Racial Interaction Effects and Student Achievement

Jeffrey Penney[1]
Department of Economics, Queen's University
penneyj@econ.queensu.ca

version 4.0.0 - October 24, 2013

This paper examines the possibility that racial complementarities exist in the education production function. Using a model of education production that accounts for the full history of inputs, a conditional differences-in-differences estimation procedure is employed to nonparametrically identify dynamic treatment effects of various sequences of interventions. The approach is applied to Tennessee's Project STAR class size experiment. Consistent with the previous literature, I find beneficial effects on student achievement across many dimensions when matching pupils to teachers of the same race. However, the timing as well as the dosage of the intervention appears to matter a great deal. The channels through which the complementarities operate are examined. This racial matching effect appears to explain a small but non-trivial portion of the Black-White test score gap.

Keywords: achievement, dynamic treatment effects, education, race

JEL Classification: I21, J15

1. Introduction

When estimating education production functions (EPF), many papers include racial dummy variables for the student or the teacher. The inclusion of such factors can be justified on the grounds that they serve as proxy variables for other inputs that are not normally observable; for example, children of certain racial backgrounds may receive considerably higher home inputs. Given the role these variables play in the estimation of EPFs, their coefficients appear to lack any clear interpretation. However, they do appear to be relevant factors in the analysis of student achievement, with most regression analyses of this topic finding statistically significant effects of race. Some studies find that non-Caucasian teachers are negatively correlated with student achievement. An analysis of four cohorts of North Carolina 10th graders using an administrative dataset shows that non-white teachers are negatively correlated with student achievement (Clotfelter et al., 2010); this result is of

[1]Acknowledgements: I have benefited from discussions with Joseph Altonji, Gigi Foster, Weili Ding, Jean-Sebastien Fontaine, Steve Lehrer, Vincent Pohl, and Caroline Weber. All remaining errors are my own.

particular of interest as the authors control for a large array of covariates. A study of 5th graders in the same state finds similar negative effects, however, some of the effects disappear under different specifications or with different subsamples (Clotfelter et al., 2006).

There exists a voluminous literature concerning a gap in test scores between white students and black students. There is evidence that minorities are often taught by teachers of lower quality, such as those with less experience or fewer educational credentials (Clotfelter et al., 2005). Unfortunately, while the black-white test score gap shrinks when observable characteristics are controlled for, a significant gulf is still present (Fryer and Levitt, 2004; Clotfelter et al., 2009; Bond and Lang, forthcoming). Not all minorities suffer from test score deficiencies relative to whites, as being of Asian descent is correlated with higher test scores (Krueger, 1999; Ding and Lehrer, 2010).

While racial characteristics of teachers and students are often analyzed in isolation, some scholars theorize the existence of interactions between them. Current research suggests that there is a positive correlation between student achievement and exposure to own-race teachers (Hanushek et al., 2005). It appears that such associations may at least partly explain the aforementioned racial correlations between teacher race and student performance: Clotfelter et al. (2010) determine that the negative effect of a black teacher on student achievement is driven by a strong negative racial interaction between black teachers and white students, while no negative relationship exists when matched with a student of the same race. In the most comprehensive study into racial interaction effects on student achievement, Dee (2004) finds strong positive effects of matching students with own-race teachers. Using data from Project STAR and examining each gender-race combination separately, he finds that the benefits are of the order of approximately 4 percentile points in all subjects, with the exception of white females in reading, where the gain is smaller but not statistically significant. There is also evidence that students may behave differently according to whether their race is the same as that of their teacher: having a teacher of another race is associated with increased odds that students will be disruptive, inattentive, or rarely complete their homework[2] (Dee, 2005).

The channels through which a common racial background increases student achievement are currently unknown. Most explanations that have been advanced can be classified as one of two types: passive teacher effects or active teacher effects. Passive teacher effects include ones that exist simply due to the teacher's racial presence, such as role-model effects. Active teacher effects reflect race-specific patterns of behaviour; for example, teachers could input less effort into helping students of a different race. These effects are not mutually exclusive. A detailed overview of the state of the literature on racial effects can be found in Section II of Dee (2004).

This paper investigates the effect of racial interactions on student achievement. It differs from previous examinations of the topic along several dimensions. The analysis is based on a dynamic model of education production, taking into account the full history of observed inputs. This approach is undertaken since both the timing and the dosage of the inputs may be relevant: for example, it may be the case that the use of a full time teacher's aide in kindergarten leads to higher scores at that grade, but is ineffective at raising achievement

[2]For the particular study cited, the effect was found only to be statistically significant in the US south census region, but all other regions found similar relationships.

in third grade. In an attempt to understand the channels through which these own-race teacher effects influence student achievement, I include teacher and student effort choices in the education production function. This expanded framework allows me to test the validity of some hypotheses regarding these interactions. To estimate the model, I employ a conditional differences-in-differences procedure to nonparametrically identify dynamic treatment effects of exposure(s) to a same-race teacher. The theoretical model and the estimation procedure are explained in Section 2.

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 employ data from a cohort that participated in the experiment from kindergarten until the end of third grade. There are three primary concerns that need to be assuaged for the analysis of this paper to be valid. The first consideration is whether randomization of students and teachers according to class size is equivalent to randomization according to race. I find that there is no evidence of racial sorting between students and teachers across classrooms in all grades. The second potential issue is whether and how to account for attrition, since Project STAR experienced sample retention problems during its implementation. An investigation into the attrition behaviour of the sample reveals that it is nonignorable; this paper implements a correction for this feature of the data. Lastly, I consider whether the unobservable determinants of achievement are correlated with the observable inputs; I find that they are, which validates the choice of estimation strategy employed here. Section 3 describes the data and performs the aforementioned preliminary robustness and specification checks.

Section 4 conducts the empirical analysis and outlines the main findings. Exposure to a teacher of the same race increases test score performance across a variety of dimensions. However, the timing appears to matter, as the effect is strongest in the early grades. The estimated dynamic treatment effects show that the benefits appear to be permanent, with early grade exposure to same-race teachers having statistically significant benefits to scores in later grades. There does not appear to be any evidence that these effects are concentrated through any one channel such as student motivation, teacher effort, or years of experience. The robustness of these findings is investigated in Section 5. There, I examine whether the results are robust to the possibility that they are merely an artifact of across-school sorting by teachers and students; this is an important consideration, as Project STAR randomized students and teachers only within schools. Repeating the analysis with classroom fixed effects, I find no substantive differences in the results.

To conclude the paper, I discuss the policy implications of the findings in Section 6. While there are statistically significant gains in achievement when students are taught by teachers of the same race, the gains in scaled test scores appear to be low or moderate in magnitude, ranging from approximately 1% to 3%. The own-race teacher effect on achievement may explain a non-trivial portion of the Black-White test score gap, since minorities are far less likely to be matched with a teacher of the same race. Overall, the benefits of racial matching on student achievement alone do not appear to be significant enough to justify hiring a more representative workforce.

2. Model

Most models of education production are based on those of Ben-Porath (1967) and Boardman and Murnane (1979), wherein output is seen as a function of both past and contemporaneous factors. Todd and Wolpin (2003) model achievement by assuming it is instead produced by a functional; that is, some inputs of education production functions are themselves functions or decision rules[3]. Typically included in these mathematical objects are familial and school inputs as well as student characteristics. Here, I augment the usual analysis of education production in two ways: first, I include the effort choices of students and teachers. Second, a set of dummy variables is used to absorb any residual effect of an own-race teacher on achievement that is not explained through the other included channels. I specify the education production function as:

$$A_{ig} = f(P, T, S, F, D, \epsilon) \tag{1}$$

where A_{ig} is the achievement of student i in grade g, $\epsilon = \{\epsilon_{ig}, \epsilon_{i,g-1}, ...\}$ are the independent random period g shocks, and $D = \{d_{ig}, d_{i,g-1}, ...\}$ are the aforementioned same-race teacher dummy variables; these take the value of 1 if the student has a teacher of the same race in grade g and 0 otherwise. The decision rules describing the inputs of the pupil, the teachers, the school, and the family are respectively given by:

$$\begin{aligned} P &= P(m_{ig}, m_{i,g-1}, ..., \mu_i) \\ T &= T(e_{jg}, x_{jg}, e_{j,g-1}, x_{j,g-1}, ...) \\ S &= S(c_{ig}, c_{i,g-1}, ...) \\ F &= F(w_{ig}, w_{i,g-1}, ...) \end{aligned}$$

where μ_i is the unobservable ability of student i, and m_{ig} is their effort in period g. The terms e_{jg} and x_{jg} denote the effort and characteristics of teacher j in grade g, respectively. The variable w_{ig} denotes family inputs. School characteristics, such as class type, are included in c_{ig}[4]. I allow all of the arguments in the above decision rules to be functions of d from the same period, inter alia.

The above formulation is very flexible, as it allows for the possibility of the timing of the inputs to matter. For example, consider the effect of small class sizes: it could be the case that attending a small class in kindergarten and first grade has a different effect than attending a small class in second and third grade. In both scenarios, the student attended a small class for two grades, but the overall effect on achievement could differ despite the level of the inputs (two doses of the small class treatment) being identical. Another advantage is that it decreases the probability of omitted variable bias affecting the results compared to a contemporaneous inputs model. Continuing with the small class example above, if being in a small class in third grade is positively correlated with being in a small class in kindergarten, and if the benefit of such an intervention only exists in kindergarten, then the estimate of

[3]For ease of exposition, I will refer to education production functionals as education production functions, as the difference is immaterial to the substance of this paper.

[4]Since schools generally match students to teachers, the racial match can be thought of as a direct school input, though I do not interpret it this way in this paper.

the small class effect in third grade would be positive despite the lack of any true benefit.

It is important to note that since many of the inputs and decision rules are unobservable, econometric estimation of (1) where the student-teacher racial matches are randomly assigned produces a policy parameter describing the expected value of the *total* effect of an own-race teacher[5]. The total effect is given by

$$\frac{dA_{ig}}{dD} = \underbrace{\frac{\partial f}{\partial P}\frac{\partial P}{\partial m}\frac{dm}{dD}}_{\text{(a)}} + \underbrace{\frac{\partial f}{\partial T}\frac{\partial T}{\partial e}\frac{de}{dD}}_{\text{(b)}} + \underbrace{\frac{\partial f}{\partial T}\frac{\partial T}{\partial x}\frac{dx}{dD}}_{\text{(c)}} + \underbrace{\frac{\partial f}{\partial S}\frac{\partial S}{\partial c}\frac{dc}{dD}}_{\text{(d)}} + \underbrace{\frac{\partial f}{\partial F}\frac{\partial F}{\partial w}\frac{dw}{dD}}_{\text{(e)}} + \underbrace{\frac{\partial f}{\partial D}}_{\text{(f)}} \tag{2}$$

where m, e, x, c, and w are vectors (e.g. $m = \{m_{ig}, m_{i,g-1}, ...\}$). The individual effects are as follows. The effect of the intervention on the student is given by (a): does the student change his or her effort when faced with a teacher of the same race? Teachers may modify the effort they input to a student if they are of the same race; this is given by (b). The effect (c) corresponds to whether experienced teachers are more or less effective at instructing students of the same race. If schools change their behaviour according to the racial match of a student with his or her teacher, this will appear in term (d). Families that modify the inputs to their child according to whether the child is not of the same race as the teacher are captured by the term (e). All the other channels in which the effect of a racial match affects achievement fall in term (f). Note again the possibility that past inputs can affect future achievement.

I make the following simplifications to allow estimation of the policy effects of interest. The production function is linearized and the error term is assumed to be additively separable. The fixed effect v_i can be correlated with the observed and unobserved determinants of achievement; it contains the effect of not only ability μ_i but also other time-invariant inputs and characteristics. Student and teacher effort are uncorrelated with the included covariates. Other unobservable inputs into the EPF 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. I assume no pretreatment effects. Teacher characteristics x_{jg} comprise of the teacher's race, years of experience, and whether the teacher has a graduate degree. Observable school inputs c_{ig} are the type of class the student attends (a small class, a regular class, or a regular class with a full-time teacher's aide) and school fixed effects (which I allow to vary by grade), while free lunch status is used a proxy variable for family inputs w_{ig}.

Since the education production function is cumulative in its inputs, each grade requires a different specification. Here, I use a similar approach to that employed in Ding and Lehrer (2010). Let X contain all of the included control variables (x, c, and w) and a constant term, and grade $g = k$ denote kindergarten. The regression equation for achievement in kindergarten is

$$A_{ik} = v_i + X_{ik}\alpha_{1k} + \alpha_{2k}d_{ik} + \epsilon_{ik}, \tag{3}$$

while for first grade the equation is given by

$$A_{i1} = v_i + X_{i1}\beta_{11} + X_{ik}\beta_{1k} + \beta_{21}d_{i1} + \beta_{2k}d_{ik} + \epsilon_{i1}. \tag{4}$$

[5]The estimated effect does not correspond to a parameter of the EPF without some very strong assumptions on the decision rules (Todd and Wolpin, 2003).

The regression equations for grades 2 and 3 follow similarly. This specification allows for the effect of inputs to vary over time: for example, the effect of having a same-race teacher in kindergarten on achievement can be different in kindergarten compared to first grade (that is, I allow the possibility that $\alpha_{2k} \neq \beta_{2k}$).

There remains the issue of the student 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 is biased and inconsistent. Therefore, I first difference the system of equations. The equations for achievement in first, second, and third grades are, respectively,

$$A_{i1} - A_{ik} = X_{i1}\beta_{11} + X_{ik}(\beta_{1k} - \alpha_{1k}) + \beta_{21}d_{i1} + (\beta_{2k} - \alpha_{2k})d_{ik} + \epsilon^*_{i1} \tag{5}$$

$$\begin{aligned} A_{i2} - A_{i1} = X_{i2}\gamma_{12} + X_{i1}(\gamma_{11} - \beta_{11}) + X_{ik}(\gamma_{1k} - \beta_{1k}) \\ + \gamma_{22}d_{i2} + (\gamma_{21} - \beta_{21})d_{i1} + (\gamma_{2k} - \beta_{2k})d_{ik} + \epsilon^*_{i2} \end{aligned} \tag{6}$$

$$\begin{aligned} A_{i3} - A_{i2} = X_{i3}\delta_{13} + X_{i2}(\delta_{12} - \gamma_{12}) + X_{i1}(\delta_{11} - \gamma_{11}) + X_{ik}(\delta_{1k} - \gamma_{1k}) \\ + \delta_{23}d_{i3} + (\delta_{22} - \gamma_{22})d_{i2} + (\delta_{21} - \gamma_{21})d_{i1} + (\delta_{2k} - \gamma_{2k})d_{ik} + \epsilon^*_{i3}. \end{aligned} \tag{7}$$

where $\epsilon^*_{ig} \equiv \epsilon_{ig} - \epsilon_{i,g-1}$. Note that the kindergarten equation remains unchanged in this transformation: while the fixed effect is still present, random assignment in this grade ensures that the fixed effect is not correlated with the included covariates. However, because of potentially non-random attrition in the following grades, we require that the fixed effect be differenced out in the other achievement equations.

Because the system of equations {(3), (5), (6), (7)} is triangular, it can be estimated using equation-by-equation OLS to produce unbiased and consistent estimates of the parameters. As the parameters enter recursively into the equations, one is required to estimate them in a sequential fashion (starting with kindergarten) if the desire is to separately identify the coefficients of interest; for example, we require the estimates of α from (3) to enter into (5) in order to obtain the estimates of β.

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 effect on the treated (DATT) estimates. The approach employed here, first outlined in Lechner and Miquel (2010), uses the estimated parameters from the system of equations to calculate the dynamic effects of various sequences of interventions. For purposes of exposition, I consider the case of two periods. Let $\tau(a, b)(w, x)$ be the DATT for the treatment sequence (a, b) with the counterfactual sequence (w, x). For example, $\tau(1, 1)(0, 0)$ refers to the DATT of having an own-race teacher in kindergarten and first grade for those who had teacher of the same race in both grades, $\tau(1, 0)(0, 0)$ describes the effect of an exposure to a teacher of the same race in kindergarten on achievement for those who have only had a same-race teacher in kindergarten, and $\tau(1, 0)(0, 1)$ is the effect of an own-race teacher in kindergarten instead of first grade for those who only had a same-race teacher in kindergarten. Using the estimated

parameters from (5), the DATT for the three examples would be calculated as follows:

$$\tau(1,1)(0,0) = \widehat{\beta_{2k}} + \widehat{\beta_{21}}$$
$$\tau(1,0)(0,0) = \widehat{\beta_{2k}}$$
$$\tau(1,0)(0,1) = \widehat{\beta_{2k}} - \widehat{\beta_{21}}$$

The standard errors of these effects are calculated using the standard formula for sums of random variables[6]. The same logic extends to more than two periods.

3. Data

The data employed in this study come from a cohort of students that participated in Project STAR, an experiment that took place in Tennessee that ran from 1985 until 1989. The experiment was legislated into existence and funded by the state government[7], at a cost of approximately $12 million over five years; this figure included the data analysis and reporting that took place in the fifth year. The primary goal of the experiment, as its acronym (Student-Teacher Achievement Ratio) 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. However, regular classes still had a part-time teacher's aide available to assist the class from approximately 25% to 33% of the time on average. It was initially intended that students stay in their assigned class type from kindergarten through third grade; however, after kindergarten, students in regular or regular with aide classes were randomly permanently reassigned between these two class types. Compliance was almost perfect in kindergarten, with only 0.3% of students enrolled in a class type that was not assigned to them. However, in first grade and beyond, 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. At the end of each year, all participating students were given a battery of academic and non-academic tests. More detailed overviews of Project STAR can be found in Krueger (1999) and Finn et. al. (2007).

In this paper, the measures of student achievement examined are the 7^{th} edition Stanford Achievement Test (SAT) scores in mathematics, reading, word recognition, and listening. The tests were designed so that the scores were comparable across grades (Finn et. al., 2007). I do not transform the scores into percentile scores, as there is some evidence that findings based on percentiles may not be the same as those based on unscaled scores; for example, Bond and Lang (forthcoming) show that the evolution of the black-white test score

[6]For example, the standard error of $\widehat{\tau}(1,1)(0,0)$ is equal to $\sqrt{\text{var}(\widehat{\beta_{2k}}) + \text{var}(\widehat{\beta_{21}}) + 2\text{cov}(\widehat{\beta_{2k}}, \widehat{\beta_{21}})}$.

[7]See House Bill (HB) 544, Tennessee Legislature, 1985.

gap from kindergarten through third grade depends on the scaling decision.

I follow the STAR cohort of students that entered the program in 1985, excluding students that joined after kindergarten. This is done to more credibly estimate the full sequence of dynamic effects (Ding and Lehrer, 2010). I only keep students and teachers whose race is either black or white, which results in a loss of 33 students and 12 teachers from the sample[8].

3.1. Random Assignment

It is generally thought that the ideal way of assessing the effectiveness of some policy or treatment is through a randomized experiment. If treatment is randomly assigned, interpretation is straightforward because treatment status is exogenous. By contradistinction, non-experimental studies require far more care in the interpretation of results due to the possibility of selection bias. For example, if parents of high-ability students are pushing school administrators to match their children with teachers of the same race, we could observe a positive relationship between racial matching and achievement, even if there is no causal mechanism from the former to the latter. This is a serious consideration in education research, as previous studies have shown that classroom sorting according to ability and race occurs even within schools (Clotfelter et. al. 2003; Clotfelter et. al., 2006).

While students and teachers were randomly matched within schools in the Project STAR experiment, the matching was performed according to class type; therefore, it is possible that randomization according to race failed. To assuage this concern, a robustness check is performed to determine whether randomization according to class type is equivalent to randomization according to race. For every grade, I run the following regression:

$$samerace_i = \beta_0 + \beta_1 small_i + \beta_2 aide_i + \theta_l + \epsilon_i \tag{8}$$

where $small_i$ and $aide_i$ are dummy variables taking the value of 1 if the student is assigned to a small class or a regular class with a teachers aide respectively and 0 otherwise, and $samerace_i$ is a dummy variable for whether the student and the teacher are of the same race. The school fixed effect for school l is given by θ_l, whose inclusion in (8) is required since randomization occurred within schools. If β_1 or β_2 are found to be statistically significant in any grade, then there is evidence that being assigned to a certain class type affects the probability of a racial match, which would indicate a failure in randomization. Using the 10% threshold for significance, I do not find any cause to believe that randomization of students and teachers failed in any grade along the dimension of race. Therefore, we can have causal interpretations of the estimated DATT parameters.

3.2. Attrition

Attrition in the Project STAR data is considerable. Of the students who were initially enrolled in kindergarten, 48.9% of them did not reach third grade. If attrition is nonrandom, naive regressions using the data may result in biased and inconsistent estimates despite random assignment of students and teachers.

[8] All of the teachers dropped from the sample are third grade teachers.

Table 1: Test of Randomization

	Small	Regular with aide
Kindergarten	-0.01	-0.01
	(0.04)	(0.04)
First grade	0.06	0.04
	(0.04)	(0.04)
Second grade	0.00	-0.03
	(0.05)	(0.05)
Third grade	0.04	0.02
	(0.04)	(0.04)

Note: the table contains the coefficients on β_1 and β_2 in regression (8). * denotes statistical significance at the 10% level, ** the 5% level, and *** the 1% level. Standard errors clustered at the level of the classroom are given in parentheses.

Past researchers have generally dealt with the attrition problem in STAR in one of four ways. The first is to limit the analysis to the kindergarten data, since randomization was successful in that grade; of course, this prevents the analysis of any dynamics. The second is to interpret the estimate of the intervention as an intent to treat (ITT) parameter. The third is to use an instrumental variables strategy, interpreting the estimated coefficient as a local average treatment effect (LATE). Frangakis and Rubin (1999) show that these two approaches may be problematic in the face of nonrandom attrition, since in this case, the ITT estimator is biased and the IV estimator cannot be interpreted as causal. The fourth method is to employ a partial identification approach and impute the missing values using a number of different assumptions, such as the procedure outlined in Horowitz and Manski (2000). However, the attrition rate is so high in these data that the bounds created using these approaches are typically uninformative. In this paper, I take a different approach that relies on whether the attrition is due to observable or unobservable factors.

I begin by testing to determine whether attrition was due to observable characteristics using a procedure developed by Becketti et al. (1988). I estimate the following regression equation:

$$A_{ik} = X_{ik}\beta_1 + L_i X_{ik}\beta_2 + \theta_l + \epsilon_{ik} \tag{9}$$

where X is a row vector of a constant term and initial characteristics, and L_i is a dummy variable taking the value of 1 if the student leaves the sample before the end of the experiment and 0 otherwise. If the interaction terms in β_2 have a jointly statistically significant effect, then selection on observables is present: based on known characteristics, those who left the experiment before it completed had different achievement scores in kindergarten compared to those who stayed.

The coefficient estimates in the β_2 vector of regression (9) are displayed in Table 2. Students who subsequently left the sample after kindergarten performed much worse in kindergarten compared to non-attritors. The significant negative coefficient on own-race teacher in the math and listening regressions shows that those who attrit at some point in the sample period exhibited lower math and listening scores when paired with a teacher of the

Table 2: Test for attrition based on observables

Variable	Mathematics	Reading	Word Recognition	Listening
Attrition dummy	-15.67***	-13.36***	-11.87**	-5.22***
	(4.54)	(2.67)	(3.35)	(2.75)
Attrition dummy interactions:				
Own-race teacher	-4.67*	0.27	-0.92	-4.00**
	(2.65)	(1.67)	(2.09)	(1.78)
Small class	-3.06	-0.18	-1.39	-1.87
	(2.92)	(1.92)	(2.22)	(2.02)
Regular with aide class	3.93	2.80	2.30	1.53
	(2.66)	(1.76)	(2.09)	(1.93)
Student receives free lunch	-5.35**	-0.62	0.50	-3.50**
	(2.29)	(1.51)	(1.77)	(1.58)
Teacher years of experience	-0.02	-0.21	-0.41	-0.34
	(0.65)	(0.40)	(0.53)	(0.42)
Teacher years of experience^2	0.01	0.01	0.02	0.01
	(0.03)	(0.02)	(0.02)	(0.02)
Teacher has a graduate degree	-1.25	0.68	1.31	0.43
	(2.48)	(1.68)	(1.99)	(1.79)
Teacher is black	3.22	2.22	1.16	2.97
	(3.22)	(2.11)	(2.56)	(1.91)
F-test on variables in β_2:				
all variables	0.0000	0.0000	0.0000	0.0000
interaction variables only	0.0317	0.6398	0.8188	0.1017
constant only	0.0006	0.0000	0.0005	0.0588

Note: the table contains the coefficients on β_2 in regression (9). * denotes statistical significance at the 10% level, ** the 5% level, and *** the 1% level. Standard errors clustered at the level of the classroom are given in parentheses. Numbers given for the F-test are the corresponding p-values of the test using clustered standard errors. Scaled test scores are used as the response variable.

same race in kindergarten; therefore, ignoring attrition may result in an upward bias of the same-race teacher coefficient in regressions involving these test scores as the response variable. Students who receive a free lunch in kindergarten that later attrit also showed lower scores in mathematics and listening. Since the coefficients on the attrition interaction variables are jointly significant in one regression and borderline significant in another, attrition due to observables is likely nonignorable. Previous research has found that attrition patterns across schools that participated in STAR did not systematically differ from those that did not, which should assuage concerns regarding selection on unobservables (Ding and Lehrer, 2010).

Because selective attrition due to observables is present, $\sqrt{N}$ consistent estimates may still be obtained through the use of inverse probability weights; since I am allowing for heterogeneous treatment effects, they are required to consistently estimate the model parameters (Wooldridge, 2002). I perform Duncan and Dumouchel (1983) tests to determine whether weighting produces systematically different estimates compared to an unweighted regression. For every grade and every subject, the null hypothesis that the estimates are not statistically distinguishable is strongly rejected (the F-statistic exceeds 19.2 in all cases, $p = 0.0000$); therefore, weighted estimates are required.

3.3. Summary Statistics

The schools in the Project STAR dataset are highly segregated. Only about 1 in 5 have a racial balance that lies between 20% and 80% of students being of a single race. Moreover, most teachers in schools that have predominantly white student bodies are themselves white, while teachers in schools with majority black student bodies have a more even distribution of teachers. The proportions of students for each grade that are taught by a teacher of the same race are displayed on Table 3.

Table 3: Proportion of students with a teacher of the same race

	Kindergarten	First grade	Second grade	Third grade
White students	0.95	0.96	0.92	0.95
Black students	0.39	0.46	0.45	0.53

The transitions that students experience is displayed in Table 4. We see that the vast majority of students have a teacher of the same race throughout the grades, while other treatment paths have far less support.

An initial look at the relationship between having an own-race teacher and test score performance is presented in Table 5. For white students, the average test score is higher for those with black teachers in only 1 of the 16 categories; for black students, it is higher in 2 of the 16 categories if they have a white teacher. In all cases where the non-racially matched students perform better on average, it occurs in second and third grade.

To better illustrate the association between having an own-race teacher and student achievement, I plot the density of the kindergarten listening test score for each race on Figure 1. In both cases, having a teacher of the same race provides a rightward shift in the distribution of test scores, but some parts of the distribution appear to benefit more than others.

Table 4: Transition tree

Kindergarten	First grade	Second grade	Third grade
			$d_{i3} = 1$, [1946]
			$d_{i3} = 0$, [117]
		$d_{i2} = 1$, [2290]	$L_{i3} = 1$, [227]
		$d_{i2} = 0$, [252]	
		$L_{i2} = 1$, [613]	$d_{i3} = 1$, [176]
			$d_{i3} = 0$, [44]
	$d_{i1} = 1$, [3155]		$L_{i3} = 1$, [32]
	$d_{i1} = 0$, [358]		
	$L_{i1} = 1$, [1301]		$d_{i3} = 1$, [101]
			$d_{i3} = 0$, [43]
		$d_{i2} = 1$, [182]	$L_{i3} = 1$, [38]
		$d_{i2} = 0$, [64]	
		$L_{i2} = 1$, [112]	$d_{i3} = 1$, [18]
			$d_{i3} = 0$, [31]
			$L_{i3} = 1$, [15]
$d_{ik} = 1$, [4814]			
$d_{ik} = 0$, [1435]			
			$d_{i3} = 1$, [101]
			$d_{i3} = 0$, [34]
		$d_{i2} = 1$, [164]	$L_{i3} = 1$, [29]
		$d_{i2} = 0$, [164]	
		$L_{i2} = 1$, [108]	$d_{i3} = 1$, [76]
			$d_{i3} = 0$, [44]
	$d_{i1} = 1$, [436]		$L_{i3} = 1$, [44]
	$d_{i1} = 0$, [502]		
	$L_{i1} = 1$, [497]		$d_{i3} = 1$, [48]
			$d_{i3} = 0$, [32]
		$d_{i2} = 1$, [115]	$L_{i3} = 1$, [35]
		$d_{i2} = 0$, [233]	
		$L_{i2} = 1$, [154]	$d_{i3} = 1$, [52]
			$d_{i3} = 0$, [136]
			$L_{i3} = 1$, [45]

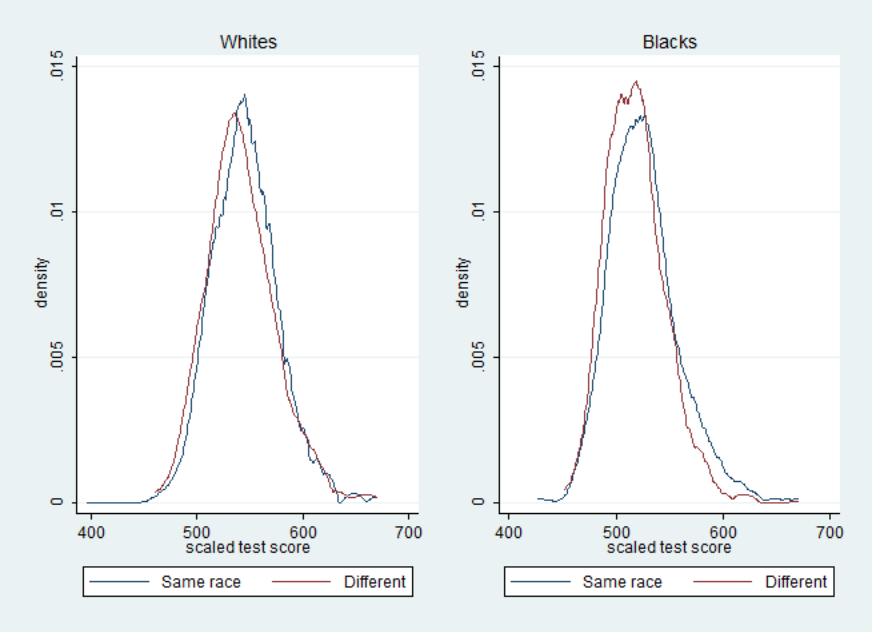
Note: numbers calculated from sample data. The number of students that experience each treatment path are in square brackets. A downward move corresponds to $d_{ig} = 0$ in the previous period, while an upward move signifies $d_{ig} = 1$ in the previous period. $L_{ig} = 1$ signifies attrition in period g. For example, 108 people had the sequence $d_{ik} = 0, d_{i1} = 1$ before attriting in the second grade, and 43 people have undergone the treatment sequence $d_{ik} = 1, d_{i1} = 0, d_{i2} = 1, d_{i3} = 0$.

Table 5: Test Score Summary Statistics by Race

	White		Black	
Variable	Same race	Different race	Same race	Different race
Kindergarten				
Mathematics	491.83	489.08	481.80	468.54
Reading	440.86	438.22	432.90	426.51
Word Skills	439.18	434.75	429.34	422.96
Listening	545.80	542.28	526.06	518.02
First grade				
Mathematics	545.63	525.22	521.44	511.55
Reading	540.72	519.13	503.77	499.83
Word Skills	529.44	514.07	502.36	499.13
Listening	579.51	562.92	557.66	550.94
Second grade				
Mathematics	596.62	593.09	570.01	566.01
Reading	602.55	595.43	571.02	566.98
Word Skills	601.73	599.65	576.76	567.37
Listening	608.56	614.68	581.39	578.41
Third grade				
Mathematics	633.08	620.39	608.06	609.48
Reading	630.19	624.09	608.67	608.06
Word Skills	627.75	620.17	603.10	604.08
Listening	635.39	626.81	614.26	610.67

Note: the table displays the average scaled scores of the subject Stanford Achievement Tests.

Figure 1: Distribution of test scores



This preliminary analysis allows us to come to a substantive conclusion: there may be reason to believe that own-race teachers increase student achievement; should this be true, it may be that this can explain part of the Black-White test score gap since whites are far more likely to be paired with a teacher of the same race compared to blacks.

4. Empirical Analysis

4.1. Results

Table 6 presents the estimates on the coefficients of the d_{ig} variables obtained by estimating the system of equations described in Section 2. These parameters correspond to dynamic average treatment effect 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 that 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 positive effects up until second grade for the case of mathematics, and listening skills tend to benefit throughout. There is a statistically significant negative effect (at the 10% level) of exposure for mathematics in the third grade, but this is likely to be an anomalous result. The case of having an own-race teacher in multiple grades is displayed in Table 7. Multiple exposures are shown to be beneficial in all cases. However, while 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 except for listening, even though both sequences give two exposures to a teacher of the same race. Moreover, 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 except listening, since the estimated DATT is about the same or lower (but well within the center of each other's confidence intervals). Hence, the benefit of a teacher of the same race for mathematics, reading, and word comprehension in third grade appears to be rather limited. We see a similar pattern in the second grade dynamic treatment effect estimates, where $\tau(0,1,1)(0,0,0)$ has at times a much lower benefit than $\tau(1,1,0)(0,0,0)$. The results here are largely in agreement with Dee (2004), with the additional insight that the timing matters for exposure to a teacher of the same race.

4.2. Channels

The method through which exposure to a teacher of the same race increases academic achievement is currently an open question, and many different avenues have been suggested. The model in Section 2, whose total derivative with respect to exposure to an own-race teacher is given by equation (2), allows me to examine some of the possible channels of influence.

The effect on student effort, given by term (a), has been cited as a possible way that own-race teachers can affect achievement. Such a possibility appears to be plausible, as an own-race effect has been found in the peer effects literature: Fruehwirth (2013) shows that peer spillover effects on achievement primarily operate through same-race peers, but is unable

Table 6: Own-race Teacher Effect on Achievement

	Mathematics	Reading	Word Recognition	Listening
Kindergarten				
d_{ik}	11.40***	5.08***	5.10***	8.75***
	(2.52)	(1.63)	(1.87)	(2.05)
First grade				
d_{ik}	4.47*	8.73***	9.12***	4.18**
	(2.55)	(2.54)	(3.15)	(1.81)
d_{i1}	12.05***	3.73	3.01	4.86***
	(2.67)	(2.63)	(3.16)	(1.70)
Second grade				
d_{ik}	6.22**	5.01*	4.73	7.33***
	(2.42)	(2.60)	(3.66)	(2.20)
d_{i1}	0.10	-1.82	0.89	1.23
	(2.97)	(3.13)	(4.16)	(2.19)
d_{i2}	6.41**	4.33	1.90	0.17
	(2.84)	(2.80)	(3.27)	(2.03)
Third grade				
d_{ik}	5.23*	10.97***	13.32***	3.46
	(3.08)	(2.35)	(3.74)	(2.49)
d_{i1}	2.41	2.13	7.14*	-1.72
	(2.94)	(2.44)	(3.89)	(2.44)
d_{i2}	5.55*	2.19	0.63	3.72
	(3.03)	(2.78)	(3.23)	(2.83)
d_{i3}	-5.02*	-2.03	0.03	4.27*
	(2.55)	(2.54)	(3.60)	(2.49)

Note: the table contains the reduced form coefficient estimates on the d_{ig} variables in the system of reduced form equations described in Section 2. These parameters correspond to DATT for single exposures to an own-race teacher; see the text for details. * denotes statistical significance at the 10% level, ** the 5% level, and *** the 1% level. Standard errors clustered at the level of the classroom are given in parentheses. Scaled test scores are used as the response variable.

Table 7: Dynamic Average Treatment Effect on the Treated Estimates

	Mathematics	Reading	Word Recognition	Listening
First grade				
$\tau(1,1)(0,0)$	16.52***	12.46***	12.13***	9.05***
	(2.76)	(2.78)	(3.20)	(1.69)
Second grade				
$\tau(1,1,1)(0,0,0)$	12.74***	7.52**	7.51**	8.73***
	(3.29)	(2.97)	(4.12)	(2.62)
$\tau(1,1,0)(0,0,0)$	6.33**	3.19	5.61	8.56***
	(3.10)	(3.37)	(4.47)	(2.37)
$\tau(0,1,1)(0,0,0)$	6.52	2.51	2.79	1.40
	(3.73)	(3.41)	(4.60)	(2.89)
Third grade				
$\tau(1,1,1,1)(0,0,0,0)$	8.17**	13.26***	21.13***	9.73***
	(3.56)	(2.99)	(5.17)	(2.83)
$\tau(1,1,1,0)(0,0,0,0)$	13.19***	15.29***	21.09***	5.47
	(3.53)	(3.40)	(5.48)	(2.99)
$\tau(1,1,0,0)(0,0,0,0)$	7.64**	13.10***	20.46***	1.75
	(3.38)	(2.87)	(4.67)	(2.78)
$\tau(0,0,1,1)(0,0,0,0)$	0.53	0.16	0.66	7.99**
	(3.88)	(3.19)	(4.66)	(3.25)

Note: the table displays the dynamic average treatment effects on the treated for exposure to a teacher of the same race. * denotes statistical significance at the 10% level, ** the 5% level, and *** the 1% level. Standard errors clustered at the level of the classroom are given in parentheses. Scaled test scores are used as the response variable.

to distinguish whether it is due to an increase in effort or due to an increase in conformity. In the channel examined here, students may find themselves more engaged and motivated due to a *role-model effect* when being the same race as their teacher (King, 1993); as Dee (2004) argues, this effect could occur through updating a student's prior beliefs, such as when an underprivileged black student is matched with a black teacher and becomes more inclined to believe that education can offer them more career possibilities. While a popular explanation, Cizek (1995) argues that there is little direct empirical support for it. A second possibility that could affect student effort is through *stereotype threat*: when students are matched with teachers of a different race, they may experience academic apprehension due to belief in a stereotype which results in reduced performance (Dee, 2004). Both these possibilities are passive teacher effects, since they do not directly operate through the teacher's actions. To examine the possibility of these effects on student motivation, I use the Self-concept and Motivation Inventory test (SCAMIN), which was administered to the Project STAR students during each year. The test produced two figures: a measure of a student's self-concept, with higher scoring students being more well-adjusted, and a score representing how motivated the student is; I use the latter score. I estimate the system of equations described in Section 2, using motivation as the response variable. It appears that there is a negative effect of having an own-race teacher in second grade, but nowhere else. Should the Holm-Bonferroni

correction be employed, we see that there are no statistically significant effects in any grade. Given the additional consideration of the lack of a pattern on the coefficients (half are positive, half are negative), it appears unlikely that motivation is affected by having an own-race teacher. Overall, racial interactions do not appear to have meaningful effects on motivation.

Teacher effort, given by term (b) in equation (2), has been thought to change according to the race of the student. There is limited experimental evidence that this active teacher effect exists (Dee, 2004). If the change in teacher effort were to drive a considerable portion of the benefit to achievement of having an own-race teacher, and assuming that the behavioural effect on teachers is independent of the age of the student for the range of interest here, we would expect the coefficients on d_{ig} in grade g to gradually increase in magnitude over the course of the grades roughly in accordance with the growth in the scaled scores. Given the results in Tables 6 and 7, there does not appear to be much support for this hypothesis, as the coefficients do not appear to exhibit this pattern. Should a teacher effort increase due to a racial match exist, the estimated regression coefficients would imply that the effect of teacher effort in the education production function varies widely by grade. If we do not accept this possibility, then the results suggest that this effort effect on achievement is small or zero.

When the race of the student and the teacher match, there may be a differential effect based on the years of experience the teacher has. For example, perhaps more experienced teachers are less likely to exhibit unconscious biases that favour students of the same race. This possibility refers to part (c) of equation (2). Past research has found a dynamic between teacher experience and the class size effect: an examination of the Project STAR data shows that the small class size effect is only present for senior teachers (Mueller, 2013). To investigate the possibility of an experience interaction, I run the following regression:

$$A_{ig} = \beta_0 + \beta_1 d_{ig} + \beta_2 X_{ig} + \beta_3 d_{ig} x_{ig} + \theta_l + \epsilon_{ig}$$

where β_3 is the coefficient of interest: if it is significant, the relationship between teacher experience (x_{ig}) and student achievement varies according to race. Running regressions for all sixteen grade and subject combinations, I have two statistically significant results: for third grade in reading, the p-value of β_3 is 0.058, while in the third grade for word recognition, it is 0.031. Using the Holm-Bonferonni correction for multiple hypothesis tests that may possibly exhibit dependence, the threshold for significance at the 10% level is 0.0063 for the smallest p-value and 0.0067 for the next smallest. The p-values here fall far short of these thresholds, so I conclude that there does not appear to be a statistically significant estimate of the effect of teacher experience on the same-race interaction effect.

The terms (d) and (e) in (2), while interesting, do not appear to be formally testable with our data. We have much reason to believe that the effect of (d) is zero, at least for the Project STAR data: since randomization according to race was successful, students were not sorted into different classes according to the potential for a racial match with their teacher. Moreover, it appears unlikely that students received different school inputs according to racial pairing either. Part (e) would require knowledge of home inputs, which is beyond the scope of these data. One can speculate that children of racist parents may input more home inputs into their education, but this would attenuate the relationship of the own-race

teacher effect on achievement instead of biasing it upward.

To summarize, there does not appear to be a clear indication that a single channel dominates for the own-race teacher effect on achievement. Given the randomization protocol and the sample size employed, I can conclude this with much confidence.

5. Robustness Checks

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 et. al., 2005). Because of de facto segregation, schools tend to predominantly have either white students or black students. If predominantly white schools primarily attract high quality white teachers and low quality black teachers and vice versa, we would see a positive effect of racial matches that is entirely driven by selection (Dee, 2004), since Project STAR only randomized teaching assignments within schools. An additional consideration is that the specification of the EPF in Section 2 did not consider the possibility of common within-grade peer effects.

To examine the possibility that the results may be driven by these considerations, I re-estimate the system of achievement equations, but include classroom fixed effects instead of school fixed effects in the controls. The results are largely similar to those that use school fixed effects, the only difference being that there are no positive statistically significant effects of an own-race teacher after first grade except for the case of listening, where the benefits of an own-race teacher end after second grade.

One of the assumptions used to justify the conditional differences-in-differences estimation strategy outlined in Section 2 was the presence of individual unobservable heterogeneity and time-invariant characteristics. If such heterogeneity is not present, one could more efficiently estimate the system of equations without differencing. To test whether such an efficiency gain can be realized, I perform a likelihood ratio test of $v_i = 0$ in all equations of the system. As we can see in Table 8, we reject the null hypothesis and conclude that accounting for unobserved heterogeneity is necessary.

Table 8: Likelihood Ratio Tests for the Presence of Unobservables

	Math	Reading	Word Recognition	Listening
Grade 1	621.32	1670.27	765.75	877.7
Grade 2	1090.15	933.01	696.25	951.49
Grade 3	866.81	971.21	262.13	700.81

Note: the table displays chi-square test statistics with 1 degree of freedom for the null hypothesis that $v_i = 0$ in the system of equations described in Section 2. Rejection of the null implies that v_i should remain in the specification. The corresponding p-values for the test statistics on the table are all below $p = 0.001$.

5.1. Subsample Analysis

Noncompliance with treatment assignment in the STAR data is quite considerable. While noncompliance was calculated to be only about 0.3% 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% do not comply in first grade, about 13% do not in second grade, and roughly 20% do not 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 are biased and inconsistent. To examine whether the results are sensitive to nonrandom switchers, I run a regression that only contains those that comply with their treatment assignment, the results of which are displayed in Table 9. The results are largely the same as the full sample, with the exception of the second grade effect on reading, which seemingly gets absorbed into kindergarten. All other cases are largely similar both qualitatively and quantitatively. This is despite the loss of a considerable number of observations. 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% in school enrollment at kindergarten) and the large schools (defined as the top 50%) show largely similar qualitative and in most cases quantitative results.

In Dee (2004), it was found that own-race teacher effects existed in almost all subjects for both blacks and whites. Here, I investigate whether there exists a differential effect of an own-race teacher treatment for black students, who comprise about a third of the sample. I estimate the regressions only using black students; the results of the estimation are in Table 10. The benefits from treatment in kindergarten are still positive, but less precisely estimated. In kindergarten and second grade, positive benefits appear but are not precisely estimated; we do see statistically significant gains in first and third grade. It is important to note that we have significantly fewer observations compared to most of the other regressions in this paper, which entails a cost in precision. One limitation of the current study (and of the STAR data in general) is a lack of information concerning the teacher's aides. Since there is no data on them, it is not possible to test whether there is an additional dynamic with their race relative to the student's and the teacher's. Excluding teacher's aide classes from the sample would not completely fix the problem, since the regular classes still had access to a part-time aide. Up to half the sample is lost in these estimations due to noncompliance and rerandomization into the regular with aide class assignment. With these limitations in mind, I estimate the system of equations outlined in Section 2, and the results are displayed in Table 11. The results are similar in kindergarten. The own-race teacher in kindergarten effect on listening in first grade loses precision but is quantitatively similar, and the other estimates in the grade are roughly the same. In second grade and beyond, the benefits are qualitatively similar, though they vary from the full sample results in terms of timing. Early grades continue to be the most important and we continue to have statistically significant positive effects.

Table 9: Own-race Teacher Effect on Achievement, Compliers

	Mathematics	Reading	Word Recognition	Listening
First Grade				
d_{ik}	6.14**	9.93***	10.40***	5.51***
	(2.60)	(3.01)	(3.58)	(1.90)
d_{i1}	12.24***	5.56**	5.03	4.85***
	(2.69)	(2.65)	(3.29)	(1.71)
Second grade				
d_{ik}	7.89***	7.96***	10.45**	9.79***
	(2.52)	(2.81)	(4.06)	(2.39)
d_{i1}	-0.02	-3.91	-5.21	-1.71
	(3.16)	(3.47)	(4.96)	(2.44)
d_{i2}	6.40**	1.78	-0.46	0.35
	(2.99)	(3.01)	(3.57)	(2.12)
Third Grade				
d_{ik}	6.30**	16.33***	18.57***	6.16*
	(3.12)	(2.95)	(4.28)	(3.18)
d_{i1}	5.02*	0.60	-1.88	-1.70
	(2.96)	(3.13)	(3.70)	(2.60)
d_{i2}	4.55	-2.07	-3.34	3.18
	(3.11)	(3.00)	(3.77)	(2.68)
d_{i3}	-2.85	-2.81	3.02	5.48**
	(2.79)	(2.97)	(3.69)	(2.60)

Note: the table contains the reduced form coefficient estimates on the d_{ig} variables in the system of reduced form equations described in Section 2 using the subpopulation of those that comply with their assigned class type. Between 5% and 20% of the sample is lost in this estimation. * denotes statistical significance at the 10% level, ** the 5% level, and *** the 1% level. Standard errors clustered at the level of the classroom are given in parentheses. Scaled test scores are used as the response variable.

Table 10: Own-race Teacher Effect on Achievement, Black Students

	Mathematics	Reading	Word Recognition	Listening
Kindergarten				
d_{ik}	5.31	3.36	2.49	6.14**
	(4.60)	(3.02)	(3.26)	(2.69)
First grade				
d_{ik}	5.33	3.03	4.13	1.76
	(3.59)	(3.20)	(4.73)	(2.58)
d_{i1}	10.92**	9.67**	7.87	7.60**
	(4.77)	(4.53)	(5.35)	(3.61)
Second grade				
d_{ik}	4.37	5.23	5.71	3.19
	(3.62)	(3.44)	(3.90)	(2.92)
d_{i1}	-3.01	-2.47	-2.84	1.03
	(4.90)	(5.29)	(6.46)	(4.24)
d_{i2}	4.03	1.02	-1.81	3.34
	(5.18)	(4.21)	(4.69)	(3.86)
Third Grade				
d_{ik}	2.50	7.74**	6.52	5.91
	(3.15)	(3.13)	(3.97)	(3.73)
d_{i1}	-2.25	5.07	12.27***	-5.82
	(4.14)	(4.38)	(4.26)	(6.34)
d_{i2}	8.78*	1.74	-2.82	7.63
	(4.99)	(4.34)	(4.46)	(6.15)
d_{i3}	-6.50	-0.75	-1.45	3.77
	(4.69)	(5.16)	(4.68)	(5.46)

Note: the table contains the reduced form coefficient estimates on the d_{ig} variables in the system of reduced form equations described in Section 2 for blacks only. Approximately 65% of the sample is lost in this estimation. * denotes statistical significance at the 10% level, ** the 5% level, and *** the 1% level. Standard errors clustered at the level of the classroom are given in parentheses. Scaled test scores are used as the response variable.

Table 11: Own-race Teacher Effect on Achievement, no teacher's aides

	Mathematics	Reading	Word Recognition	Listening
Kindergarten				
d_{ik}	13.08***	5.64**	4.32***	8.24***
	(2.74)	(1.79)	(2.03)	(2.57)
First grade				
d_{ik}	3.66	8.18**	10.87**	4.21
	(4.22)	(3.66)	(4.27)	(2.95)
d_{i1}	10.66**	1.54	-0.44	3.88
	(4.38)	(3.32)	(3.48)	(2.90)
Second grade				
d_{ik}	-0.42	1.81	10.47*	9.05***
	(3.63)	(4.12)	(5.52)	(3.37)
d_{i1}	7.14*	-0.8	1.79	0.51
	(3.99)	(3.80)	(4.92)	(3.50)
d_{i2}	8.68**	1.13	-4.02	-2.37
	(3.54)	(3.07)	(4.14)	(3.18)
Third Grade				
d_{ik}	-1.83	4.95	9.93*	1.26
	(4.06)	(3.72)	(5.64)	(2.74)
d_{i1}	8.87**	6.35	13.94**	6.31**
	(3.86)	(4.06)	(6.01)	(2.70)
d_{i2}	1.1	0.93	1.1	1.95
	(4.31)	(3.12)	(4.40)	(3.08)
d_{i3}	-5.59*	-5.64	-8.07	-0.54
	(3.29)	(4.17)	(5.43)	(2.70)

Note: the table contains the reduced form coefficient estimates on the d_{ig} variables in the system of reduced form equations described in Section 2 for the subsample that excludes students that ever attend a regular class with a full-time teacher's aide. Between 35% and 53% of the sample is lost in this estimation. * denotes statistical significance at the 10% level, ** the 5% level, and *** the 1% level. Standard errors clustered at the level of the classroom are given in parentheses. Scaled test scores are used as the response variable.

5.2. Bounds Analysis

Because the achievement tests were not all administered on the same day, students may at times not be present to complete them. If these missing values are missing at random for all students, the only consequence is decreased precision when estimating the regression coefficients. Examining the rates of absenteeism for kindergarten, first grade, and third grade[9], the rates are 6.4%, 5%, and 4.6% respectively. I compare these figures to the various absenteeism rates for the tests in various grades, displayed below in Table 12. The rates

Table 12: Test Absenteeism in Project STAR

	Total	Absent	Absenteeism
Kindergarten			
Mathematics	6325	454	7.18%
Reading	6325	536	8.47%
Word recognition	6325	474	7.49%
Listening	6325	488	7.72%
First grade			
Mathematics	6829	231	3.38%
Reading	6829	434	6.36%
Word recognition	6829	857	12.55%
Listening	6829	273	4.00%
Second grade			
Mathematics	6840	775	11.33%
Reading	6840	763	11.15%
Word recognition	6840	493	7.21%
Listening	6840	797	11.65%
Third grade			
Mathematics	6802	725	10.66%
Reading	6802	802	11.79%
Word recognition	6802	453	6.66%
Listening	6802	728	10.70%

Note: these figures correspond to the number of students absent for the various Stanford Achievement Tests given at the end of each grade. Author's calculations. The numbers are from the whole Project STAR sample.

of test absenteeism seem elevated compared to the absenteeism rates at every grade level. Because of this, it is possible that students are avoiding testing dates on purpose. To examine the limits to which this can affect the results of this paper, I perform a Horowitz and Manski (2000) assumption-free bounds analysis to determine the upper and lower limits on the reduced form regression parameters[10].

[9] Attendance data is not available for the second grade in the public release version of Project STAR.

[10] Because the response variable does not have finite support, the bounds technically carry the assumption that the observed imputed values are the limits that would be observed with an infinite sample size.

The lower bound of the reduced form parameter estimates are created by replacing the missing values of the response variable with their minimum value when $d_{ig} = 1$, and with their maximum value when $d_{ig} = 0$. Similarly, the upper bound of the reduced form parameter estimates are derived by replacing the missing values of the response variable with their maximum value when $d_{ig} = 1$, and with their minimum value when $d_{ig} = 0$. The imputed values I enter into the missing data are the respective test scores within school and within class type. The results of this bounds exercise are displayed in Table 13. While the

Table 13: Horowitz and Manski Assumption-Free Bounds

	Mathematics		Reading		Word Recognition		Listening	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Kindergarten								
d_{ik}	-2.39	26.15‡	-5.57‡	17.47‡	-5.84‡	17.23‡	-0.98	19.71‡
	(2.84)	(2.61)	(2.01)	(1.81)	(2.12)	(2.10)	(2.06)	(2.23)
First grade								
d_{ik}	-8.64‡	19.09‡	-0.96	22.10‡	-0.26	18.99‡	-4.80‡	14.42‡
	(2.56)	(2.64)	(2.74)	(2.57)	(3.61)	(3.31)	(1.83)	(1.84)
d_{i1}	8.52‡	15.64‡	-2.72	9.66‡	-22.08‡	24.24‡	1.73	8.82‡
	(2.60)	(3.01)	(2.67)	(2.87)	(3.91)	(3.44)	(1.80)	(1.82)
Second grade								
d_{ik}	-6.06†	20.40‡	-3.59	17.77‡	-3.67	13.73‡	-1.71	17.13‡
	(2.94)	(2.54)	(3.02)	(2.97)	(3.91)	(3.79)	(2.51)	(2.47)
d_{i1}	-3.71	4.36	-8.12†	3.59	-24.32‡	21.12‡	-2.25	5.21†
	(3.39)	(2.87)	(3.27)	(3.44)	(4.34)	(4.14)	(2.49)	(2.40)
d_{i2}	1.54	10.10‡	-1.03	8.42‡	-2.36	5.50	-4.32*	4.17*
	(3.25)	(2.97)	(3.21)	(3.07)	(3.66)	(3.52)	(2.43)	(2.45)
Third Grade								
d_{ik}	-4.41	23.67‡	5.39*	29.28‡	4.43	23.68‡	-2.62	15.14‡
	(3.75)	(5.48)	(2.77)	(5.05)	(3.71)	(3.95)	(3.07)	(4.56)
d_{i1}	-5.02	10.28†	-6.57†	10.64‡	-18.11‡	27.95‡	-8.36‡	5.10*
	(3.24)	(3.98)	(2.78)	(3.60)	(3.95)	(3.99)	(3.06)	(2.92)
d_{i2}	1.78	4.35	-2.85	0.85	-3.86	5.59*	0.31	2.98
	(3.02)	(4.27)	(2.74)	(4.18)	(3.25)	(3.23)	(2.94)	(3.95)
d_{i3}	-11.22‡	2.54	-6.74†	4.23	-2.28	1.43	-0.92	10.96‡
	(3.28)	(3.00)	(3.12)	(2.93)	(3.62)	(3.62)	(2.94)	(3.07)

Note: the table displays Horowitz and Manski (2000) bounds for censored outcome data. * denotes statistical significance at the 10% level, † the 5% level, and ‡ the 1% level. Standard errors clustered at the level of the classroom are given in parentheses. Scaled test scores are used as the response variable.

assumptions used to create the bounds are quite weak, they are still informative in this case. Throughout the table we see that, in general, early exposure to a same-race teacher appears

Nonetheless, I use the term assumption-free to refer to the specific technique in use here. For a discussion of the limitations of this sort of analysis, see Lee (2009).

to matter more. Most of the bounds lean heavily positive, and even the lower bounds of the own race teacher effects are either imprecise or only mildly negative compared to the magnitude of the upper bounds. The upper bounds also tend to be much more precise. One particularly interesting coefficient is the precisely estimated positive effect of an own-race teacher in first grade in the *lower* bound of the first grade mathematics test score. The results of this exercise give strong credence to the theory that teachers increase the academic achievement of their students that share their racial heritage.

6. Policy Discussion

I have thus far shown 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. Using the estimates of the DATT from Table 7, the gains from an own-race teacher appear to be moderate at best. In kindergarten, an own-race teacher increases the mean mathematics test score by 2.3%. First grade students with a teacher of the same race in both kindergarten and first grade obtain a test score gain of 1.7% in reading. Continuous treatment in second grade yields a 1.3% increase in the word recognition test score. In third grade, continuous treatment results in a test score increase of 1.6% for listening, and 3.4% in word recognition. Overall, these benefits are low to modest in magnitude.

The positive influence of a teacher of the same race on student achievement may help explain a small but non-trivial part of the racial test score gap between black and white students, since black students are far less likely to be matched with an own-race teacher compared to white students. Table 14 below displays the data concerning the racial test score gap in the Project STAR data, where the figures are in standard deviation units[11]. The raw gap for math is much smaller than what is typically seen in the literature, while the raw gap in reading is only slightly less. Including student and teacher covariates does not appreciably change the gap in math but decreases it considerably in reading. Augmenting the model further with an own-race teacher variable moderately narrows the racial gap for mathematics, and gives a drop of roughly half that reduction in reading. Overall, accounting for racial matches appears to explain a non-trivial portion of the gap in test scores.

Since most 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, 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. However, 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 be successful, it may result in a lower average quality of teachers from the underrepresented races if we assume that the highest quality

[11] School fixed effects are included in the adjusted gaps to account for the fact that schools have a high level of racial segregation; since schools whose student bodies are white are much more likely to have own-race teacher matches, the contribution to the same-race teacher gap may be overestimated if this is not controlled for.

Table 14: 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	

Note: 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.

teachers are hired first (and a higher average quality for the majority race teachers). This later 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 that lacked experience and credentials in order to implement the policy (Imazeki, 2003). The test score gains from own-race matching are likely not to be significant enough to merit aggressive hiring of minorities for the purposes of raising test scores, even absent these other considerations.

References

[1] Bond, Timothy N. and Kevin Lang. (forthcoming) "The Evolution of the Black-White Test Score Gap in Grades K-3: The Fragility of Results," *Review of Economics and Statistics*.

[2] Cizek, Gregory J. (1995) "On the Limited Presence of African-American Teachers: An Assessment of Research, Synthesis and Policy Implications," *Review of Educational Research*, vol. 65: 78-92.

[3] Clotfelter, Charles T., Helen F. Ladd, and Jacob Vigdor. (2003) "Who teaches whom? Race and the distribution of novice teachers," *Economics of Education Review*, vol. 24: 377-392.

[4] Clotfelter, Charles T., Helen F. Ladd, and Jacob Vigdor. (2006) "Teacher-Student Matching and the Assessment of Teacher Effectiveness," *Journal of Human Resources*, vol. 41: 778-820.

[5] Clotfelter, Charles T., Helen F. Ladd, and Jacob Vigdor. (2009) "The Academic Achievement Gap in Grades 3 to 8," *Review of Economics and Statistics*, vol. 91: 398-419.

[6] 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*, vol. 45: 655-681.

[7] Dee, Thomas S. (2004) "Teachers, Race, and Student Achievement in a Randomized Experiment," *Review of Economics and Statistics*, vol. 86: 195-210.

[8] Dee, Thomas S. (2005) "A Teacher Like Me: Does Race, Ethnicity, or Gender Matter?" *American Economic Review Papers and Proceedings*, vol. 95: 158-165.

[9] 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*, vol. 92: 31-42.

[10] Ding, Weili, and Steven F. Lehrer. (2011) "Experimental estimates of the impacts of class size on test scores: robustness and heterogeneity," *Education Economics*, vol. 19: 292-252.

[11] DuMouchel, William H., and Greg J. Duncan. (1983) "Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples," *Journal of the American Statistical Association*, vol. 78: 535-543.

[12] Finn, Jeremy D., Jayne Boyd-Zaharias, Reva M. Fish, and Susan B. Gerber. (2007) "Project STAR and Beyond: Database Users Guide," HEROS, Incorporated, January 1.

[13] Frangakis, Costas E., and Donald B. Rubin. (1999) "Addressing Complications of Intention-to-Treat Analysis in the Presence of All-or-None Treatment-Noncompliance and Subsequent Missing Outcomes," *Biometrika*, vol 86: 365-379.

[14] Fruehwirth, Jane Cooley. (2013) "Identifying peer achievement spillovers: Implications for desegregation and the achievement gap," *Quantitative Economics*, vol 4: 85-124.

[15] Fryer, Roland G., and Steven D. Levitt. (2004) "Understanding the Black-White Test Score Gap in the First Two Years of School," *Review of Economics and Statistics*, vol. 86: 447-464.

[16] Hanushek, E.A., J.F. Kain, D.M. O'Brien, and S.G. Rivkin. (2005) "The market for teacher quality," Working Paper 11154. National Bureau of Economic Research, Cambridge, MA (February).

[17] Horowtiz, Joel L., and Charles F. Manski. (2000) "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data," *Journal of the American Statistical Association*, vol. 95: 77-84.

[18] . Class-Size Reduction and Teacher Quality: Evidence from California, in School Finance and Teacher Quality: Exploring the Connections, David Monk and Margaret Plecki, eds. (2003), 159-178.

[19] King, Sabrina H. (1993) "The Limited Presence of African-American Teachers," *Review of Educational Research*, 63: 115-149.

[20] Krueger, Alan B. (1999) "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics*, vol. 114: 497-532.

[21] Lechner, Michael, and Ruth Miquel. (2010) "Identification of the effects of dynamic treatments by sequential conditional independence assumptions," *Empirical Economics*, vol. 39: 111-137.

[22] Lee, David S. (2009) "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, vol. 76: 1071-1102.

[23] Mueller, Steffen. (2013) "Teacher experience and the class size effect - Experimental evidence," *Journal of Public Economics*, vol. 98: 44-52.

[24] Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. (2005) "Teachers, schools, and academic achievement," *Econometrica*, vol. 73: 417-458.

[25] Wooldridge, Jeffrey M. (2002) "Inverse probability weighted M-estimators for sample selection, attrition, and stratification," *Portugese Economic Journal*, vol. 1: 117-139.