Another Look at the New York City School Voucher Experiment

Alan B. Krueger
Princeton University and NBER

and

Pei Zhu
Princeton University

April 2003

Another Look at the New York City School Voucher Experiment

ABSTRACT

This paper reexamines data from the New York City school choice program, the largest and best implemented private school scholarship experiment yet conducted. In the experiment, low-income public school students in grades K-4 were eligible to participate in a series of lotteries for a private school scholarship in May 1997. Data were collected from students and their parents at baseline, and in the Spring of each of the next three years. Students with missing baseline test scores, which encompasses all those who were initially in Kindergarten and 11 percent of those initially in grades 1-4, were excluded from previous analyses of achievement, even though these students were tested in the follow-up years. In principle, random assignment would be expected to lead treatment status to be uncorrelated with all baseline characteristics. Including students with missing baseline test scores increases the sample size by 44 percent. For African American students, the only group to show a significant, positive effect of vouchers on achievement in past studies, the difference in average follow-up test scores between the treatment group (those offered a voucher) and control group (those not offered a voucher) becomes statistically insignificant at the .05 level and much smaller if the full sample is used. In addition, the effect of vouchers is found to be sensitive to the particular way race/ethnicity was defined. Previously, race was assigned according to the racial/ethnic category of the child's mother, and parents who marked "other" and wrote in Black/Hispanic were typically coded as non-Black and non-Hispanic. If children with a Black father are added to the sample of children with a Black mother, the effect of vouchers is small and statistically insignificant at conventional levels.

Alan B. Krueger
Economics Department
Princeton University
Princeton, NJ 08544
(609) 258-4046
akrueger@princeton.edu

Pei Zhu
Economics Department
Princeton University
Princeton, NJ 08544
(609) 258-5694
pzhu@princeton.edu

Now that the Supreme Court has ruled in the *Zelman* case that public funds may be used to support vouchers to enroll children in private religious schools, many states, school districts and parents will seriously consider the desirability of school vouchers. This decision naturally depends on many factors, not least of which is whether vouchers are likely to raise student achievement. The best currently available evidence on the effect of school vouchers on students' performance is from a series of three randomized experiments conducted in Washington, D.C., Dayton, OH and New York City by Paul Peterson and his collaborators. This paper reexamines evidence from the New York City voucher experiment, which was conducted by Mathematica Policy Research (MPR) and the Program on Education Policy and Governance at Harvard University.

The New York City experiment was selected because it is the only one of the three experiments for which data have been made available to outside researchers so far. Two additional reasons argue for a detailed evaluation of the New York experiment, however. First, the New York experiment is the best documented of the three experiments, and had the lowest attrition rate, highest voucher take-up rate, and largest sample size. Second, New York is the only one of the three cities to show significant gains in test scores for voucher recipients relative to non-recipients for African American students at the conclusion of the experiment.[1] In all three experiments, there is no significant difference in student performance between those offered a voucher and the control group for other racial and ethnic groups, or overall.

---

[1] See Howell and Peterson (2002), Table 6-3. In Washington, vouchers had a statistically insignificant, negative effect on Black students' scores after three years; in Dayton, vouchers had a positive effect that was statistically insignificant at the 0.10 level holding constant family background controls (but significant at the 0.10 level without family background controls) in the second and final year of the experiment.

The New York City school choice experiment worked as follows.[2] In February

1997, the School Choice Scholarships Foundation (SCSF), a private foundation, offered

1,300 scholarships worth up to $1,400 a year for three years to children from low-income

families (i.e., qualified for free lunch) who were enrolled in Kindergarten through fourth

grade in New York City public schools. Some 11,105 eligible students applied for a

scholarship between February and late April 1997. Recipients were selected in a series of

lotteries in May 1997, and began attending private schools the next fall. Mathematica

randomly selected the students offered vouchers (subject to the SCSF requirement that 85

percent of recipients be from public schools in the bottom half of the city-wide test score

distribution) and a control group from the eligible applicants. About three quarters of the

students who were offered vouchers used them in at least one year; these students

overwhelmingly attended religious schools.[3] Information from the students and their

parents was collected prior to the lottery and in the spring of each of the ensuing three

years. Base weights were constructed so the students in the sample were representative

of the pool of eligible applicants (which had 70 percent from schools with below-median

scores), and the weights were subsequently adjusted for nonresponse each year.

Students were given the Iowa Test of Basic Skills (ITBS) at baseline and in the

spring of each of the three follow-up years. A decision was made not to test the cohort of

Kindergarten students applying for scholarships for first grade at baseline, however.

(Henceforth, the five cohorts of students in different grades will be referred to by their

grade level at baseline.) The Kindergarten cohort was nonetheless given follow-up ITBS

tests, along with other students, when they were in grades 1, 2 and 3. In addition, about

---

[2] See Mayer, Peterson and Myers, et al. (2002) and Hill, Rubin and Thomas (2000) for further details.
[3] Only 11 percent of the controls attended private school in at least one year.

11 percent of students initially in grades 1-4 lacked baseline scores.[4] These students were also given the ITBS in the three follow-up waves. The sample weights do not attempt to adjust for missing baseline scores.

A contribution of our paper is that we include students with missing baseline scores in much of our analysis. Previous analyses of achievement omitted students with missing baseline test data.[5] Howell and Peterson (2002), for example, report, "A handful of additional families were offered vouchers, but they were not included in the evaluation for lack of baseline [test] information." Because of random assignment, however, estimates are unbiased even without conditioning on baseline information, so there is an efficiency loss from excluding these students. For the subsample with baseline scores, omitting the baseline score only trivially affects the estimated treatment effect, as one would expect with random assignment. Including students with missing baseline test data increases the sample size by 44 percent in the third and final follow-up year; nearly 30% of those with missing baseline scores were in grades 1-4 when the experiment started and should have had baseline scores. An argument can be made that including students with missing baseline scores – both those in the Kindergarten cohort and those in the other cohorts – is desirable because the weights make no provision for sample exclusion due to missing baseline scores, and, more importantly, because using a sample that encompasses more grade levels enhances the generalizability of the results.

---

[4] This is in addition to students who had scores of 0 on the baseline test – of the 1,851 students with baseline scores, 199 (10.7%) received a zero score in reading and 324 (17.5%) received a zero score in math; 97 (5.2%) received a zero on both exams.

[5] When Mayer, Peterson and Myers, et al. (2002) examine outcomes such as parental satisfaction, however, they include observations on students without baseline test data. It is unclear whether Howell and Peterson (2002) include or exclude students with missing baseline tests when they study parental responses. Their Table 2-3 reports that they include students entering grades 1-4, but that apparently is a misprint, as their sample includes students entering grades 2-5. Nevertheless, we could only replicate some of their results based on the parental survey if we include the Kindergarten cohort.

For African American students, the estimated effect of being offered a voucher is much weaker if students with missing baseline scores are included in the analysis. In the third follow-up year, for example, the effect of being offered a voucher on composite test scores is 2.78 percentile points (t = 1.64) if baseline test scores are dropped from the model and the larger sample is used; controlling for baseline covariates other than test scores, the effect is 2.08 points with a t-ratio of 1.24.[6] For comparison, Mayer, Peterson and Myers report an effect of 5.50 points with a t-ratio of 3.42. Results are also weaker in the first two follow-up years if students with missing baseline scores are included. These findings raise doubts about the robustness of earlier findings of a significant positive effect of offering vouchers on the test scores of African American students.

In the next section, we discuss and evaluate the random assignment procedures in more detail. In the following section we explore the sensitivity of the results to including and excluding students with missing baseline scores, and examine differences in the treatment effect across cohorts.

Although results for the larger sample encompassing students with and without baseline scores cast doubt on the inference that vouchers had a positive impact on Black students' test scores, we explore reasons why vouchers might have been more effective at raising scores for Black students than for other students in the grade 1-4 cohorts. Results presented in Section 4 suggest that differential characteristics of the initial public schools Black and non-Black students attended are *not* responsible for any differential effect by race, even in the sample analyzed by Howell and Peterson.

---

[6] These estimates also control for 30 dummies indicating the original strata students were placed in for random assignment and utilize bootstrap standard errors. We weighted the underlying data using the revised follow-up weights that Mathematica provided on April 3, 2003.

4

The data suggest that race itself independently affects the gain from vouchers in the Howell-Peterson sample. This leads us to examine the particular definition of race that was used. There is no universally accepted definition of race. Mathematica assigned students to a racial/ethnic group based on a single question on the parental survey that asked respondents to *select only one* of the following categories for the mother or female guardian: Black/African American (non-Hispanic), White (non-Hispanic), Puerto Rican, Dominican, Other, etc. Students were assigned the mother's race/ethnicity, regardless of the father's race or ethnicity. Thus, if a student is reported as having a Black mother and a Hispanic father, he or she was classified as Black. If the same student had a Black father and Hispanic mother, however, he or she would have been classified as Hispanic. Arguably, father's race is also relevant. If we augment the sample to include students for whom *either* parent is classified as Black/African American and those whose parents marked "Other" and wrote in Black/Hispanic in the blank provided, the effect of offering a voucher on the composite score in year three falls to 1.44 points with a t-ratio of 1.01. So in the broader sample the effect of offering vouchers to students with a Black parent is small and statistically insignificant.

Throughout the paper, we focus mainly on intent-to-treat estimates; that is, the impact of offering students a voucher on their test performance, as opposed to the effect of *attending private school* on test performance. We focus on intent-to-treat estimates because offering a voucher – as opposed to compelling students to switch to private school -- is the policy decision that is most relevant, and because there is a cleaner statistical interpretation of the intent-to-treat estimates in this case. Nevertheless, the effect of attending a private school for varying lengths of time is also of interest. In

Section 5 we present Instrumental Variables results that estimate the impact of the *number of years* in private school on student achievement. These results differ from those emphasized by Howell and Peterson (2002), who examine the effect of attending private school for three full years, and implicitly make strong assumptions about the effect of switching to private school on achievement for those who attend private school for fewer than three years.

1. Randomization and Data

The procedures used to randomly assign students to treatment and control status, and select control group members for follow up, are described in Hill, Rubin and Thomas (2000). Because multiple children from many families applied for scholarships, and it was desired to assign all family members to the same treatment status, students were assigned to control and treatment groups in a lottery in which *families* were the unit of observation. Two methods of random assignment were used.

Briefly, for students from 1,000 families a Propensity Match Pairs Design (PMPD) was used, and for students from 960 families a Stratified Block design was used. The PMPD method, which introduces considerable complexity to the design, was used because in the first lottery many more potential control group members were available to be followed up than was money to follow them up. Rather than select a random sample of the controls to follow up, it was decided that it would be more efficient to follow up the subset that, in some sense, is most alike to treatment group members. Consequently, after the treatment group was randomly selected, the members of the control group who were followed up were selected by estimating a propensity score model to identify those with attributes that were closest to members of the treatment group.

6

According to Hill, Rubin and Thomas (2000), variables used in this model included, in order of importance: family size, a dummy indicating above versus below median schools, grade level, and initial test scores.[7] Students with missing data were also included in the selection. Once the propensity score model was estimated, a matched pair for each treatment group family was selected by choosing the nearest available neighbor Mahalanobis match from among those with propensity scores close to that of each treatment group member.

The Stratified Block design was much simpler. Samples of screened applicants were invited to participate in four sessions at which baseline data were collected and ITBS tests were administered. In these lotteries, by design approximately 85 percent of the invitees were from schools with below the city median test score. Treatments and controls were randomly assigned in four lotteries, and a random sample of participants were included in the follow-up sample. Each of these lotteries constitutes a block.

One can define 30 mutually exclusive "random assignment strata": 5 lottery blocks (PMPD block plus 4 stratified blocks) x 2 school types (above and below city median test) x 3 family size groups (1, 2 or 3 or more students). Assignment is random within these strata. After the lotteries were held, Mathematica discovered that some families had misreported their family size and were thus placed in the "wrong" strata; revised strata were created with the latest family size information. Nevertheless, assignment to treatment status is random within the *original* strata that were actually used to apportion the sample at the time of random assignment. Howell and Peterson (2002)

---

[7] Other variables included, in order of importance, were: ethnicity, mother's education, participation in special education, participation in a gifted and talented program, language spoken at home, welfare receipt, food stamp receipt, mother's employment status, educational expectations, number of siblings, and an indicator for whether the mother was foreign born.

and Mayer, Peterson and Myers, et al. (2002), however, controlled for dummies indicating membership in the revised strata, not the actual strata used to make assignments. Unless otherwise specified, our results condition on the *actual* strata that were used when random assignments were made. Fortunately, the choice of original or revised strata has relatively little impact on the results.

Because assignments were over families, not students, and children from the same family tend to have correlated outcomes, we compute bootstrap standard errors that use families instead of individual students as the unit for resampling to allow for dependence across family members.[8] (Forty-six percent of students in the sample had at least one sibling in the sample as well.) Moulton (1990) provides a nice illustration of inference problems that can arise from ignoring correlated errors.

Table 1 reports the mean of several baseline characteristics for the treatment and control groups for the full sample, separately for Black and Latino students, and disaggregated by cohort for Black students. Because random assignment was implemented within strata, regressions were estimated to condition on the 30 original randomization strata, and conditional treatment-control differences and t-tests are reported as well. For the overall sample, the results indicate small and statistically insignificant differences between the treatment and control groups. For example, the conditional t-test for the difference in the mean baseline composite test score between treatments and controls is only -0.65. Mayer, Peterson and Myers, et al. (2002) present similar results. Hill, Rubin and Thomas (2000) report treatment-control differences by PMPD versus Stratified Block strata, and find slightly better balance in the PMPD strata,

---

[8] The sample was drawn with replacement 10,000 times to compute the bootstrap standard errors.

but in both cases there are not systematic differences between the treatments and controls.

In the first-year follow-up report, however, Peterson, Myers and Howell (1998) reported highly statistically significant differences in *baseline* test scores between the treatment and control group, with control group members scoring significantly higher on both the math and reading exams. Evidently, this was a result of inaccurate weights. The baseline weights did not adjust correctly for the size of the underlying assignment strata.[9] In the subsequent reports, the baseline weights were revised "to include post-stratification adjustments," and the baseline differences in test scores were no longer statistically significant. The complexity of the design of the experiment undoubtedly contributed to the initially inaccurate baseline weights, as the average discrepancy between the initial and corrected baseline weights was 316 percent for the PMPD strata and only 24 percent for the Stratified Block strata.[10]

Because of a small inconsistency in the baseline weights that we pointed out to MPR, another mistake was discovered in the baseline weights. Mathematica subsequently provided us with revised baseline and follow-up weights. The follow-up weights adjust the baseline weights for nonresponse, by using estimates from a logit equation to predict nonresponse from observed characteristics; a stepwise procedure was used to help select covariates in the logit model. Unless otherwise noted, we utilize the revised weights that Mathematica recommended on April 3, 2003 throughout this paper.

The results in Table 1 suggest that the assignment groups were well balanced, as one would expect with random assignment. One exception, however, is the oldest cohort of African American students. For this group, at baseline the treatments' mean score is

---

[9] Email communication from David Myers of Mathematica, February 5, 2001.
[10] In particular, the weight accorded to the control PMPD strata increased considerably.

higher than the controls', and the treatments are more likely to come from higher income families with better educated mothers. For African Americans as a whole, however, the differences in baseline characteristics are typically insignificant.

The Addendum to the table reports follow-up information: the mean test score in year three and the proportion of students who ever attended a private school (during the three follow-up years of the experiment). Eighty-one percent of Black students offered a voucher used it in at least one year, compared with 10 percent of those not offered a voucher, leading to a 71 percentage point increase in the likelihood that those offered a voucher enrolled in private school. The increase in private school enrollment was smaller for the Kindergarten and third grade cohorts, but still in excess of 60 percentage points. Test scores are considered extensively below.

1.1 Precision

The PMPD design was intended to reduce sampling variance and improve the balance between treatments and controls. Did the precision of the estimates improve? Table 2 reports separate estimates of the treatment effect and associated standard errors for the two methods of random assignment. The treatment effect in this instance is the average impact of being offered a voucher on students' test scores, measured in National Percentile Ranks (NPR's). Formally, the treatment effect, α, is computed by estimating the following regression model for each follow-up year:

(1) $\qquad Y_{if} = \alpha\, Z_f \;+\; \delta^j_{if} + \varepsilon_{if},$

where Y_{if} is the composite test score (that is, average NPR on reading and math), Z_f is a dummy variable that equals one if the student was offered a voucher and zero if not, δ^j_{if} is a fixed effect for the randomization strata ($j = 1, \ldots, 30$), and ε_{if} is an error term that is

possibly correlated among members of the same family. The subscript i indexes students

and f indexes families. The strata fixed effects are controlled for by including 30 dummy

variables indicating each original randomization stratum.

We also present another set of estimates in which the regression model is

augmented to control for baseline math and reading test scores, using as a sample the

subset of students who have baseline test scores. Formally:

(2) $Y_{if} = \forall Z_f + \gamma_M M_{if} + \gamma_R R_{if} + \delta^j_{if} + \varepsilon'_{if,}$

where M is the math NPR score, R is the reading NPR score, and the γ's are coefficients.

Both equations (1) and (2) provide unbiased estimates of the effect of offering a voucher;

hence, they are specified with the same parameter, α.[11]

Estimates of α for the full sample are in Panel A, and for African Americans in

Panel B. The rows of the table indicate whether the underlying regression model

controlled for baseline test scores or omitted baseline scores and used a larger sample.

The second to last column reports the standard error in the Stratified Block design

relative to the standard error in the PMPD design. Because the sample sizes differ

slightly, in the last column the relative standard errors are scaled by the ratio of the

square root of the relative sample sizes.[12]

For the full sample, in most cases the standard error of the estimated treatment

effect is about 10 percent larger in the Stratified Block design than in the PMPD design.

The adjustment for sample size has little effect on this ratio. The treatment effect is

---

[11] See R.A. Fisher (1951) for a thoughtful discussion of whether baseline covariates (e.g., crop yields)
should be held constant in agricultural experiments. Recognizing that estimates with or without baseline
covariates are unbiased, he concludes that in practice the gain in precision from adding covariates is not
worth the extra cost of collecting the data. Coincidentally, the magnitudes are similar to the voucher
experiment.

insignificantly different from zero under either design. Thus, there is a small gain in power from the PMPD design.

For the subsample of Black students the standard errors are not consistently smaller in the PMPD design than in the Stratified Block design. This result is somewhat surprising because, as Hill, Rubin and Thomas point out, analyses of subgroups are expected to have more power in the PMPD design because matching should lead to a more equal representation of subgroups in the treatment and control groups.

One could question whether the gain in precision from the PMPD random assignment procedure was worth the cost of added complexity. For practical purposes, the difference in precision between the two designs is fairly small, even in the full sample: in year 3, for example, the width of a 95% confidence interval increases trivially, from +/-3.1 points in the PMPD design to +/-3.5 points in the Stratified Block design. If the complexity of the experimental design contributed to the incorrect initial inference about the baseline difference in test scores (because of inaccurate weights), then in this case the PMPD design would seem to have been hardly worth the gain in power. This likelihood seems to us to underscore Cochran and Cox's (1957) advice: "A good working rule is to use the simplest experimental design that meets the needs of the occasion."

2. Sensitivity of Earlier Results

Next we explore the sensitivity of the estimated impact of offering a voucher to controlling or not controlling for baseline scores, and to including the cohort of students who were in Kindergarten when the experiment began. The first set of columns in Tables

---

[12] At baseline, the sample size was 1,341 for the PMPD sample and 1,325 for the Stratified Block sample. The sample sizes in Table 2 differ because of missing follow-up test data. If attrition is endogenous to the design, then the unadjusted comparison is more appropriate.

3a and 3b provide estimates of equation (2): a regression of the test score NPR on a dummy indicating whether the student was randomly selected to be offered a voucher, baseline test scores, and 30 dummy variables indicating the actual stratum the family was allocated to for random assignment. The coefficient (α) on the voucher offer dummy, also known as the intent-to-treat (ITT) estimate, is presented in the table, along with its standard error and t-ratio.

In Table 3a we use the same sample, follow-up weights and strata definition as Mayer, Peterson and Myers, et al., Tables 23-25. Our results exactly replicate theirs. As they found, for Black students the results indicate significantly higher test scores for those offered a voucher than for the controls. For Latino students the voucher effect is negative but statistically insignificant.

The first estimates in Table 3b correct two minor flaws. First, as mentioned, the previous researchers did not control for the actual strata that students were placed in when random assignments were made; instead, they controlled for strata based in part on revised family size information. To exploit the experiment, it is more appropriate to control for the strata that were actually used at the time of random assignment, so we use the original strata in Table 3b and elsewhere in the paper (except Table 3a). Second, we weight the data using revised follow-up weights that Mathematica provided after the earlier work was completed. The revised weights correct for a minor problem in the baseline weights. These two changes result in a slightly larger estimate of the ITT effect for Black students in the first and second follow-up years, and a slightly smaller one in

the third follow-up year. Henceforth, we focus on estimates using the more appropriate weights and strata controls.[13]

The middle set of columns in Tables 3a and 3b report estimates for models that use the same sample of students – i.e., those who were in grades 1-4 *and* have baseline test data – but omit the baseline test scores (i.e., equation (1)). Omitting the baseline scores has a trivial effect on the ITT estimate, as one would expect given random assignment. For the year-three composite score for African Americans in Table 3b, for example, the estimated treatment effect is 5.03 points controlling for baseline scores and 5.00 points without controlling for baseline scores. Such stability is expected in a randomized experiment with a reasonably large sample.

The third set of columns report estimates of the same model for the largest possible sample, including students initially in Kindergarten (none of whom were tested at baseline) and those initially in grades 1-4 with missing baseline scores.[14] Seventy-one percent of the additional observations are from the Kindergarten cohort, and 29 percent from the other cohorts. The results for African American students are notably weaker in the larger sample. In year three, for example, the estimated treatment effect in Table 3b falls about in half, to 2.78 points (s.e. = 1.69), with a t-ratio that is about equal to the threshold value for statistical significance at the 0.10 level.[15] By subject, the effect on the

---

[13] Myers and Mayer (2003) acknowledge that the original strata controls we use in Table 3b are preferable to the ones used in their previous estimates.

[14] Note that the cohort in Kindergarten at baseline was in grade 1, 2 or 3 during the follow ups, so testing them in the follow-up waves presumably did not pose greater challenges than testing the other cohorts when they were in those grades.

[15] The conventional OLS standard error in this case is 1.64, not very different than the one we compute, which allows for within family correlated errors. The standard errors we have computed are likely an underestimate because we have not taken into account added uncertainty that arises from the estimation of the follow-up weights or from dependence among paired observations in the PMPD strata.

reading test is even smaller, while the treatment effect for the mathematics test is larger and statistically significant at the 0.05 level. A similar pattern is found in Table 3a.

The notably smaller estimate when students with missing baseline scores are included in the sample is due to both the inclusion of students who were in Kindergarten and those who were in grades 1-4 at baseline. If the Kindergarten cohort is excluded from the sample in the model in Table 3b that omits baseline scores, the treatment effect in the composite score regression for African Americans is 4.37 (s.e. = 1.85; N = 577). So about 30 percent of the reduction in the impact effect from including students with missing baseline scores comes about from adding students in grades 1-4, and about 70 percent from including Kindergarten students.

Interestingly, the standard errors are only trivially different between the estimates that condition on baseline test scores and those that do not but use a larger sample. In the third follow up, for example, the standard error for the treatment effect on African American students' reading scores is 0.09 points smaller using the full sample and omitting baseline scores than it is in the subsample that controls for baseline scores, and for math scores it is the same whether baseline scores are omitted or included. For the composite scores, the standard error is .02 points larger when baseline scores are omitted. The reason for this similarity is that the reduction in residual variance from controlling for baseline scores is roughly offset by the gain in precision from having a larger sample.

One can improve the precision of the estimates in Table 3b by controlling for baseline information other than test scores. Table 4 presents results from models that control for the original 30 randomization strata, cohort (i.e., grade at baseline), gender, log family income, mother's years of schooling, indicators for whether the mother works

15

full- or part-time, an indicator for whether the student has had special education, an indicator for whether the student has been in a gifted or talented class, student's age, and dummies for English spoken at home, mother born in the United States, family welfare receipt, mother resided in residence for more than one year, and whether the mother's religion was Catholic.[16] The first set of columns present results from models that also control for baseline test scores, for the subsample of observations with baseline scores.

Results shown in the second set of columns are based on the larger sample and omit baseline test scores from the model. The treatment effect of vouchers is smaller in these models. In year 3, for example, the impact of offering a voucher on the composite score for African Americans is 2.08 NPR's (t-ratio = 1.24). Again, the impact on math scores is larger than the impact on reading scores.

Finally, it is possible to incorporate the information on baseline scores for those who have it, and still use the larger sample. In particular, we created a dummy variable that equals one if baseline scores are available and zero if they are missing.[17] Define this variable as D. We then interacted D with the baseline math and reading scores, which imposes a value of zero for those with missing scores regardless of the values one would have used to impute those scores. The following "hybrid" model was estimated using the full sample of observations, both those with and without baseline scores:

(3) $Y_{if} = \alpha Z_f + \gamma_M M_{if} \bullet D_{if} + \gamma_R R_{if} \bullet D_{if} + \lambda D_{if} + X_{if}' \beta + v^c_{if} + \delta^j_{if} + \varepsilon_{if,}$

where X is a vector of other baseline covariates, such as gender, mother's education and log family income, and v^c is a set of cohort fixed effects. This model has the effect of

---

[16] So as to not reduce the sample size, missing values of income and mother's education were replaced by their means, and dummy variables were included indicating whether income, mother's education, students' gender, mother's employment status, and each of the other covariates were missing.
[17] It is never the case that math scores are missing and reading scores are available, or vice versa.

controlling for baseline scores for students who have them, and not controlling for baseline scores for student who lack them. Because baseline scores can be missing for two distinct reasons – either students are in the Kindergarten cohort and not tested by design or they are in the other cohorts and failed to comply with testing – controlling for cohort dummies and D in the hybrid model removes differences between these groups.

Results of estimating $\forall$ from equation (3) without **X-**covariates (but with cohort and lottery strata dummies) are reported in the third set of columns in Table 4, and with additional **X-**covariates in the fourth set of columns of the table. Notice that the standard errors from both hybrid models are smaller than those from the other models in the table. The hybrid model with the additional **X-**covariates in the right-most set of columns, it turns out, typically yields the most powerful estimates of the models we have considered. Controlling for baseline test scores when they are available results in a slightly higher estimated effect of vouchers when the **X-**covariates are excluded. In the model without the **X-**covariates, the year-three impact of offering a voucher on the composite test scores of African American students is just below the threshold for statistical significance at the 0.05 level. In the model with additional baseline covariates – which yields a slightly more precise estimate -- the effect is somewhat smaller, 2.13 NPR's, and not statistically significant at the 0.10 level. This point estimate is less than half as large as the estimate in Mayer, Peterson and Myers, et al. (2002) and more precisely estimated.

2.1 To Control, or Not Control

Mayer, Peterson and Myers, et al. (2002, p. 9) give the following justification for why they controlled for baseline test scores: "Baseline characteristics were included to adjust for chance differences between the characteristics of treatment- and control-group

17

members and to increase the precision of the estimated impacts." But because both sets of estimates – those with or those without baseline characteristics – provide an unbiased estimate of the treatment effect, only the latter rationale is potentially germane.[18] Any improvement in the estimates from controlling for covariates is *fully* reflected in the sampling variance of the estimates. With or without covariates, the estimates are unbiased and approximately normally distributed, so the coefficient and standard error completely characterize the sampling distribution of the estimated treatment effect.

The key question then is whether including baseline covariates reduces the residual variance by enough to justify the loss of degrees of freedom and the fact that, in practice, some observations must be dropped from the sample because they have missing baseline data. On this basis, there is little reason to chose between the estimates in Table 3b that control for baseline test scores and the ones that do not. Both are unbiased with about equal power. However, an advantage of estimates using the larger sample is that the results potentially can be generalized to a broader population – namely, students initially in grades K-4, as opposed to those in grades 1-4 without missing baseline data.

Moreover, estimates without baseline covariates are simple and transparent. And unless the specific covariates that are to be controlled are fully described in advance of analyzing the data in a project proposal or planning document, there is always the possibility of specification searching. To avoid the charge of specification searching, we have provided results without covariates (except for those that define the randomization

---

[18] Another way to see that the first rationale is inappropriate is to note that if there is a chance difference in a baseline characteristic between treatments and controls, there could also be an erroneous correlation (due to chance or misspecification) between the baseline characteristic and the outcome variable that would sway the estimated treatment effect if covariates are included. A correlation between treatment status and baseline covariates in an experiment is potentially problematic regardless of whether baseline covariates are controlled in a regression. Mayer, Peterson and Myers, et al. seem to imply that an unbiased estimator can be made less biased by controlling for covariates.

process), as well as others with a varying set of baseline covariates, so the reader can judge the sensitivity of the results for him or herself.

Lastly, notice that equation (4) in the far right of Table 4, which incorporates the baseline covariates for those who have them and still uses observations for those with missing baseline data, provides the most precise estimates and pertains to those initially in grades K-4. An argument can be made that, of the estimates presented so far, the most credence should be placed in these because they are the most precise. As we explain in Section 4 below, however, we think the sample of Black students used in Table 4 is unnecessarily restrictive. We also think it is wise to draw conclusions from the whole panoply of results, rather than one particular specification.

2.2 Cohort Interactions

The results in Tables 3a, 3b and 4 suggest that including students who were enrolled in Kindergarten at baseline in the sample qualitatively weakens the impact of school vouchers on achievement scores of African American students. These students were in grades 1-3 when the follow-up tests were administered (assuming they were not held back), and were among the youngest students in the experiment. One possibility is that vouchers are effective for older African American students, but not younger ones. Mathematica (2000) raised this concern after the second-year follow-up study. Howell and Peterson (2002; p. 222) strongly argue against this interpretation, however, maintaining that "no decipherable pattern suggests that impacts vary by grade level." Our analysis of inter-cohort differences in treatment effects also suggests that the grade at which students are offered vouchers is unrelated to the magnitude of the treatment effect in the third year of the experiment when we test the null hypothesis of any unrestricted

pattern of cohort effects, although we find some tendency for older students to have a larger treatment effect when Kindergarten students are included in the sample and the cohort-treatment-status interaction is constrained to be linear.

The treatment effect on the third-year composite test score for each cohort can be viewed in the Addendum at the bottom of Table 1. The treatment effects are not uniform, and it is particularly large for the 4th grade cohort, but they individually have large standard errors. To conduct a formal test for inter-cohort variability in treatment effects, we interacted five dummy variables indicating membership in one of the initial grade cohorts with the treatment status dummy in the models in Table 4, and conducted an F-test of the null hypothesis that the treatment effects were equal for all cohorts. In the third follow-up year, there was no more than a chance occurrence that the treatment effect differed across the cohorts for African American students. For example, in the model for the composite score in the far right of Table 4, the p-value for an F-test of uniform treatment effects is 0.33. This F-test puts no structure on the pattern of interactions and does not depend on first screening the pattern of coefficients to test a particular hypothesis, so in this sense it is "hands above the table."

Observing the pattern of treatment effects by cohort, however, it appears that there may be a tendency for the older cohorts to have larger coefficients than younger cohorts. If we perform a simple test for differential cohort effects that recognizes the *order* of the cohorts by including an interaction between a student's linear grade level at baseline (0, 1, 2, 3, 4) and the voucher offer dummy, as well as the voucher offer main effect, the p-value for a test of the interaction term in the most powerful model in Table 4 is 0.06 for the composite score. Thus, there is some evidence that the treatment effect

rises with the entry grade of the cohort of African American students in the full sample, but it is not overwhelming and it requires the imposition of a linear cohort-treatment interaction. Moreover, it is unclear why the treatment effect would vary with age for African American students but not Hispanic students.

2.3 Attrition

One commonly voiced criticism of the voucher experiments is that attrition was high, especially for the control group (e.g., Neal, 2002). The response rate in the third year of the New York experiment was 68.8 percent for treatment group members and 64.6 percent for control group members in cohorts 1-4. The rates were slightly higher for the Kindergarten cohort: 71.7 percent for treatments and 69.2 percent for controls. Although, in principle, the follow-up sample weights adjust for attrition based on observable characteristics, it is possible that those who exited the sample differ along unobserved dimensions, and the average achievement in year three for non-responding treatments and controls could be different than it was for respondents. Nonrandom attrition that is correlated with treatment status would bias the weighted estimates.

In results not reported here, we examined the impact of attrition in a couple of different ways. First, we computed two-step normal selection correction estimates to adjust for those with missing follow-up test scores. Second, we used the technique employed in Krueger (1999) of carrying forward the last available test score for those with missing data to impute values. We also used these techniques to adjust for students with scores of zero. These adjustments typically led to only slightly smaller estimates of the effect of vouchers. Thus we conclude that nonrandom attrition was probably not a major problem in the experiment.

3. The Relevance of Race and Residence

Our interpretation of the results so far is that the impact of vouchers on achievement for Black students in New York City is smaller and less robust than has been previously acknowledged. Nevertheless, we examine reasons why there may have been gains using the sample that displayed the largest impact, Black students in the grade 1-4 cohorts with baseline test data.

The inference that African American students benefited academically from scholarships to attend private school would be more credible if there were a compelling explanation for why African American students benefited while other, equally poor Hispanic and White students did not benefit. Howell and Peterson (2002) offer the following *ex post* explanation: "African Americans, more than other groups, live in the poorest, least attractive, and most dangerous communities within metropolitan regions. … Precisely because African Americans suffer most under a system of public education based on residency, they stand to benefit the most from the new education opportunities that vouchers afford."

In short, they hypothesize that constrained residential choices for African American families – but not Dominican or Puerto Rican families – account for the poor performance of public schools attended by African American students. As a result, African American students are disproportionately confined to public schools that under perform vis-à-vis private schools. This is a plausible and potentially quite important hypothesis, with many striking implications.[19] Indeed, if discrimination in the housing market were responsible for the alleged poorer performance of Black students in public

school, then enforcing anti-discrimination laws could possibly be more effective at raising achievement than school reform.

Howell and Peterson's hypothesis is testable. In particular, if differential residential location and therefore differential school attendance explains the results, then the minority of non-Black students who attend the *same* public schools as Black students should benefit from being offered private school scholarships as much as Black students. We find no support for this hypothesis.

To test the hypothesis, we identified all the non-Black students in the sample who, at the time of baseline, attended a public school that was also attended by a Black student in the sample that year. In the third follow-up, a total of 470 non-Black students and 333 Black students in cohorts 1-4 came from overlapping baseline schools. We then adjusted the sample weights for this subsample of non-Black students so that weighted counts would correspond to the (weighted) distribution of Black students in the overlapping schools. Consequently, we can produce results for a sample of non-Black students in the experiment who attended the *same* distribution of schools as a subsample of Black students in the experiment. If differential residence and school attendance patterns by race are responsible for the differential effect of school vouchers, we would expect the sample of non-Black students in this subsample to exhibit the same treatment effect of vouchers as the Black students.

This was not the case. If anything, the results point in the opposite direction. For third year math scores, for example, the treatment effect of vouchers for the sample of Blacks from overlapping schools is 5.11 (s.e. = 2.39) and for non-Black students is –3.00

---

[19] One fact that is inconsistent with this hypothesis, however, is that, on average, the Hispanic students have lower baseline test scores and lower family incomes than the Black, non-Hispanic students in the sample.

(s.e. = 3.32).[20] Even when attention is limited to the set of overlapping schools that Black and non-Black students attended, the non-Black students do not appear to gain from being offered a voucher; indeed, they do worse, and the differential treatment effect is statistically significant at the 0.05 level. In addition, we find an insignificant effect if we interact the percent of students in the baseline school who are Black (derived from Common Core Data reported by schools) with the treatment status dummy for non-Black or for Black students. These findings suggest that differential characteristics of the initial public school that students with different racial backgrounds attended do *not* account for any gain in test scores that Black students may have reaped from attending private school.[21]

This conclusion is a contrast to Krueger and Whitmore's (2002) findings on racial differences of the effect of class size. They find that the Black students benefited more from attending smaller classes than White students, but when they estimate effects for White students who attended the same mix of schools as Black students, the White students benefited equally. They concluded, "These findings suggest that small classes matter for Blacks because of something having to do with the schools they attend, rather than something inherent to individual Black students *per se*." In the case of New York City school vouchers, if one accepts that there is an impact for Black students, something about race itself would seem to be part of the reason.

---

[20] These results control for baseline test scores and 30 strata dummies. Similar results arise if reading data are analyzed, or if data from all five cohorts are used and baseline scores are dropped from the model. We focus here on the older cohorts to give the hypothesis a stronger chance of being supported by the data.

[21] From examining administrative data on school characteristics and parental reports on the school environments, Mayer, Myers and Tuttle (2002) similarly conclude, "Differences in school characteristics do not appear to explain the differences in test scores between the African American and Latino controls."

4. Hispanic and Non-Hispanic Black Mothers and Fathers

If race matters, it is important to know how race is defined. Race is a social construct that varies from survey to survey. In Howell and Peterson (2002), African American students are defined as children whose mother or female guardian is reported as Black/African American (non-Hispanic) in the parental survey. Specifically, the Mathematica baseline survey asked respondents to "MARK ONLY ONE" of the following racial/ethnic categories for each parent or guardian: "Black/African American (non-Hispanic); White (non-Hispanic); Puerto Rican; Dominican; Other Hispanic"; American Indian; Chinese; etc. At the end of the list, respondents were given an opportunity to mark "Other" and write in a response. Mathematica then coded a child's race as Black if the mother or female guardian marked "Black/African American (non-Hispanic)". Consequently, a child's race and Hispanic ethnic origin are by definition mutually exclusively in these data. This deviates from the Office of Management and Budget's guidelines (Statistical Policy Directive No. 15), which recommend separate questions for race and ethnicity self-identification questions.[22]

We find from the 1990 Census that in the New York metropolitan area 15 percent of low-income ($\leq$ $20,000) children age 5-14 whose race is reported as "Black, African American or Negro" are also reported as Hispanic.[23] And 28 percent of those who indicate their ethnicity as Dominican indicate their race as Black/African American. A total of 541 students whose race/ethnicity was classified as Dominican participated in the

---

[22] In other surveys, such as the National Job Corps Study, Mathematica has used the OMB format of asking separate race and ethnicity questions.
[23] This is based on a sample of children living in the New York-Northeastern New Jersey Consolidated Metropolitan Area.

voucher experiment, almost half as many as the number of non-Hispanic Blacks who participated. Many of those who were categorized as Dominican probably would have identified themselves as Black if the Census question on race had been used in the study. Because the treatment effect of vouchers was negative for students whose race was classified as Dominican Republic, including Black Dominicans in the sample of Black students would likely lead to smaller estimated treatment effects.

Another problem is that the practice of assigning a child the race/ethnicity of his or her mother irrespective of his or her father's race is asymmetric and restrictive.[24] Some countries, such as China, legally assign a child's race based on the father's race, while others use the mother's race, and still others allow families to choose. In part, race may matter because society treats individuals with different skin tones differently (see, e.g., Darity, Hamilton and Dietrich, 2001, Keith and Herring, 1991 and Scarr, et al., 1977). If this is the case, then one could argue that treating mothers and fathers symmetrically is more sensible than assigning race according to the mother's race, regardless of father's race. Moreover, race was reported as Black for 85 percent of the children with a Black father and a Hispanic mother present in the New York metropolitan region, according to our tabulation of the 1990 Census. Because most of the children with a Black father who were not classified as Black in the Mathematica sample had a Hispanic (possibly Black) mother, it is quite likely that these students would have been classified as Black had they or their parents been given the opportunity to report the race of the child on the questionnaire.

---

[24] As an example of the types of problems created by this procedure, note that 8 students in the data set had a Black (non-Hispanic) father and were missing information on their mother's race/ethnicity; these students were classified as non-Black. In three of these cases, there is no indication the mother lived at home.

Yet another concern is that the data were coded in such a way that parents who checked "Other" and then wrote in a response in the blank were typically classified as non-Black *and* non-Hispanic, regardless of what they wrote in. This is problematic because in many cases the child's mother wrote "Black/Hispanic" or "Black/Puerto Rican," and a few even wrote in just "African American" or "Black." Clearly, children in these households would be classified as Black/African American in most surveys.

In Table 5 we present results for the group of students for whom *either* parent's race/ethnicity is identified as Black/African American (non-Hispanic). We also include in the sample students whose parents responded "Other" for race/ethnicity and then indicated that their race was Black in their write-in response.[25] These changes increase the sample by about 10 percent. This is a broader definition of Black students' race than the one employed previously, although it still treats race and Hispanic origin as mutually exclusive unless such a response was written in, contrary to the OMB guidelines.

Notably, the results for the broader sample of Black students are weaker than those in the original sample. In the full sample, including students with missing baseline scores, the effect of offering a voucher on the third-year composite score is quite small (1.44 NPR's) and statistically insignificant at conventional levels (t-ratio=1.01). To put this in context, note that the standard deviation of percentile ranks in the national population is 28.9 (because percentiles are uniform), so the estimated effect size is only 0.05 standard deviations. If we use only students in grades 1-4 and the larger group of Black students, the fourth model in Table 5 yields a coefficient of 2.67 points and a

---

[25] Of the 78 students who are added in Table 5 model 3, 43 were added because their father was reported as Black and 35 because a written response for the mother's race/ethnicity indicated that her race was Black, usually by writing Black/Hispanic or Black combined with a specific Latin country. Mathematica used the written-in responses in an inconsistent way, sometimes using them to assign race and sometimes not.

t-ratio of 1.78, which suggests that the previous results for grades 1-4 are fragile if a more conventional definition of race is used.

Moreover, if we interact treatment status with three dummies indicating combinations of parents' race (mother Black, father Black, both parents Black), we do not reject that the treatment effect is the same for all three groups. The coefficients, with standard errors in parentheses, for the voucher-race interactions in the last model of Table 5 in year three, for example, are: 1.37 (1.73) for both parents Black; 0.34 (5.91) for father Black, mother non-Black or Hispanic; and 1.53 (2.36) for mother Black, father non-Black or Hispanic. The p-value for an F-test of the null hypothesis that the three effects are equal is 0.98, about as far from rejecting the null of constant treatment effects as possible.

These results cast further doubt on the extent to which one can generalize previous findings from the voucher experiment to the population of low-income Black students. We would stress that the qualitatively different results in Table 5 stem primarily from two plausible changes in the sample from previous studies: students with missing baseline scores are included and parental race is treated symmetrically for assigning children's race. *We regard the results in Table 5 as the most relevant estimates of the impact of offering private school vouchers on achievement for the broadest sample of Black elementary school children.* This sample comes closer to meeting the OMB guidelines, although some Black/Hispanic children are undoubtedly excluded. The finding that the results are so sensitive to these defensible changes in the sample leads us to conclude that the provision of vouchers in New York City probably had no more than a trivial effect on the average test performance of participating Black students. Furthermore, if parents were not asked to give a mutually exclusive response to their race

28

and ethnicity in a single question, contrary to the OMB guidelines – or if data on the students' race were directly collected – the effect of vouchers for Black students of *any* ethnicity would likely be even smaller.

5. Effect of Time in Private School – Instrumental Variables Estimates

Although the ITT estimates of the effect of offering a student a voucher are of much interest, another question concerns the effect of *attending* private school on student achievement. Howell and Peterson (2002) and Mayer, Peterson and Myers, et al. (2002) report Instrumental Variables (IV) estimates to assess the impact of attending private school on student achievement.[26] Specifically, they create a dummy variable that equals one if the student attended private school for three consecutive years, and zero otherwise. They then use the voucher offer dummy as an instrument for private school attendance. Because the voucher was randomly offered, if the variables are correctly measured this approach yields a consistent estimate of the population parameter. (If there are heterogeneous treatment effects across students, additional assumptions are necessary to interpret the parameter as the causal effect on compliers; see Angrist, Imbens and Rubin, 1996.) As Rouse (1997) and Mayer, Peterson and Myers, et al. point out, if switching to private school for one or two years raises (lowers) achievement, this approach will overstate (understate) the impact of attending private school for three years versus not at all. Mayer, Peterson and Myers, et al. also present estimates where the endogenous regressor is a dummy indicating whether the student ever attended private school.

Intuitively, all the IV estimates involve scaling the ITT estimate emphasized

---

[26] The technique of Instrumental Variables was first used by the economist P.G. Wright in 1928; see Stock and Trebbi (2003).

29

previously by the difference in some quantity between those offered and not offered a voucher. To see this, note that if there are no covariates the IV estimate of the effect of attending private school for three years, θ, has probability limit:

(4) $\text{plim } \theta = \{E[Y \mid Z = 1] - E[Y \mid Z = 0]\} / \{E[P3 \mid Z = 1] - E[P3 \mid Z = 0]\}$,

where Y is test scores, Z is a dummy that equals one if a voucher was offered and zero if not, and P3 is a dummy variable that equals one if the student was enrolled in private school for all three years of the experiment and zero otherwise. The numerator is the ITT estimate. In a model in which attending private school for at least one year is the endogenous regressor, the numerator is the same but the denominator is the treatment-control difference in the probability of attending private school at least one year.

An alternative approach is to treat the number of years in private school as the relevant variable of interest. Assuming years in private school have a linear effect on achievement, the equations of interest are:

(5) $Q_{if} = \alpha_0 + \alpha_1 Z_f + X_{if}'\alpha_2 + \gamma^j_{if} + \mu_{if}$
(6) $Y_{if} = \beta_0 + \beta_1 Q_{if} + X_{if}'\beta_2 + \delta^j_{if} + \varepsilon_{if}$,

where Q_{if} is the number of years the student has spent in a private school up to that point of the experiment, Z_f is a dummy indicating whether the student was offered a voucher, X_{if} is a vector of baseline covariates, Y_{if} is the test score in the relevant follow-up year, γ^j_{if} and δ^j_{if} are randomization strata fixed effects, and μ_{if} and ε_{if} are equation errors. Interest is in the parameter β_1, which measures the marginal impact on performance of attending private school an additional year. An advantage of the linear model is that the coefficient β_1 can be compared in different years. Furthermore, for comparison to θ, an estimate of the effect of attending private school for all three years is $3\beta_1$.

Two-Stage Least Squares estimates of β_1 using various samples are presented in Table 6. Similar to our earlier findings, the results indicate that spending more time in private school has a positive and statistically significant effect on achievement for children of Black (non-Hispanic) mothers when the analysis is confined to those with baseline test scores, but an insignificant effect in the larger samples. In the third follow-up, the implied effect of attending a private school for three years is 6.4 points ($t = 2.58$) in the sample with baseline test scores, and 3.3 points ($t = 1.41$) in the sample that is not restricted to those with baseline scores. These effects are, respectively, 31 and 65 percent smaller than the estimated impact of spending three years in a private school emphasized by Howell and Peterson (2002; Table 6-3).[27]

The third set of columns in Table 6 use the race/ethnicity of the father as well as the mother to delineate the samples. That is, as in Table 5, the sample of Black (non-Hispanic) students includes children for whom *either* parent is reported as Black (non-Hispanic). The sample of Latino students likewise includes students for whom either parent is reported as Puerto Rican, Dominican, or other Hispanic. (These samples are not mutually exclusive, as is normally the case with race and ethnicity.) For students with a Black parent, the estimated effect of attending a private school for three years is 2.2 points ($t=1.01$), 76 percent smaller than Howell and Peterson's estimate for students with a Black (non-Hispanic) mother.

One final consideration involves measurement error in private school attendance. For the controls, information on whether students attended private or public school was primarily inferred from the parents' report of the name of the school, and public/private

---

[27] These estimates use the revised weights and original strata controls. If we use the same model, weights and strata controls as Howell and Peterson (2002), we can replicate their results.

school attendance is missing in at least one year for 42 percent of control group members.[28] Administrative records were used to determine private school enrollment for those offered a voucher, so this variable is never missing for the treatment group. To avoid dropping observations, Howell and Peterson and Mathematica imputed a value for public or private school enrollment if this information was missing. We followed their procedures. However, there is clearly measurement error in private school attendance for many students. The imputed values are unlikely to be correct in every case. Even for the treatment group, several parents indicated that their child attended a parochial school, yet they were classified as attending public school that year, or vice versa.

Random measurement error would not affect the consistency of IV estimates if the endogenous variable were continuous. In the case of a binary endogenous variable (e.g., if the variable being instrumented for is a dummy indicating private school attendance for three years), however, Kane, Rouse and Staiger (1999) show that IV estimates are asymptotically biased, in an upwards direction under conditions that are likely to hold. This results because measurement error is necessarily correlated with the true value of the endogenous variable in the case of a dummy variable, and therefore correlated with the instrument. Because misreporting and imputation create measurement error in their private school enrollment variable, Howell and Peterson's IV estimates are likely to overstate any effect of switching to private school on achievement. This is another reason to favor the ITT estimates.

---

[28] For control group members with available third year follow-up test data and baseline test scores, 18.5 percent are missing information on whether they attended a private or public school in at least one year.

6. Conclusion

Our reanalysis of the New York City school voucher experiment suggests that the positive effect of vouchers on the achievement of African American students emphasized by previous researchers is less robust than commonly acknowledged. Most importantly, if the cohort of students who were enrolled in Kindergarten when the experiment began is included in the sample, the effect of vouchers is greatly attenuated. As the results in Table 5 indicate, treating mother and father's race symmetrically further attenuates the effect of school vouchers for African American children. The evidence is stronger that the availability of private school vouchers raised achievement on math than on reading exams after three years, but both effects are relatively small if the sample includes students with missing baseline test scores and students who have at least one Black parent.

Below is a list of several issues about experimental program evaluation that we believe our analysis raises:

• Researchers are often unsure as to whether they should or should not control for baseline characteristics when a treatment is randomly assigned. We would advise that key results be presented *both* ways, with and without baseline characteristics (and with and without varying samples).[29] In expectation, the treatment effect should not change; if it does, more work is needed to understand why.

• Controlling for baseline characteristics can be justified if their inclusion increases the precision of the key estimates. As a practical matter, however, controlling for baseline

---

[29] This is often done in the academic economics literature. For example, Angrist, Bettinger, and Bloom, et al. (2002) present results with and without baseline controls in their analysis of a natural experiment involving vouchers in Colombia. Researchers are on stronger grounds for presenting results that do not condition on baseline covariates when they have an actual randomized experiment.

characteristics tends to reduce the sample size, which could well offset the decline in residual variance and create a non-representative sample.

- Simplicity and transparency are valuable in their own right and can help prevent mistakes. These benefits may be well worth the loss of some precision. A complicated design increases the likelihood of error down the road – for example, in the derivation of weights or in the delineation of strata within which the treatment is randomly assigned. An under appreciated virtue of presenting results *without* baseline covariates is that the results are transparent and simple, and therefore less prone to human error.

- Having a broader sample expands the population to which the results can be generalized. One wonders why the foundations that funded the voucher experiment paid the additional costs to grant scholarships and administer follow-up tests to Kindergarteners and other students with missing baseline tests if they were not to be used in the analysis.[30] Because these students make up more than 40 percent of the sample used in previous analyses, there was also a loss in efficiency from excluding them from the sample. (It is unclear whether follow-up tests were administered to students with missing baseline test scores in the Dayton and Washington experiments.) In addition, because decisions by parents about whether to send their children to private school in a publicly funded voucher program are particularly likely to be made around the time of school entry, one could argue that the experience of the Kindergarten cohort is the most relevant of the five cohorts for extrapolating the results to a broader, permanent voucher program.

---

[30] Certainly the participating students would wonder why they were inconvenienced on three weekends to take exams that were not used in the analysis published by the research team.

• Researchers often make data available to other scholars, but limit the data to the sample and variables used in their previous analysis. Mathematica deserves much credit for quickly making data available (for a small fee) to outside researchers who agreed to adhere to certain confidentiality requirements, and for providing all their data, not just the subset used to generate results presented in previously published reports.

• The definition of race can be more than an incidental detail. Researchers should think carefully about the assignment of a child's race, especially if the child or parent is not asked to directly report the child's race. Because race is inherently a subjective concept, there may be benefits from exploring the sensitivity of the findings to alternative definitions of race. Government statistics typically define race and ethnicity separately. Using a consistent and clear definition of race is necessary if results are to be compared across studies and data sets. For example, NAEP data from NCES define race to include students of any ethnicity, so the Black-White gap from NAEP is not a comparable benchmark for statistics pertaining to students with a Black (non-Hispanic) mother from the voucher experiments.

Computing separate results by race was not an original goal of the New York City voucher experiment. Indeed, in the first-year follow-up report, results were not reported separately by race. Only after the results indicated no impact for the full sample were the data disaggregated by race. Had race been a priority of the study from the beginning, we suspect the survey designers would have asked directly about students' race and their Hispanic ethnicity.

• Although we do not think that a fully specified and empirically verified theoretical model is necessary to interpret experimental results (let alone achievable), a plausible and

testable theoretical explanation can help avoid mistakes in interpretation and policy. For example, the lack of a convincing theory for why African American students would benefit from vouchers while Hispanic students from the same schools would not is a cause for concern in interpreting the New York City experiment. In its simplest form, Howell and Peterson's model of constrained residential choice would predict that poor Hispanic students attending the same initial public schools as African American students would experience a rise in test scores from vouchers, at variance with the data. Combined with our finding that the effect of vouchers for African American students is more fragile than previous analyses of the data have suggested, we would counsel caution in concluding that vouchers raised achievement for African American students in New York City. The safest conclusion is probably that the provision of vouchers did not lower the scores of African American students.

References

Angrist, Joshua D., Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer, "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment," *American Economic Review*, Vol. 92, No. 5, December 2002, pp. 1535-58.

Angrist, Joshua D., Guido Imbens, and Donald B. Rubin, "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, vol. 91, 1996, pp. 444-455.

Cochran, William G. and Gertrude M. Cox, *Experimental Designs*, New York: John Wiley and Sons, 2nd edition, 1957.

Darity, William, Darrick Hamilton and Jason Dietrich, "Passing on Blackness: Latinos, Race and Earnings in the USA," Mimeo., University of North Carolina, September 2001.

Fisher, Ronald A., *The Design of Experiments*, New York, Hafner, 6th ed., 1951.

Hill, Jennifer L., Donald B. Rubin, and Neal Thomas, "The Design of the New York School Choice Scholarship Program Evaluation," in *Donald Campbell's Legacy*, edited by Leonard Bickman, Sage Publications, 2000.

Howell, William G. and Paul E. Peterson, *The Education Gap: Vouchers and Urban Schools*, Washington, DC: Brookings Institute Press, 2002 (Advance Reading Copy).

Kane, Thomas, Cecilia E. Rouse, and Douglas Staiger, "Estimating Returns to Schooling When Schooling is Misreported," Industrial Relation Section Working Paper No. 419, Princeton University, June 1999. (Available from www.irs.princeton.edu.)

Keith, Verna M., and Cedric Herring, "Skin Tone and Stratification in the Black Community," *American Journal of Sociology*, Vol. 97, No. 3. (Nov., 1991), pp. 760-778.

Krueger, Alan, "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics*, vol. 114, no. 2, May, 1999, pp. 497-532.

Krueger, Alan and Diane Whitmore, "Would Smaller Classes Help Close the Black-White Achievement Gap?" In *Bridging the Achievement Gap*, edited by John Chubb and Tom Loveless, Washington, DC: Brookings Institute Press, 2002.

Mathematica Policy Research, Inc., Press Release, "Voucher Claims of Success Are Premature in New York City," Washington, D.C., September 2000.

Mayer, Daniel P., David E. Myers, and Christina Clark Tuttle, "Appendix E: African American and Latino Test-Score Differences," in "School Choice in New York City After Three Years: An Evaluation of the School Choice Scholarships Program," edited by Daniel P. Mayer, et al.,Washington, DC: Mathematica Policy Research, Inc., 2002.

Mayer, Daniel P., Paul E. Peterson, David E. Myers, Christina Clark Tuttle, and William G. Howell, "School Choice in New York City After Three Years: An Evaluation of the School Choice Scholarships Program," Washington, DC: Mathematica Policy Research, Inc., 2002.

Myers, David and Daniel P. Mayer, "Comments on 'Another Look at the New York City Voucher Experiment'," mimeo., April 1, 2003.

Moulton, Brent R., "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units," *Review of Economics and Statistics*, vol. 72, no. 2, May 1990, pp. 334-338.

Neal, Derek, "How Vouchers Could Change the Market for Education," forthcoming *Journal of Economic Perspectives*, 2002.

Peterson, Paul E., David Myers and William G. Howell, "An Evaluation of the New York City Scholarships Program: The First Year," Washington, DC: Mathematica Policy Research, Inc., 1998.

Rouse, Cecilia E., "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program," NBER Working Paper Number 5964 (March, 1997).

Scarr, Sandra, Andrew J. Pakstis, Solomon Katz and William Barker, "Absence of a relationship between degree of white ancestry and intellectual skills within a black population," *Human Genetics* 39, 1977, pp. 69-86.

Stock, James and Francesco Trebbi, "Who Invented IV Regression?" February 2003, forthcoming, *Journal of Economic Perspectives*.

Table 1: Treatment-Control Differences at Baseline by Assignment Status, Selected Variables

Group	Mean treatment	Mean control	Raw T-C Difference	t-ratio	Conditional on Lottery Strata: T-C Difference	t-ratio	Sample size
MOTHER'S YEARS OF EDUCATION							
Overall	12.85	12.78	0.08	0.72	0.08	0.76	2,476
Black- all	13.03	12.96	0.07	0.50	0.11	0.74	1,081
Black-ch0	12.83	12.91	-0.09	-0.31	-0.10	-0.35	233
Black-ch1	13.10	13.22	-0.12	-0.45	-0.03	-0.12	215
Black-ch2	13.22	13.07	0.15	0.52	0.19	0.62	247
Black-ch3	12.83	12.87	-0.03	-0.12	0.01	0.02	216
Black-ch4	13.20	12.58	0.62	1.92	0.52	1.45	169
Latino-all	12.48	12.40	0.08	0.53	0.09	0.58	1,186
MOTHER GRADUATED COLLEGE							
Overall	0.11	0.10	0.02	0.95	0.01	0.90	2,476
Black- all	0.12	0.09	0.03	1.28	0.04	1.71	1,081
Black-ch0	0.13	0.05	0.07	1.82	0.06	1.53	233
Black-ch1	0.07	0.12	-0.05	-1.12	-0.01	-0.28	215
Black-ch2	0.17	0.13	0.03	0.57	0.04	0.67	247
Black-ch3	0.07	0.08	-0.01	-0.29	0.00	-0.11	216
Black-ch4	**0.20**	**0.04**	**0.16**	**2.93**	**0.13**	**2.25**	**169**
Latino-all	0.06	0.07	0.00	-0.18	0.00	-0.11	1,186
ANNUAL INCOME							
Overall	$10,247	$9,984	$263	0.69	$248	0.66	2,433
Black- all	10,582	10,310	272	0.45	153	0.25	1,068
Black-ch0	11,847	10,326	1,521	1.21	1,193	0.87	228
Black-ch1	8,688	10,396	-1,709	-1.51	-1,395	-1.10	213
Black-ch2	10,688	10,075	613	0.61	450	0.41	246
Black-ch3	9,747	10,778	-1,031	-0.99	-1,053	-0.95	211
Black-ch4	**12,466**	**9,915**	**2,551**	**2.19**	2,589	1.84	169
Latino-all	9,871	9,459	412	0.78	275	0.54	1,161
BASELINE TEST SCORES							
COMPOSITE MATH AND READING SCORE (AVERAGE NPR)							
Overall	20.02	20.83	-0.82	-0.79	-0.65	-0.61	1,851
Black- all	20.05	20.41	-0.36	-0.25	-0.34	-0.23	806
Black-ch0	NA	NA	NA	NA	NA	NA	NA
Black-ch1	17.31	21.70	-4.39	-1.53	-4.46	-1.38	203
Black-ch2	23.35	22.45	0.90	0.34	-0.30	-0.10	230
Black-ch3	16.17	18.03	-1.86	-0.75	-1.91	-0.74	214
Black-ch4	**24.81**	**18.40**	**6.41**	**2.11**	5.73	1.78	159
Latino-all	18.09	19.97	-1.87	-1.31	-1.78	-1.15	876
READING (NPR)							
Overall	22.90	24.55	-1.65	-1.35	-1.45	-1.16	1,851
Black- all	24.33	25.37	-1.04	-0.58	-0.90	-0.48	806
Black-ch0	NA	NA	NA	NA	NA	NA	NA
Black-ch1	24.46	31.09	-6.63	-1.64	-7.45	-1.57	203
Black-ch2	26.40	25.94	0.46	0.13	0.72	0.19	230
Black-ch3	19.75	21.69	-1.94	-0.71	-2.80	-1.00	214
Black-ch4	**27.75**	**20.75**	**7.00**	**2.12**	**7.59**	**2.09**	**159**
Latino-all	19.82	23.26	-3.44	-1.98	-3.18	-1.70	876

MATH (NPR)							
Overall	17.13	17.11	0.02	0.01	0.16	0.14	1,851
Black- all	15.76	15.45	0.32	0.22	0.22	0.15	806
Black-ch0	NA	NA	NA	NA	NA	NA	NA
Black-ch1	10.16	12.31	-2.15	-0.95	-1.47	-0.62	203
Black-ch2	20.29	18.96	1.33	0.51	-1.31	-0.47	230
Black-ch3	12.59	14.37	-1.78	-0.67	-1.02	-0.37	214
Black-ch4	21.88	16.06	5.81	1.60	3.87	1.10	159
Latino-all	16.37	16.67	-0.30	-0.20	-0.39	-0.25	876
MOTHER EMPLOYED FULL TIME							
Overall	0.22	0.22	0.00	0.14	0.00	0.10	2,479
Black- all	0.28	0.26	0.02	0.66	0.01	0.33	1,083
Black-ch0	0.36	0.27	0.09	1.26	0.04	0.53	233
Black-ch1	0.19	0.29	-0.11	-1.61	-0.08	-1.12	216
Black-ch2	0.27	0.20	0.06	1.09	0.07	1.19	248
Black-ch3	0.25	0.31	-0.06	-0.94	-0.10	-1.28	216
Black-ch4	**0.36**	**0.19**	**0.17**	**2.40**	**0.17**	**2.02**	**169**
Latino-all	0.19	0.19	0.00	-0.02	0.00	-0.01	1,187
STUDENT SEX (MALE)							
Overall	0.50	0.48	0.01	0.67	0.01	0.64	2,617
Black- all	0.47	0.53	-0.06	-1.73	-0.06	-1.72	1,134
Black-ch0	**0.42**	**0.60**	**-0.18**	**-2.56**	**-0.22**	**-3.27**	**241**
Black-ch1	0.50	0.59	-0.09	-1.17	-0.14	-1.60	225
Black-ch2	0.51	0.54	-0.03	-0.41	-0.08	-1.22	256
Black-ch3	0.46	0.43	0.02	0.35	0.00	-0.06	232
Black-ch4	0.48	0.46	0.02	0.25	0.05	0.51	180
Latino-all	0.49	0.44	0.05	1.50	0.04	1.37	1,253
MOTHER'S PLACE OF BIRTH USA							
Overall	0.53	0.55	-0.03	-0.98	-0.03	-0.98	2,593
Black- all	0.75	0.82	-0.06	-1.95	**-0.08**	**-2.47**	**1,132**
Black-ch0	0.74	0.77	-0.03	-0.47	0.00	-0.01	240
Black-ch1	0.76	0.86	-0.11	-1.66	**-0.15**	**-2.31**	**225**
Black-ch2	0.77	0.78	-0.01	-0.17	-0.05	-0.86	261
Black-ch3	0.82	0.82	0.00	-0.06	-0.05	-0.86	227
Black-ch4	**0.64**	**0.87**	**-0.23**	**-3.49**	**-0.22**	**-3.04**	**178**
Latino-all	0.32	0.36	-0.04	-1.17	-0.05	-1.29	1,245
FOOD STAMP RECIPIENT							
Overall	0.66	0.68	-0.03	-1.09	-0.02	-0.98	2,474
Black- all	0.63	0.69	-0.06	-1.59	-0.04	-1.10	1,074
Black-ch0	0.61	0.70	-0.08	-1.22	-0.08	-1.08	227
Black-ch1	0.71	0.68	0.03	0.38	0.07	0.80	216
Black-ch2	0.59	0.69	-0.10	-1.52	-0.13	-1.82	248
Black-ch3	0.69	0.67	0.01	0.21	0.00	0.03	217
Black-ch4	**0.54**	**0.73**	**-0.18**	**-2.31**	**-0.20**	**-2.28**	**165**
Latino-all	0.67	0.70	-0.02	-0.60	-0.02	-0.43	1,190
AFDC RECIPIENT							
Overall	0.57	0.60	-0.03	-1.13	-0.03	-0.95	2,375
Black- all	0.58	0.65	-0.06	-1.64	-0.05	-1.32	1,030
Black-ch0	0.57	0.63	-0.06	-0.83	-0.07	-0.89	218
Black-ch1	0.64	0.63	0.01	0.16	0.03	0.33	204
Black-ch2	**0.51**	**0.66**	**-0.15**	**-2.22**	**-0.18**	**-2.42**	**241**
Black-ch3	0.64	0.60	0.05	0.64	0.03	0.34	206
Black-ch4	**0.53**	**0.73**	**-0.20**	**-2.59**	**-0.24**	**-2.77**	**160**
Latino-all	0.54	0.59	-0.05	-1.21	-0.04	-1.08	1,143

MEDICAID RECIPIENT

Overall	**0.62**	**0.68**	**-0.06**	**-2.37**	**-0.06**	**-2.15**	**2,299**
Black- all	**0.59**	**0.68**	**-0.09**	**-2.25**	-0.07	-1.84	**1,003**
Black-ch0	0.61	0.69	-0.08	-1.10	-0.07	-0.91	217
Black-ch1	0.57	0.60	-0.03	-0.41	-0.01	-0.15	195
Black-ch2	0.58	0.71	-0.13	-1.93	**-0.15**	**-1.97**	**239**
Black-ch3	0.67	0.69	-0.02	-0.22	-0.02	-0.31	203
Black-ch4	**0.52**	**0.73**	**-0.22**	**-2.70**	**-0.22**	**-2.39**	**148**
Latino-all	0.63	0.70	-0.07	-1.83	-0.06	-1.55	1,098

EVER ATTENDED GIFTED STUDENT CLASSES

Overall	0.12	0.12	0.00	-0.13	0.00	-0.07	2,547
Black- all	0.15	0.13	0.02	0.65	0.01	0.51	1,115
Black-ch0	0.08	0.09	-0.02	-0.40	0.00	-0.10	239
Black-ch1	0.12	0.11	0.01	0.21	-0.01	-0.19	222
Black-ch2	0.21	0.17	0.04	0.63	0.00	0.05	254
Black-ch3	0.14	0.11	0.03	0.59	0.06	1.37	227
Black-ch4	0.21	0.19	0.01	0.21	0.01	0.17	173
Latino-all	0.09	0.09	-0.01	-0.32	-0.01	-0.28	1,208

EVER RECEIVED SPECIAL EDUCATION SERVICES

Overall	0.11	0.10	0.01	0.73	0.01	0.74	2,574
Black- all	0.11	0.10	0.01	0.66	0.02	0.78	1,111
Black-ch0	0.08	0.09	-0.02	-0.38	-0.02	-0.33	238
Black-ch1	0.07	0.11	-0.03	-0.60	-0.03	-0.53	222
Black-ch2	0.13	0.10	0.03	0.70	-0.01	-0.14	252
Black-ch3	**0.16**	**0.06**	**0.09**	**2.13**	0.08	1.92	**226**
Black-ch4	0.14	0.15	-0.01	-0.25	0.03	0.44	173
Latino-all	0.11	0.12	-0.01	-0.49	-0.01	-0.51	1,239

ADDENDUM: THIRD YEAR FOLLOW-UP TEST RESULTS
COMPOSITE MATH AND READING SCORE (AVERAGE NPR)

Overall	26.94	27.09	-0.15	-0.12	-0.25	-0.21	1,801
Black- all	25.64	23.78	1.86	1.11	2.78	1.64	733
Black-ch0	22.19	27.04	-4.85	-1.26	-5.80	-1.50	156
Black-ch1	26.69	23.73	2.96	0.73	1.38	0.30	139
Black-ch2	23.02	21.47	1.55	0.52	2.59	0.77	177
Black-ch3	28.26	26.40	1.86	0.53	2.11	0.55	139
Black-ch4	**28.54**	**20.31**	**8.23**	**2.21**	10.28	2.55	**122**
Latino-all	26.37	28.12	-1.75	-1.08	-1.52	-0.90	932

EVER ATTENDED PRIVATE SCHOOLS IN THREE YEARS

Overall	**0.77**	**0.11**	**0.66**	**35.73**	**0.66**	**35.98**	**2,117**
Black- all	**0.81**	**0.10**	**0.71**	**28.15**	**0.70**	**27.23**	**903**
Black-ch0	**0.81**	**0.18**	**0.63**	**11.23**	**0.57**	**9.63**	**192**
Black-ch1	**0.82**	**0.04**	**0.78**	**16.01**	**0.78**	**15.21**	**176**
Black-ch2	**0.82**	**0.07**	**0.75**	**17.69**	**0.75**	**16.01**	**206**
Black-ch3	**0.76**	**0.11**	**0.65**	**12.25**	**0.65**	**12.63**	**182**
Black-ch4	**0.82**	**0.06**	**0.76**	**15.77**	**0.77**	**15.44**	**146**
Latino-all	**0.77**	**0.11**	**0.66**	**25.23**	**0.66**	**25.53**	**1,021**

Notes: ch-0 refers to Kindergartners at baseline, ch-1 to first graders at baseline, and so on. Observations with missing values for a particular variable are dropped from the sample in the relevant row. Bootstrap standard errors used to compute the t-ratios account for dependent observations within families. Lottery strata are the 30 groups students were placed in at the time of random assignment. Observations are weighted using revised baseline weights.

A bold font indicates that the absolute t-ratio for the within-strata treatment-control difference exceeds 1.96.

Table 2: Efficiency Comparison of Two Methods of Random Assignment
Estimated Treatment Effect and Standard Error By Method of Random Assignment

A. All Students

Year	Model	Propensity Score Match Subsample			Stratified Blocks Subsample			Relative Std. Error	Sample-Size-Adjusted Relative Std. Error
		Coeff.	Std. Error	Nobs.	Coeff.	Std. Error	Nobs.		
Year 1	Control for baseline	-0.12	1.51	721	2.50	1.52	734	1.007	1.016
	Omit baseline	0.19	1.63	1048	-0.87	1.70	1032	1.043	1.035
Year 2	Control for baseline	0.89	1.52	603	0.30	1.84	596	1.207	1.200
	Omit baseline	0.76	1.61	908	-1.05	1.81	846	1.123	1.084
Year 3	Control for baseline	1.04	1.58	613	0.78	1.79	637	1.133	1.155
	Omit baseline	-0.18	1.57	900	-0.31	1.79	901	1.140	1.141

B. African American Students

Year	Model	Propensity Score Match Subsample			Stratified Blocks Subsample			Relative Std. Error	Sample-Size-Adjusted Relative Std. Error
Year 1	Control for baseline	-0.10	1.79	302	9.29	1.80	321	1.006	1.037
	Omit baseline	2.42	2.00	436	4.40	2.23	447	1.115	1.129
Year 2	Control for baseline	2.81	2.10	244	3.80	2.87	253	1.369	1.394
	Omit baseline	4.33	2.25	360	0.63	2.53	362	1.129	1.132
Year 3	Control for baseline	3.78	2.56	247	6.45	2.17	272	0.848	0.890
	Omit baseline	2.23	2.47	347	3.32	2.36	386	0.955	1.008

Notes: Treatment effect coefficient is from a regression of test scores on a dummy indicating assignment to receive a voucher (1=yes), original 30 lottery randomization strata dummies, and in some models baseline test scores. Bootstrap standard errors account for within-family correlation in residuals. Year 1, 2, or 3 refers to follow-up year. Revised follow-up weights are used to weight observations.

Table 3a: Estimated treatment effects, with and without controlling for baseline scores
Controls for 30 randomization strata as defined by Mayer, Peterson and Myers, et al. (2000) and uses their follow-up weights

Test	Group	SubSample with Baseline Scores; Controls For Baseline Scores				SubSample with Baseline Scores; Omits Baseline Scores				Full Sample; Omits Baseline Scores			
First Follow-up Test:		NOBS	Coefficient	S.E.	t-ratio	NOBS	Coefficient	S.E.	t-ratio	NOBS	Coefficient	S.E.	t-ratio
Composite	Overall	1,455	1.20	0.97	1.24	1,455	0.23	1.30	0.18	2,080	-0.40	1.06	-0.38
	Black	623	**4.43**	**1.27**	**3.48**	623	**3.87**	**1.73**	**2.24**	883	2.64	1.42	1.86
	Latino	709	-0.66	1.38	-0.48	709	-2.59	1.77	-1.46	1,021	-2.05	1.56	-1.31
Reading	Overall	1,455	1.01	1.05	0.96	1,455	-0.18	1.38	-0.13	2,080	-1.17	1.18	-0.99
	Black	623	**3.47**	**1.57**	**2.21**	623	2.74	1.98	1.38	883	1.46	1.72	0.85
	Latino	709	-0.70	1.45	-0.48	709	-3.02	1.88	-1.60	1,021	-2.60	1.72	-1.51
Math	Overall	1,455	1.39	1.17	1.19	1,455	0.65	1.45	0.45	2,080	0.37	1.16	0.31
	Black	623	**5.39**	**1.52**	**3.56**	623	**5.01**	**1.92**	**2.60**	883	**3.82**	**1.54**	**2.48**
	Latino	709	-0.63	1.70	-0.37	709	-2.17	2.00	-1.08	1,021	-1.50	1.67	-0.90
Second Follow-up Test:													
Composite	Overall	1,199	0.46	1.13	0.41	1,199	0.14	1.44	0.10	1,754	0.05	1.22	0.04
	Black	497	3.27	1.68	1.95	497	3.51	1.94	1.81	722	2.58	1.70	1.52
	Latino	612	-0.60	1.50	-0.40	612	-2.34	1.97	-1.19	902	-1.83	1.62	-1.13
Reading	Overall	1,199	1.35	1.11	1.22	1,199	0.83	1.45	0.57	1,754	0.62	1.30	0.48
	Black	497	**3.44**	**1.74**	**1.97**	497	3.50	2.06	1.70	722	2.52	1.90	1.32
	Latino	612	0.17	1.55	0.11	612	-1.75	2.05	-0.85	902	-1.17	1.73	-0.68
Math	Overall	1,199	-0.43	1.51	-0.28	1,199	-0.55	1.73	-0.32	1,754	-0.52	1.39	-0.38
	Black	497	3.10	2.18	1.42	497	3.53	2.36	1.49	722	2.64	1.97	1.35
	Latino	612	-1.37	1.92	-0.72	612	-2.93	2.29	-1.28	902	-2.48	1.82	-1.36
Third Follow-up Test													
Composite	Overall	1,250	0.93	1.13	0.82	1,250	0.24	1.40	0.17	1,801	-0.31	1.18	-0.26
	Black	519	**5.50**	**1.61**	**3.42**	519	**5.30**	**1.91**	**2.77**	733	2.87	1.69	1.71
	Latino	637	-0.95	1.57	-0.60	637	-1.86	2.00	-0.93	932	-1.35	1.66	-0.81
Reading	Overall	1,250	0.27	1.25	0.21	1,250	-0.56	1.52	-0.37	1,801	-0.95	1.25	-0.77
	Black	519	**3.97**	**1.87**	**2.13**	519	3.56	2.18	1.63	733	1.53	1.83	0.84
	Latino	637	-1.85	1.73	-1.07	637	-2.90	2.15	-1.35	932	-1.81	1.78	-1.01
Math	Overall	1,250	1.59	1.28	1.24	1,250	1.05	1.52	0.69	1,801	0.33	1.32	0.25
	Black	519	**7.03**	**1.82**	**3.86**	519	**7.04**	**2.06**	**3.41**	733	**4.22**	**1.85**	**2.28**
	Latino	637	-0.05	1.82	-0.03	637	-0.82	2.19	-0.38	932	-0.89	1.83	-0.48

Notes: Dependent variable is test score NPR. Reported coefficient is coefficient on voucher offer dummy. All regressions control for 30 revised randomization strata. Bootstrap standard errors are robust to correlation in residuals among students in the same family. Bold font indicates that the absolute t-ratio exceeds 1.96.

Table 3b: Estimated treatment effects, with and without controlling for baseline scores
Controls for *original* 30 strata used to assign students to random assignment blocks and uses revised weights

Test	Group	SubSample with Baseline Scores; Controls For Baseline Scores				SubSample with Baseline Scores; Omits Baseline Scores				Full Sample; Omits Baseline Scores			
First Follow-up Test:		NOBS	Coefficient	S.E.	t-ratio	NOBS	Coefficient	S.E.	t-ratio	NOBS	Coefficient	S.E.	t-ratio
Composite	Overall	1,455	1.29	1.05	1.22	1,455	0.15	1.45	0.11	2,080	-0.34	1.18	-0.29
	Black	623	**4.70**	**1.30**	**3.63**	623	**4.35**	**1.76**	**2.47**	883	**3.42**	**1.47**	**2.33**
	Latino	709	-0.14	1.36	-0.11	709	-1.31	1.69	-0.78	1,021	-1.06	1.49	-0.71
Reading	Overall	1,455	1.20	1.08	1.11	1,455	0.02	1.46	0.01	2,080	-0.84	1.25	-0.67
	Black	623	**3.76**	**1.61**	**2.34**	623	3.35	2.05	1.64	883	2.30	1.82	1.26
	Latino	709	-0.14	1.45	-0.10	709	-1.64	1.83	-0.90	1,021	-1.50	1.65	-0.91
Math	Overall	1,455	1.37	1.32	1.03	1,455	0.29	1.69	0.17	2,080	0.17	1.34	0.12
	Black	623	**5.65**	**1.54**	**3.66**	623	**5.35**	**1.94**	**2.75**	883	**4.54**	**1.55**	**2.94**
	Latino	709	-0.15	1.70	-0.09	709	-0.99	1.94	-0.51	1,021	-0.62	1.60	-0.39
Second Follow-up Test:													
Composite	Overall	1,199	0.64	1.18	0.54	1,199	0.38	1.43	0.26	1,754	-0.14	1.22	-0.11
	Black	497	3.33	1.79	1.86	497	3.49	1.96	1.78	722	2.47	1.69	1.46
	Latino	612	-0.52	1.58	-0.33	612	-1.64	1.97	-0.83	902	-2.03	1.61	-1.25
Reading	Overall	1,199	1.49	1.15	1.29	1,199	1.11	1.47	0.76	1,754	0.41	1.30	0.32
	Black	497	**3.72**	**1.77**	**2.10**	497	3.77	2.02	1.86	722	2.35	1.82	1.29
	Latino	612	0.38	1.66	0.23	612	-0.94	2.09	-0.45	902	-1.39	1.76	-0.79
Math	Overall	1,199	-0.20	1.57	-0.13	1,199	-0.36	1.73	-0.21	1,754	-0.69	1.40	-0.50
	Black	497	2.94	2.39	1.23	497	3.21	2.49	1.29	722	2.59	2.05	1.27
	Latino	612	-1.42	1.98	-0.72	612	-2.34	2.27	-1.03	902	-2.66	1.79	-1.48
Third Follow-up Test													
Composite	Overall	1,250	0.95	1.16	0.81	1,250	0.52	1.41	0.37	1,801	-0.25	1.20	-0.21
	Black	519	**5.03**	**1.67**	**3.02**	519	**5.00**	**1.95**	**2.57**	733	2.78	1.69	1.64
	Latino	637	-1.00	1.61	-0.62	637	-1.63	1.98	-0.82	932	-1.52	1.68	-0.90
Reading	Overall	1,250	0.40	1.30	0.31	1,250	-0.16	1.55	-0.11	1,801	-0.73	1.26	-0.58
	Black	519	3.65	1.95	1.88	519	3.37	2.24	1.50	733	1.55	1.86	0.84
	Latino	637	-1.71	1.80	-0.95	637	-2.50	2.15	-1.16	932	-1.81	1.81	-1.00
Math	Overall	1,250	1.49	1.33	1.12	1,250	1.21	1.54	0.78	1,801	0.23	1.35	0.17
	Black	519	**6.42**	**1.87**	**3.42**	519	**6.62**	**2.11**	**3.13**	733	**4.00**	**1.87**	**2.14**
	Latino	637	-0.29	1.84	-0.16	637	-0.77	2.18	-0.38	932	-1.22	1.83	-0.67

Notes: Dependent variable is test score NPR. Reported coefficient is coefficient on voucher offer dummy. All regressions control for original 30 randomization strata. Bootstrap standard errors are robust to correlation in residuals among students in the same family. Bold font indicates that the absolute t-ratio exceeds 1.96.

Table 4: Estimated treatment effects, with varying controls for baseline characteristics

Test	Group	SubSample with Baseline Scores Controls for baseline scores, cohort & other covariates				Full Sample Controls for cohort & other covariates (omits baseline test scores)				Full Sample Controls for cohort, baseline scores & interaction for missing scores			Full Sample Controls for cohort, baseline scores, interaction for missing scores & other covariates		
First Follow-up Test:		NOBS	Coefficient	S.E.	t-ratio	NOBS	Coefficient	S.E.	t-ratio	Coefficient	S.E.	t-ratio	Coefficient	S.E.	t-ratio
Composite	Overall	1,455	1.72	0.99	1.75	2,080	0.15	1.06	0.15	0.68	0.95	0.71	0.88	0.90	0.97
	Black	623	**4.12**	**1.35**	**3.06**	883	**3.28**	**1.43**	**2.30**	**3.69**	**1.25**	**2.96**	3.47	1.26	**2.76**
	Latino	709	0.05	1.34	0.04	1,021	-1.21	1.36	-0.89	0.17	1.20	0.14	-0.02	1.18	-0.02
Reading	Overall	1,455	1.25	1.06	1.18	2,080	-0.60	1.15	-0.53	0.16	1.05	0.15	0.17	1.01	0.17
	Black	623	2.67	1.66	1.61	883	1.75	1.74	1.01	2.49	1.61	1.55	2.04	1.58	1.29
	Latino	709	-0.10	1.39	-0.07	1,021	-1.47	1.58	-0.93	0.06	1.37	0.04	-0.14	1.37	-0.10
Math	Overall	1,455	2.20	1.25	1.76	2,080	0.91	1.20	0.76	1.20	1.10	1.09	1.58	1.06	1.49
	Black	623	**5.58**	**1.57**	**3.55**	883	**4.80**	**1.51**	**3.17**	**4.89**	**1.33**	**3.66**	4.90	1.37	**3.59**
	Latino	709	0.20	1.74	0.12	1,021	-0.96	1.48	-0.65	0.28	1.35	0.20	0.09	1.35	0.07
Second Follow-up Test:															
Composite	Overall	1,199	0.53	1.14	0.46	1,754	-0.70	1.16	-0.60	0.07	1.05	0.07	-0.41	1.00	-0.41
	Black	497	**3.42**	**1.69**	**2.02**	722	2.39	1.68	1.42	2.42	1.59	1.52	2.19	1.50	1.46
	Latino	612	-1.31	1.57	-0.83	902	-2.83	1.49	-1.90	-1.35	1.33	-1.02	-1.86	1.29	-1.44
Reading	Overall	1,199	1.57	1.16	1.35	1,754	0.07	1.24	0.06	0.70	1.11	0.63	0.40	1.08	0.37
	Black	497	**3.94**	**1.88**	**2.09**	722	2.32	1.84	1.26	2.54	1.65	1.54	2.23	1.68	1.32
	Latino	612	-0.07	1.72	-0.04	902	-1.83	1.70	-1.08	-0.57	1.49	-0.38	-0.77	1.48	-0.52
Math	Overall	1,199	-0.51	1.50	-0.34	1,754	-1.47	1.34	-1.09	-0.56	1.27	-0.44	-1.21	1.22	-1.00
	Black	497	2.90	2.10	1.38	722	2.45	1.97	1.25	2.30	1.99	1.15	2.15	1.81	1.19
	Latino	612	-2.56	1.99	-1.29	902	**-3.82**	**1.64**	**-2.33**	-2.13	1.54	-1.38	-2.94	1.49	-1.98
Third Follow-up Test															
Composite	Overall	1,250	1.25	1.10	1.13	1,801	-0.32	1.16	-0.28	0.15	1.04	0.15	-0.02	1.01	-0.02
	Black	519	**4.28**	**1.64**	**2.61**	733	2.08	1.69	1.24	2.76	1.55	1.78	2.13	1.51	1.41
	Latino	637	-0.88	1.56	-0.56	932	-1.50	1.61	-0.94	-0.79	1.41	-0.56	-0.83	1.41	-0.59
Reading	Overall	1,250	0.62	1.25	0.50	1,801	-0.83	1.23	-0.67	-0.21	1.10	-0.19	-0.46	1.09	-0.42
	Black	519	3.02	1.93	1.57	733	0.78	1.84	0.42	1.72	1.70	1.01	1.02	1.66	0.61
	Latino	637	-1.76	1.76	-1.00	932	-1.78	1.78	-1.00	-0.90	1.55	-0.58	-1.06	1.58	-0.67
Math	Overall	1,250	1.88	1.28	1.47	1,801	0.18	1.30	0.14	0.52	1.22	0.43	0.43	1.19	0.36
	Black	519	**5.54**	**1.88**	**2.94**	733	3.39	1.89	1.79	**3.79**	**1.75**	**2.16**	3.24	1.74	1.86
	Latino	637	0.01	1.81	0.01	932	-1.23	1.76	-0.70	-0.68	1.60	-0.43	-0.60	1.59	-0.38

Notes: Dependent variable is test score NPR. Reported coefficient is coefficient on voucher offer dummy. All regressions control for original 30 randomization strata and cohort (initial grade) dummies. The first set of columns also control for baseline test scores. The last two sets of columns control for baseline test scores and a dummy indicating whether baseline scores are missing and an interaction between that dummy and baseline test scores. When indicated, "other covariates" included are gender, mother's education, mother's full-time or part-time employment status, special education, gifted or talented class, welfare, log family income, student's age, mother's place of birth, English spoken at home, mother in current residence less than one year, mother's religion Catholic, and dummies indicating whether the covariates are missing.

Bootstrap standard errors account for dependent observations among students in the same family. Bold font indicates absolute t-ratio in excess of 1.96.

Table 5: Estimated treatment effects for all Black students
Sample includes students whose mother or father is Black/African American

Test/Year	SubSample w/ Baseline Test — Controls for baseline scores & covariates				Full Sample — No controls except randomization strata				Full Sample — Controls for covariates except baseline scores			Full Sample — Controls for covariates, baseline scores & interaction for missing scores		
	N	Coefficient	S.E.	t-ratio	N	Coefficient	S.E.	t-ratio	Coefficient	S.E.	t-ratio	Coefficient	S.E.	t-ratio
Year One														
Composite	687	3.52	1.30	**2.71**	974	2.35	1.53	1.54	2.24	1.46	1.53	2.63	1.21	2.17
Reading	687	2.27	1.53	1.48	974	1.36	1.82	0.75	0.83	1.72	0.48	1.28	1.49	0.86
Math	687	**4.77**	**1.62**	**2.94**	974	**3.34**	**1.63**	**2.05**	**3.64**	**1.59**	**2.29**	**3.98**	**1.38**	**2.89**
Year Two														
Composite	548	2.04	1.65	1.23	793	1.36	1.63	0.83	0.96	1.63	0.59	1.10	1.43	0.77
Reading	548	2.91	1.77	1.15	793	1.57	1.81	0.87	1.20	1.80	0.67	1.39	1.61	0.86
Math	548	1.16	2.04	0.57	793	1.15	1.93	0.59	0.72	1.89	0.38	0.81	1.70	0.48
Year Three														
Composite	573	**4.00**	**1.60**	**2.50**	811	2.01	1.67	1.20	1.14	1.64	0.69	1.44	1.43	1.01
Reading	573	2.98	1.87	1.59	811	0.99	1.84	0.54	0.01	1.77	0.00	0.45	1.59	0.28
Math	573	**5.02**	**1.87**	**2.68**	811	3.04	1.85	1.64	2.27	1.86	1.22	2.43	1.66	1.46

Notes: Dependent variable is test score NPR. Reported coefficient is coefficient on voucher offer dummy. Regressions in Column 1, 3, and 4 control for 30 randomization strata,cohort dummies, gender, mother's education, mother's full-time or part-time employment status, special education, gifted or talented class, welfare, log family income, student's age, mother's place of birth, English spoken at home, mother in current residence less than one year, and mother's religion. The first set of columns also control for baseline test scores. The last set of columns also control for a dummy indicating baseline scores missing and an interaction between that dummy and baseline test scores.

Bootstrap standard errors account for dependent observations among students in the same family. Bold font indicates absolute t-ratio in excess of 1.96.

Table 6: Two-Stage Least Squares estimates of effect of years in private school, with varying controls for baseline covariates

Test	Group	SubSample with Baseline Scores Controls for baseline scores & covariates				Full Sample; Mother's Race/Ethnicity Controls for covariates, baseline test scores & interaction for missing scores				Full Sample; Either Parent's Race/Ethnicity Controls for covariates, baseline test scores & interaction for missing scores			
First Follow-up Test:		NOBS	Coefficient	S.E.	t-ratio	NOBS	Coefficient	S.E.	t-ratio	NOBS	Coefficient	S.E.	t-ratio
Composite	Overall	1,449	2.36	1.35	1.74	2,080	1.23	1.32	0.93	2,080	1.23	1.32	0.93
	Black	622	**5.39**	**1.79**	**3.01**	883	**4.70**	**1.73**	**2.71**	974	**3.57**	**1.65**	**2.17**
	Latino	704	0.05	1.87	0.03	1,021	-0.04	1.72	-0.02	1,145	0.26	1.72	0.15
Reading	Overall	1,449	1.73	1.43	1.21	2,080	0.24	1.46	0.17	2,080	0.24	1.46	0.17
	Black	622	3.49	2.15	1.62	883	2.77	2.15	1.29	974	1.75	2.00	0.87
	Latino	704	-0.11	1.96	-0.06	1,021	-0.21	2.00	-0.11	1,145	-0.38	1.98	-0.19
Math	Overall	1,449	2.99	1.74	1.72	2,080	2.22	1.55	1.43	2,080	2.22	1.55	1.43
	Black	622	**7.29**	**2.11**	**3.46**	883	**6.62**	**1.89**	**3.51**	974	**5.40**	**1.90**	**2.84**
	Latino	704	0.22	2.43	0.09	1,021	0.14	1.95	0.07	1,145	0.90	1.93	0.47
Second Follow-up Test:													
Composite	Overall	1,199	0.35	0.76	0.46	1,754	-0.28	0.69	-0.41	1,754	-0.28	0.69	-0.41
	Black	497	**2.13**	**1.05**	**2.03**	722	1.43	0.99	1.44	793	0.72	0.93	0.77
	Latino	612	-0.91	1.08	-0.84	902	-1.29	0.89	-1.46	1,006	-1.37	0.89	-1.54
Reading	Overall	1,199	1.04	0.77	1.35	1,754	0.27	0.75	0.36	1,754	0.27	0.74	0.37
	Black	497	**2.45**	**1.17**	**2.10**	722	1.45	1.12	1.30	793	0.91	1.06	0.86
	Latino	612	-0.05	1.18	-0.04	902	-0.54	1.01	-0.53	1,006	-1.12	1.01	-1.11
Math	Overall	1,199	-0.34	0.99	-0.34	1,754	-0.83	0.83	-1.00	1,754	-0.83	0.83	-1.00
	Black	497	1.80	1.29	1.39	722	1.41	1.18	1.20	793	0.53	1.11	0.48
	Latino	612	-1.76	1.36	-1.30	902	**-2.05**	**1.03**	**-1.99**	1,006	-1.62	1.04	-1.57
Third Follow-up Test													
Composite	Overall	1,250	0.62	0.55	1.13	1,801	-0.01	0.53	-0.02	1,801	-0.01	0.53	-0.02
	Black	519	**2.12**	**0.82**	**2.58**	733	1.11	0.79	1.41	811	0.74	0.73	1.01
	Latino	637	-0.44	0.78	-0.56	932	-0.43	0.72	-0.59	1,039	-0.46	0.70	-0.66
Reading	Overall	1,250	0.31	0.62	0.50	1,801	-0.24	0.57	-0.42	1,801	-0.24	0.57	-0.42
	Black	519	1.50	0.95	1.57	733	0.53	0.86	0.62	811	0.23	0.82	0.28
	Latino	637	-0.89	0.88	-1.00	932	-0.54	0.81	-0.67	1,039	-0.51	0.78	-0.65
Math	Overall	1,250	0.93	0.64	1.46	1,801	0.23	0.62	0.36	1,801	0.23	0.62	0.36
	Black	519	**2.74**	**0.95**	**2.89**	733	1.69	0.92	1.85	811	1.25	0.86	1.46
	Latino	637	0.01	0.91	0.01	932	-0.31	0.82	-0.38	1,039	-0.41	0.78	-0.53

Notes: Dependent variable is test score NPR. Reported coefficient is coefficient on years in private school. All regressions control for 30 randomization strata, cohort dummies, gender, mother's education, mother's full-time or part-time employment status, special education, gifted or talented class, welfare, log family income, student's age, mother's place of birth, English spoken at home, mother in current residence less than one year, mother's religion, and dummies indicating whether the covariates are missing. The first set of columns also control for baseline test scores. The last two sets of columns also control for a dummy indicating whether baseline scores are missing and an interaction between that dummy and baseline test scores.

Bootstrap standard errors account for dependent observations among students in the same family. Bold font indicates absolute t-ratio in excess of 1.96.