Online Appendix for Racial Interaction Effects and Student Achievement by Jeffrey Penney

(dr.jeffrey.penney@gmail.com)

Identification and Threats to Validity

This appendix investigates technical issues related to identification and the various threats to validity. I discuss the former since randomization only occurs in the first period; I determine that the initial randomization suffices to identify the various treatment paths. I then turn to issues relating to threats to validity due to randomization, attrition, student sorting, and unobservable variables. I find that the randomization process was successful, and provide a formal mathematical proof that randomizing students across class types is equivalent to randomizing according to the student-teacher racial match. An investigation into the attrition behaviour of the sample reveals that it is likely nonignorable; this paper implements a correction for this feature of the data. Student sorting across schools by race may be a concern if these sorting patterns are also related to academic ability; for example, able black students primarily attending schools whose students are predominantly black. Examining these patterns, I find no evidence that student sorting by ability is taking place.

A.1. Identification

The coefficients of interest (those on the d_{ig} variables) in the system of equations 1, 2, 3, 4 and the derived causal effects of the treatment paths discussed in section 2 are identified using the experimental variation in Project STAR that was induced via the random assignment of both students and teachers to classrooms of different types. This variation provides clean identification of the causal effects of the various treatment paths because assigning students to a

particular class type for the duration of the experiment is the equivalent of assigning them to a particular treatment path with respect to racial matches because teacher race can vary within a class type assignment. For example, consider two students assigned to the small class treatment in kindergarten, each at a different school: one student may experience the treatment path $t(1,1,1,1)$ while the other may encounter the path $t(0,0,0,0)$. In short, the initial assignment to class type is the randomization mechanism for the treatment paths. Because of this, it is safe to claim that a form of weak dynamic conditional independence assumption (WDCIA) applies, that is, for all possible paths τ,

$$\tau \perp D_k | \, S_k = s_k$$

$$\tau \perp D_1 | \, S_1 = s_1, D_k = d_k$$

$$\tau \perp D_2 | \, S_2 = s_2, D_1 = d_1, D_k = d_k$$

$$\tau \perp D_3 | \, S_3 = s_3, D_2 = d_2, D_1 = d_1, D_k = d_k$$

where D_i is a random variable for treatment status in period i, S denotes a random variable of school attended, s represents the particular draw of a school S, and d a particular draw of D. Moreover, I also assume common support:

$$1 > P(D_k = 1 \mid S_k = s_k) > 0$$

$$1 > P(D_1 = 1 \mid S_1 = s_1, D_k = d_k) > 0$$

$$1 > P(D_2 = 1 \mid S_2 = s_2, D_1 = d_1, D_k = d_k) > 0$$

$$1 > P(D_3 = 1 \mid S_3 = s_3, D_2 = d_2, D_1 = d_1, D_k = d_k) > 0$$

$\forall d_i = D_i$ and $\forall s_i = S_i$, where $P(\cdot)$ denotes probability. Table 2 shows that there is indeed support over all possible treatment paths. More details about these technical requirements can be found in Lechner (2009).

A.2. Randomization

Previous studies have shown that classroom sorting according to ability and race occurs even within schools (Clotfelter, Ladd, and Vigdor 2005, 2006). This is a serious concern in research on racial matching and its effect on academic achievement, as sorting patterns can be so pronounced as to distort the direction (and magnitude) of the causal effect. 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.

The general consensus of the literature is that the randomization process of Project STAR was successful. Krueger (1999) shows that, examining within school variation along student characteristics, there is no evidence that randomization was unsuccessful across the different class types. Since the number of observable student characteristics of the student is rather scarce in the data, some skepticism still remained about whether randomization was indeed properly executed (Hanushek 2003); Chetty et al. (2011), who merged the Project STAR data with income tax records, found that the additional observables afforded through the merge were still balanced within class types.

While students and teachers were randomly matched within schools in the Project STAR experiment, the matching was performed according to class type and not racial match. What

follows is a formal proof that randomization into distinct groups (such as a classroom) effectively randomizes along other observable dimensions.

Proposition. *Let there be a pool of objects Y, whose n individual elements y have a characteristic b and whose probability distribution f is well defined. Let j be an arbitrary particular value of the characteristic b. Suppose Y is divided into a finite number of k distinct sets $Y_1, Y_2, \dots, Y_k$ such that $\cup_i Y_i = Y$ and each y is randomly allocated to a single subset. Then,* $\forall Y_i, pr_n(y(b = j)|y \in Y_i) \xrightarrow{a.s.} pr(y(b = j))$ *as* $n \to \infty$.

Proof. Since randomization into the k subsets of Y is independent of the value of b, $pr_n(y(b = j) \in Y_i) = pr_n(y \in Y_i)$. Therefore, $f_n(y|y \in Y_i) \xrightarrow{d} f(y)$ as $n \to \infty \, \forall i$, which implies $pr_n(y(b = j)|y \in Y_i) \xrightarrow{a.s.} pr(y(b = j)) \; \forall Y_i$. ■

Roughly speaking, randomly sorting objects from a group into different subgroups will give the same expected distribution of characteristics of the objects as the original group in each of the subgroups. Since there is no reason to believe that randomization into different class types leads to an asymptotically dissimilar distribution along the dimension of racial matches between students and teachers in the different classes, all that remains to verify is whether randomization failed in the finite sample. To examine this question, I run the following regression for every grade g and every current and future grade j:

$$samerace_{ig} = \beta_0 + \beta_1 small_{ij} + \beta_2 aide_{ij} + \theta_l + \varepsilon_i \qquad \text{(A.1)}$$

where $small_{ij}$ and $aide_{ij}$ are dummy variables taking the value of 1 if student i is assigned to a small class or a regular class with a teacher's aide in grade j respectively and 0 otherwise, and

$samerace_{ig}$ is a dummy variable for whether the student and the teacher are of the same race in grade g. The school fixed effect for school l is given by θ_l, whose inclusion in equation A.1 is required since randomization occurred within schools; standard errors were clustered at this level. If the coefficients are jointly significant in the contemporaneous regressions (where $g = j$), then there is evidence that being assigned to a certain class type is associated with a change in the probability of a racial match, which would indicate a possible failure in randomization. If the coefficients β_1 and β_2 are jointly significant for any case where $g < j$, then this means randomization may have failed in a dynamic fashion: past racial matches should not be indicative of current and future class assignments (see table A.1).

Examining the results, I do not find any cause to believe that randomization of students and teachers failed in any current or future grade along the dimension of race.[1]

A.3. Attrition

Attrition in the Project STAR data is considerable. Of the students who were initially enrolled in kindergarten, 48.9 percent of them have left the experiment by 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.

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

[1] Note that given the large number of tests, there was a 41 percent chance of at least one false rejection of the null of randomization if the null is true.

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 instrumental variable 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 is correlated with observable characteristics using a procedure developed by Becketti et al. (1988). I estimate the following regression equation:

$$A_{ik} = X_{ik}\beta_1 + L_i X_{ik}\beta_2 + \theta_l + \varepsilon_{ik} \quad \text{(A.2)}$$

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 A.2 are displayed in table A.2. Students who subsequently left the sample after kindergarten performed much worse in kindergarten compared to those who do not attrit. The only statistically significant input is free lunch status in the mathematics regression: students who receive a free lunch in kindergarten that later attrit showed lower scores in mathematics. Since the coefficients on the attrition interaction

variables are jointly significant in one regression, attrition due to observables is likely nonignorable. Encouragingly, there is no statistically significant difference in the effectiveness of the same-race teacher treatment on test scores in kindergarten between those who attrit and those who do not.

Because selective attrition due to observables may be present, $\sqrt{N}$ consistent estimates[2] may still be obtained through the use of inverse probability weights (Wooldridge 2002). Since I am allowing for heterogeneous treatment effects, they are required to consistently estimate the model parameters. I perform Dumouchel and Duncan (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 19.37 in all cases, $p<0.0001$); therefore, weighted estimates are required. The estimates throughout this paper make use of inverse probability weights.

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).

A.4. Student Sorting

Student sorting across schools is an additional challenge above and beyond the usual randomization and attrition concerns of Project STAR, because sorting patterns across schools could potentially bias the results, even though students were randomized within schools. For example, if high-ability white students tend to attend schools with a high percentage of white

[2] Here, I use N to denote sample size.

teachers, while low ability white students tend to enroll in schools with a lower percentage of white teachers, then the magnitude of the racial interaction effect may be overestimated. We can examine whether there is any evidence of across-school sorting by graphing the relationship between the percentage of white teachers and the Black-White test score gap at school entry. Here I define the black-white test score gap as the difference in the average test score in a school between whites and blacks, with positive numbers indicating the number of scaled score points in which whites are ahead. If high-ability students tend to attend schools staffed by teachers of the same race, and low-ability students tend to enroll in schools largely staffed by teachers of a different race, we would expect the black-white test score gap to be increasing as the percentage of white teachers in a school increases; that is, we expect a positive relationship between the percentage of white teachers in a school and the size of the Black-White test score gap.

Figure A.1 plots the relationships between staffing and racial test score gaps using scatterplots for test scores at school entry for mathematics and reading scores.[3] We see there appears to be no relationship between the composition of teachers and the black-white test score gap. Both figure panels hint at a slightly negative relationship, which is the opposite of what we would expect if students were sorting in a way to overestimate the racial interaction effect; in fact, there are many schools staffed entirely by white teachers in the sample where blacks are ahead of whites in reading scores. Naive ordinary least squares regressions show a negative relationship between percent of white teachers and the black-white test score gap for all three

[3] The findings for reading scores and word recognition scores are almost identical (the latter are omitted for reasons of space). Racially homogeneous schools are not included in figure A.1 because black-white test score gaps are not identified in these cases.

subjects, but none is statistically significant. Given this evidence, it appears unlikely that across-school sorting of students is driving the results.[4]

[4] The issue of across-school sorting of teachers is addressed in section 4 since it is a robustness check of the regression results.

References

Becketti, Sean, William Gould, Lee Lillard, and Finis Welch. 1988. The panel study of income dynamics after fourteen years: An evaluation. *Journal of Labor Economics* 6(4): 472-492.

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.

Clotfelter, Charles T., Helen F. Ladd, and Jacob Vigdor. 2005. Who teaches whom? Race and the distribution of novice teachers. *Economics of Education Review* 24(4): 377-392.

Clotfelter, Charles T., Helen F. Ladd, and Jacob Vigdor. 2006. Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources* 41(4): 778-820.

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* 78(383): 535-543.

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* 86(2): 365-379.

Hanushek, Eric A. 2003. The failure of input-based schooling policies. *Economic Journal* 113(485): F64-F98.

Horowitz, Joel L., and Charles F. Manski. 2000. Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association* 95(449): 77-84.

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

Lechner, Michael. 2009. Sequential causal models for the evaluation of labor market programs. *Journal of Business and Economic Statistics* 27(1): 71-83.

Wooldridge, Jeffrey M. 2002. Inverse probability weighted M-estimators for sample selection, attrition, and stratification. *Portuguese Economic Journal* 1(2): 117-139.

Table A.1. Tests of Randomization

	Kindergarten	Grade 1	Grade 2	Grade 3
Kindergarten	0.9168			
Grade 1	0.9236	0.2623		
Grade 2	0.8049	0.3977	0.8031	
Grade 3	0.9107	0.5446	0.8586	0.6393

Note: The table contains the p-values of the test of joint significance of the coefficients β_1 and β_2 in regression A.1 where the column is the current grade i and the rows indicate current or future grades j.

Table A.2. Test for Attrition Based on Observables

Variable	Mathematics	Reading	Word Recognition
Attrition dummy	-15.67**	-13.36**	-11.87**
	(4.54)	(2.67)	(3.35)
Attrition dummy interacted			
Student and teacher are same	-4.67	0.27	-0.92
	(2.65)	(1.67)	(2.09)
Small class	-3.06	-0.18	-1.39
	(2.92)	(1.92)	(2.22)
Regular with aide class	3.93	2.80	2.30
	(2.66)	(1.76)	(2.09)
Student receives free lunch	-5.35*	-0.62	0.50
	(2.29)	(1.51)	(1.77)
Years of experience	-0.02	-0.21	-0.41
	(0.65)	(0.40)	(0.53)
Years of experience^2	0.01	0.01	0.02
	(0.03)	(0.02)	(0.02)
Teacher has a graduate degree	-1.25	0.68	1.31
	(2.48)	(1.68)	(1.99)
Teacher is black	3.22	2.22	1.16
	(3.22)	(2.11)	(2.56)
F-test on variables in β_2:			
all variables	0.0000	0.0000	0.0000
interaction variables only	0.0317	0.6398	0.8188
constant only	0.0006	0.0000	0.0005

Notes: The table contains the coefficients on β_2 in regression A.2. 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.
*Statistical significance at the 5% level; **statistical significance at the 1% level.

Figure A.1. Student Sorting by Teacher Race

Notes: The figures display scatterplots of the black-white test score gap in a school (defined as the number of points whites are ahead on average) at kindergarten entry and the percentage of white teachers in the school participating in the Project STAR program for kindergarten through third grade. If good students of a given race had a propensity to attend schools largely staffed by teachers of their own race, we would see a pronounced positive slope on the fitted regression lines.