include free lunch status as additional covariates, since the classroom fixed effects absorb all classroom-invariant control variables. Student fixed effects are included in the specification. Observations are weighted using inverse probability weights; see section A.3 of the online appendix.

*Statistical significance at the 5% level; **statistical significance at the 1% level.

do not comply in second grade, and roughly 20 percent do not comply in third grade. If students nonrandomly switched class types based on the race of the teacher they would have been assigned, estimates of the teacher effects would be biased and inconsistent. To examine whether the results are sensitive to nonrandom switchers, I estimate the system of equations using only those that comply with their treatment assignment, the results of which are displayed in table 7.[20] Despite the loss of a considerable number of observations, the results are largely similar to those from the full sample.[21]

20. These estimates may be biased in the presence of nonrandom attrition—they are meant to serve as a sanity check.

21. To account for the possibility of nonrandom sorting across classrooms (such as noncompliance based on the student's lack of a racial match with his teacher in the assigned class type), Dee (2004) uses an instrumental variables strategy using the probability that a student is assigned to a teacher of the same race as the instrument. Compared with the ordinary least squares estimates, the results are almost unchanged, providing strong

Table 7. Own-race Teacher Effect on Achievement, Compliers of Treatment Assignment

	Mathematics	Reading	Word Recognition
First grade			
$\tau(1, 1)(0, 0)$	16.88**	11.42**	7.38
	(3.91)	(3.99)	(4.33)
$\tau(1, 0)(0, 0)$	6.26*	10.43**	10.91**
	(2.71)	(3.05)	(3.59)
Observations	3,660	3,572	3,121
Second grade			
$\tau(1,1,1)(0,0,0)$	16.02**	8.83*	7.73
	(3.58)	(3.69)	(4.57)
$\tau(1,1,0)(0,0,0)$	9.10*	7.49	8.40
	(3.64)	(4.10)	(5.06)
$\tau(0,1,1)(0,0,0)$	6.29	0.14	−3.98
	(4.15)	(4.31)	(5.67)
Observations	1,996	1,997	2,005
Third grade			
$\tau(1,1,1,1)(0,0,0,0)$	12.26**	11.27**	15.19**
	(3.62)	(3.56)	(5.81)
$\tau(1,1,1,0)(0,0,0,0)$	15.13**	15.09**	12.56*
	(3.79)	(4.06)	(5.75)
$\tau(1,1,0,0)(0,0,0,0)$	10.98**	17.36**	16.96**
	(3.84)	(3.75)	(4.92)
$\tau(0,0,1,1)(0,0,0,0)$	1.28	−6.09	−1.77
	(4.15)	(3.45)	(4.78)
Observations	1,457	1,468	1,487

Notes: The table displays the dynamic average treatment on the treated estimates for exposure to a teacher of the same race for a given treatment path $\tau(\cdot)$ using the subpopulation of those that comply with their assigned class type. Standard errors clustered at the level of the classroom are given in parentheses. Scaled test scores are used as the response variable. Regressions as controls include class type, free lunch status, teacher years of experience and its square, whether the teacher has a graduate degree, and whether the teacher is black. Student fixed effects are included in the specification. Kindergarten estimates are not included because students are assumed to have complied in this initial grade; they would be identical to those in table 5. Observations are weighted using inverse probability weights; see section A.3 of the online appendix.

*Statistical significance at the 5% level; **statistical significance at the 1% level.

Past research has found that the effect of small classes may vary according to the school characteristics (Ding and Lehrer 2011). Given this, I examine whether there exists a differential effect of racial matching according to school size. Both small schools (defined as the bottom 50 percent in school enrollment at kindergarten) and large schools (defined as the top 50 percent) show largely similar qualitative and, in most cases, quantitative results. Unfortunately, robustness checks according to a school's racial composition are not possible because data are only available for the current grade of each school in Project STAR.

Dee (2004) finds that own-race teacher effects existed in almost all subjects for both blacks and whites, and that the magnitude of the effects was similar. Here, I investigate whether there exists a differential effect of an own-race teacher treatment for black students, who constitute about a third of the sample. I estimate the regressions only

evidence that this type of sorting was absent in the data. An analogous approach is not possible here because of the estimation strategy utilized.

Table 8. Own-race Teacher Effect on Achievement, Black Students

	Mathematics	Reading	Word Recognition
Kindergarten			
$\tau(1)(0)$	5.31	3.36	2.49
	(4.60)	(3.02)	(3.26)
Observations	1,889	1,852	1,889
First grade			
$\tau(1, 1)(0, 0)$	16.07*	12.76*	12.50
	(6.55)	(5.72)	(7.07)
$\tau(1, 0)(0, 0)$	10.87*	9.60*	7.99
	(4.96)	(4.57)	(5.42)
Observations	1,156	1,150	990
Second grade			
$\tau(1,1,1)(0,0,0)$	5.37	3.91	0.90
	(6.96)	(6.62)	(8.22)
$\tau(1,1,0)(0,0,0)$	0.87	2.87	2.74
	(5.64)	(5.77)	(6.70)
$\tau(0,1,1)(0,0,0)$	1.33	−1.17	−5.09
	(6.29)	(6.60)	(8.48)
Observations	642	641	646
Third grade			
$\tau(1,1,1,1)(0,0,0,0)$	3.69	15.84*	14.67
	(7.56)	(7.26)	(8.07)
$\tau(1,1,1,0)(0,0,0,0)$	10.67	17.01**	17.07*
	(6.60)	(6.28)	(7.22)
$\tau(1,1,0,0)(0,0,0,0)$	0.56	14.52**	19.22**
	(5.40)	(5.17)	(5.72)
$\tau(0,0,1,1)(0,0,0,0)$	3.13	1.32	−4.56
	(6.20)	(5.09)	(5.47)
Observations	436	441	448

Notes: The table displays the dynamic average treatment on the treated estimates for exposure to a teacher of the same race for a given treatment path $\tau(\cdot)$ using black students only. Standard errors clustered at the level of the classroom are given in parentheses. Scaled test scores are used as the response variable. Regressions as controls include class type, free lunch status, teacher years of experience and its square, whether the teacher has a graduate degree, and whether the teacher is black. Student fixed effects are included in the specification. Observations are weighted using inverse probability weights; see section A.3 of the online appendix.

*Statistical significance at the 5% level; **statistical significance at the 1% level.

using black students—the results of the estimation are in table 8. It is important to note that there are significantly fewer observations compared with most of the other regressions in this paper, which entails a substantial cost in precision. The benefits from treatment appear to exhibit a qualitatively similar pattern to the results in table 5. Nevertheless, the lack of precision means that we cannot reject the hypotheses that the estimates are different from zero in many cases, even though they may be numerically similar to the DATT estimates of the full sample.

Additional robustness checks were also performed to see whether the effects vary by gender and by socioeconomic status. I repeat the analysis of the main section by gender and do not find any substantive changes in the results. An analysis of those of low socioeconomic status (proxied by receiving free or reduced-price lunches in kindergarten) also reveals that the numbers are largely unchanged. For reasons of space, these tables are not included in this paper.

5. POLICY DISCUSSION

Because much of the benefit from an own-race teacher comes from kindergarten and first grade in most subjects and the benefit appears to persist for at least a few years, some may argue that it may be justifiable to sort teachers and students according to race in the first few grades if the goal is to maximize student achievement.[22] Such a policy is especially attractive because effectively costless gains may potentially be obtained by simply reallocating students and teachers across classrooms. Though academic benefits of sorting students and teachers across race within classrooms are present, they come with an important caveat: Additional research should first be conducted on the source of these complementarities to determine *why* they exist before incorporating such a policy.[23] For example, if they are present because teachers exert more effort to students who are of the same race, and such effort comes at the intensive margin, then teachers are engaging in favoritism towards students of the same race at the expense of students who do not share their racial background. In short, we do not yet know if the same-race effects are a "free lunch." Moreover, such racial sorting could have pernicious effects on student noncognitive skills, such as the ability to socialize and interact with students of different races or the willingness to respect authority figures of a different race. General equilibrium issues may also be relevant because of supply constraints. Should a concerted effort to hire a more representative workforce in order to more effectively incorporate this policy be successful, it may result in a lower average quality of teachers from the underrepresented races if we assume that the highest quality teachers are hired first (and a higher average quality for the majority race teachers). This latter assumption seems plausible—California's experiment with class size reductions led to considerable decreases in teacher quality and exacerbated inequalities across school districts because educational institutions were forced to hire teachers who lacked experience and credentials in order to implement the policy (Imazeki 2003; Jepsen and Rivkin 2009). At this time, there appears to be insufficient evidence to support a policy of sorting students and teachers across classrooms by race.

The positive influence of a teacher of the same race on student achievement may help explain a small but nontrivial part of the racial test-score gap between black and white students, because black students in the sample are far less likely to be matched with an own-race teacher compared with white students. This pattern continues to this day because of the continuing shortage of minority teachers (Ingersoll 2015). Table 9 displays the data concerning the racial test-score gap in the Project STAR data, where the figures are in standard deviation units.[24] The raw gap for math is slightly over half the size, as in Fryer and Levitt (2006) (where they find a gap of 0.663 standard deviations), and the raw gap in reading is minimally smaller (where it is 0.4 standard deviations). Including student and teacher covariates does not appreciably change the

22. Racial sorting at the level of the classroom, although not wholesale segregation, may nonetheless be subject to legal challenges under Title IX, for example.
23. Moreover, segregation at the school level (rather than at the level of the classroom) is not a policy that is being recommended here. Historically, such policies have had many pernicious social and economic effects when implemented.
24. School fixed effects are included in the adjusted gaps to account for the fact that the kindergarten grades in the sample have a high level of racial segregation. In this analysis school grades whose student bodies are white are much more likely to have own-race teacher matches, and therefore the contribution to the same-race teacher gap may be overestimated if this is not controlled for.

Table 9. Estimated Black–White Test Score Gap in Kindergarten

	Raw Gap	Adjusted	With Same Race	% of Gap Explained
Mathematics	−0.37	−0.36	0.28	21.72
Reading	−0.36	−0.25	−0.21	13.55
School fixed effects?	no	yes	yes	

Notes: This table displays regression results where a normalized test score is the response variable, and the displayed coefficient is the black student dummy. Numbers are in standard deviations, save the final column. The adjusted column includes student and teacher covariates, and the column following adds a same-race teacher dummy.

gap in math but decreases it considerably in reading—although these adjusted gaps are much larger than in Fryer and Levitt (2006).[25] Augmenting the model further with an own-race teacher variable moderately narrows the racial gap for mathematics, and provides a drop of roughly half that reduction in reading. Overall, accounting for racial matches appears to explain a nontrivial portion of the gap in test scores between black and white students.

ACKNOWLEDGMENTS

This paper is based on a chapter of my doctoral thesis and I thank my thesis committee for their feedback. I graciously thank my advisors Steve Lehrer and James MacKinnon for their supervision and comments on this project. I have benefited from discussions with Joseph Altonji, Gigi Foster, Weili Ding, Jean-Sebastien Fontaine, Vincent Pohl, and Caroline Weber. I would like to thank seminar participants at Queen's University, the University of Toronto, and SOLE 2014 for their feedback.

REFERENCES

Bond, Timothy N., and Kevin Lang. 2013. The evolution of the black-white test score gap in grades K–3: The fragility of results. *Review of Economics and Statistics* 95(5):1468–1479.

Cascio, Elizabeth U., and Douglas O. Staiger. 2012. Knowledge, tests, and fadeout in educational interventions. NBER Working Paper No. 18038.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *Quarterly Journal of Economics* 126(4):1593–1660.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review* 104(9):2593–2632.

Clotfelter, Charles T., Helen F. Ladd, and Jacob Vigdor. 2010. Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects. *Journal of Human Resources* 45(3):655–681.

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

25. Caution should be taken in comparing the adjusted gaps, however. A direct comparison between these gaps and the gaps of Fryer and Levitt (2004, 2006) is not possible because the latter use a different set of covariates.

Ding, Weili, and Steven F. Lehrer. 2010. Estimating treatment effects from contaminated multiperiod education experiments: The dynamic impacts of class size reductions. *Review of Economics and Statistics* 92(1):31–42.

Ding, Weili, and Steven F. Lehrer. 2011. Experimental estimates of the impacts of class size on test scores: Robustness and heterogeneity. *Education Economics* 19(3):229–252.

Ding, Weili, and Steven F. Lehrer. 2014. Understanding the role of time-varying unobserved ability heterogeneity in education production. *Economics of Education Review* 40:55–75.

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

Finn, Jeremy D., Jayne Boyd-Zaharias, Reva M. Fish, and Susan B. Gerber. 2007. Project STAR and beyond: Database user's guide. Lebanon, TN: HEROS, Incorporated.

Fryer, Roland G., and Steven D. Levitt, 2004. Understanding the black–white test score gap in the first two years of school. *The Review of Economics and Statistics* 86(2):447–464.

Fryer, Roland G., and Steven D. Levitt. 2006. The black–white test score gap through third grade. *American Law and Economics Review* 8(2):249–281.

Hanushek, Eric A. 1999. Some findings from an independent investigation of the Tennessee STAR Experiment and from other investigations of class size effects. *Educational Evaluation and Policy Analysis* 21(2):143–163.

Hanushek, Eric A., John F. Kain, Daniel M. O'Brien, and Steven G. Rivkin. 2005. The market for teacher quality. NBER Working Paper No. 11154.

Imazeki, Jennifer. 2003. Class-size reduction and teacher quality: Evidence from California. In *School finance and teacher quality: Exploring the connections*, edited by David Monk and Margaret Plecki, 159–178. Abingdon, UK: Routledge.

Ingersoll, Richard. 2015. What do the national data tell us about minority teacher shortages? In *The state of teacher diversity in American education*, 14–22. Washington, DC: Albert Shanker Institute.

Ingersoll, Richard, and Henry May. 2011. *Recruitment, retention and the minority teacher shortage*. Philadelphia, PA: Consortium for Policy Research in Education.

Jackson, C. Kirabo. 2013. Matching quality, worker productivity, and worker mobility: Direct evidence from teachers. *Review of Economics and Statistics* 95(4):1096–1116.

Jepsen, Christopher, and Steven Rivkin. 2009. Class size reduction and student achievement: The potential tradeoff between teacher quality and class size. *Journal of Human Resources* 44(1):223–250.

Krueger, Alan B. 1999. Experimental estimates of education production functions. *Quarterly Journal of Economics* 114(2):497–532.

Lechner, Michael, and Ruth Miquel. 2010. Identification of the effects of dynamic treatments by sequential conditional independence assumptions. *Empirical Economics* 39(1):111–137.

Miquel, Ruth. 2003. Identification of effects of dynamic treatments with a difference-in-differences approach. Unpublished paper, University of St. Gallen, Switzerland.

Mueller, Steffen. 2013. Teacher experience and the class size effect—Experimental evidence. *Journal of Public Economics* 98: 44–52.

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

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

Rothstein, Jesse. 2010. Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics* 125(1):175–214.

Rothstein, Jesse. 2014. Revisiting the impacts of teachers. Unpublished paper, University of California, Berkeley.

Word, Elizabeth, John Johnston, Helen Pate Bain, B. DeWayne Fulton, Jayne Boyd Zaharias, Charles M. Achilles, Martha Nannette Lintz, John Folger, Carolyn Breda. 1990. *The state of Tennessee's Student/Teacher Achievement Ratio (STAR) Project. Technical report 1985–1990*. Available www.classsizematters.org/wp-content/uploads/2016/09/STAR-Technical-Report-Part-I.pdf. Accessed 18 January 2017.