Efficiency, Bias, and Classification Schemes: Estimating Private-School Impacts on Test Scores in the New York City Voucher Experiment

by

Paul E. Peterson and William G. Howell

Harvard University

June 2003

Paper prepared for publication in *The American Behavioral Scientist* (forthcoming).

Efficiency, Bias, and Classification Schemes: Estimating Private-School Impacts on Test Scores in the New York City Voucher Experiment

Paul E. Peterson and William G. Howell
(Executive Summary)

School vouchers are perhaps the most controversial policy reform in education today. Much public debate on this issue is polarized, as ideological posturing regularly substitutes for social scientific inquiry. Fortunately, in the last decade a wealth of new information has become available about the educational experiences of students in small, targeted voucher programs, including the program operated by the School Choice Scholarships Foundation (SCSF) in New York City.

A new analysis of data from the SCSF evaluation yields results consistent with the findings reported in *The Education Gap: Vouchers and Urban Schools* (Brookings 2002). As shown in Table 1, 108 out of 120 separate estimates indicate that attendance at a private school had significantly positive effects on African Americans' test scores but had no effects, positive or negative, on those of Hispanics or other ethnic groups. Princeton University's Alan Krueger and Pei Zhu have purported to show otherwise, and their research has been given wide media coverage. Unfortunately, the analytical methods they used in reaching these conclusions are either unreliable, at risk of bias, and/or based upon a problematic classification system that deviates from federally approved guidelines.

In the spring of 1997, more than 1,200 New York City public school students in grades K–4 were offered vouchers worth up to $1,400 annually to help pay the cost of private school. The vouchers were initially guaranteed for three years. As vouchers were awarded randomly, SCSF could be studied as a randomized field trial. To facilitate the evaluation, the research team collected baseline test scores and other data prior to the lottery, administered the lottery, and then collected follow-up information one, two, and three years later.

We previously reported that African Americans in private schools who were retested after one, two, and three years scored, on average, 6.1, 4.2, and 8.4 National Percentile Rank (NPR) points higher than their peers in public schools on the combined reading and math portions of the Iowa Test of Basic Skills.

Krueger and Zhu, however, say that positive estimates of voucher impacts are overstated. In making this claim, they highlight two aspects of their re-analysis: the inclusion of students for whom no baseline test scores are available and the reclassification of students' ethnicity. But neither of these two alterations to the analysis, separately or together, changes the results for African Americans. When students without baseline scores are added to the study, the impact of attending a private school remains significantly positive. And even when students are defined as African American when either their mother or their father is African American, as Krueger and Zhu recommend, significant effects are observed in all three years for students with baseline scores, and in years one and three when all students together are considered.

Both of these procedural changes are nonetheless problematic. Including students without baseline scores is risky because it is then unknown whether students in the treatment and control

groups have similar benchmark scores before the experiment began. And when Krueger and Zhu define some Hispanics as African Americans, they deviate from clear federal guidelines as to how persons are to be classified—the very guidelines, in fact, they cite to support their case. Krueger and Zhu go on to classify students of mixed ethnic background as African American, even when the parental caretaker is from another ethnic group. This classification decision, applied only to African American students (and not to those from any other ethnic group) ignores the fact that most students live with their mothers and only 20 percent of the participating students lived in families where the parents were married.

Still, observed impacts drop below standard thresholds of statistical significance only when Krueger and Zhu introduce a third manipulation, the addition of 28 variables to statistical models of test-score gains. Adding all these variables to models is acceptable practice only when a theoretical justification is provided *ex ante*, the estimate becomes substantially more precise, and the treatment and control groups are shown to be balanced. Otherwise, missing information and imbalances in the data can worsen the estimates. Also, their inclusion may, as Krueger and Zhu point out, give rise to concerns that analysts are searching for model specifications that yield preferred results. In this case, such concerns are heightened because the models in the Krueger-Zhu paper presented at the National Press Club in April, 2003 differ from those presented by them at a conference at Yale University in August, 2002.

In sum, except for estimations based upon problematic model specifications and classification schemes, positive and significant impacts on the test score performance of African American students are routinely observed.

Efficiency, Bias, and Classification Schemes: Estimating Private-School Impacts on Test Scores in the New York City Voucher Experiment

Paul E. Peterson and William G. Howell

School vouchers are perhaps the most controversial policy reform in education today.[1] Fortunately, in the last decade a wealth of new information has become available about the educational experiences of students in small, targeted voucher programs. The School Choice Scholarships Foundation (SCSF) in New York City has furnished some of this information, as we and our colleagues originally reported in *The Education Gap: Vouchers and Urban Schools*.[2]

In the spring of 1997, students in grades K–4 who attended a public school and who were eligible for participation in the free-lunch program were invited to apply to SCSF for a school voucher that would help defray the cost of private-school tuition. More than 20,000 students expressed an interest in the program. Lotteries were held in May, and that fall students began using vouchers to attend private schools. Over 1,200 students were offered vouchers, which were worth up to $1,400 per annum and were initially guaranteed for three years. During the program's first year, 74 percent of families offered vouchers actually used them to send their children to private schools; after two and three years, 62 and 53 percent of the treatment group continued to attend private schools, respectively.[3]

Evaluation Procedures

Since vouchers were awarded by lot, the SCSF program could be evaluated as a randomized field trial. To facilitate the evaluation, the research team collected baseline test scores and other data prior to the lottery, administered the lottery, and then collected follow-up information one, two, and three years later. During the 1997 eligibility verification sessions attended by voucher applicants, students in grades 1–4 took the Iowa Test of Basic Skills (ITBS) in reading and mathematics.[4]

Scheduled during the months of February, March, and April immediately prior to the voucher lottery, sessions were held in private school classrooms, where schoolteachers and administrators served as proctors under the overall supervision of the evaluation team and program sponsors. While children were being tested, accompanying adults completed surveys of their satisfaction with their children's current public schools, their involvement in their children's education, and their demographic characteristics.[5] Over 5,000 students attended baseline sessions in New York City. Mathematica Policy Research (MPR) then administered the lottery in May and SCSF announced the winners.

To assemble a control group, approximately 960 families were randomly selected from those who did not win the lottery.[6] In the absence of administrative error, those offered vouchers should not differ significantly from members of the control group (those who did not win a voucher). Baseline test-score data, easily the best predictor of test-score outcomes (see below), confirm this expectation for those students for whom such data are available. For those students with baseline test scores, therefore, observed differences between the two groups' downstream test scores can safely be attributed to the programmatic intervention.

In the spring of 1998, the annual collection of follow-up information commenced. Testing and questionnaire procedures were similar to those administered at the baseline sessions. Adults accompanying the children again completed surveys that asked a wide range of questions about the educational experiences of their oldest child within the eligible age range. Students completed tests and short questionnaires in schools different from those they were then attending.

To ensure as high a response rate as possible, SCSF conditioned the renewal of scholarships on participation in the evaluation. Members of the control group and students in the treatment group who initially declined a voucher were compensated for their expenses and told that they could

automatically reenter a new lottery if they participated in follow-up sessions. Overall, 82 percent of students in the treatment and control groups attended the year-one follow-up session, as did 66 percent in year two, and 67 percent in year three.

Private-School Impacts on Test Scores

For non–African Americans, and for students taken as a whole, private schools did not have any discernible impact, positive or negative, on test scores. But for African Americans, substantial differences were observed in all three years. African Americans in private schools who were retested after one, two, and three years scored, on average, 6.1, 4.2, and then fully 8.4 National Percentile Rank (NPR) points higher on the combined reading and math portions of the Iowa Test of Basic Skills than their peers in public schools (see Table 2, row 1).[7] All of these effects are statistically significant. These findings are robust to alternative specifications and classification schemes. As summarized in Table 1, 108 of 120 different statistical models yield positive and significant effects.

These findings from New York are consistent with those of prior studies using observational data. Surveying the literature on school sector effects and private school vouchers, Princeton Economist Cecilia Rouse says that "the overall impact of private schools is mixed, [but] it does appear that Catholic schools generate higher test scores for African-Americans."[8] Jeffrey Grogger and Derek Neal, economists from the University of Wisconsin and the University of Chicago, respectively, find little in the way of detectable attainment gains for whites, but conclude that "urban minorities in Catholic schools fare much better than similar students in public schools."[9] Similarly, in our own research in Washington D.C., and Dayton, Ohio, we did not find any evidence of positive impacts for white students. In the second year of these evaluations, however, we did observe moderately large positive impacts for African Americans.[10] And in the tables of Alan Krueger and Pei Zhu's secondary analysis of the New York City voucher program, it can be seen that a clear

majority—30 of 51—of the estimations of the voucher impacts on the overall (composite) test scores of African Americans is significantly positive.[11]

Despite the weight of evidence available from the extant literature and from their own estimations, Krueger and Zhu express strong doubts that African Americans benefited from the New York City voucher intervention.[12] At one point in their essay, they suggest "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." In the end, however, Krueger and Zhu back away from this statement, asserting only that "the safest conclusion is probably that the provision of vouchers did not lower the test scores of African Americans"—or, equivalently, that African American students who used vouchers to attend private schools performed as well or better than their peers in public school.[13]

How did Krueger and Zhu generate findings that justify their conclusion? Three analytical decisions stand out as most consequential: 1) add students without baseline scores to the analysis, despite the risk of obtaining a biased estimate of the program's effects; 2) employ an unusual, problematic classification scheme that identifies as "black/African American (non-Hispanic)" two select groups of students, one whose parents have identified themselves as black/Hispanic or a similar self-designation, and another whose fathers are African American but the parental caretaker is not, despite the fact that other ethnic groups are not classified in this manner; and 3) add 28 additional variables to the statistical models, despite their own admitted warnings against "specification searching," rummaging theoretically barefoot through data in the hopes of finding desired results.

The mere addition of students without baseline scores—the analytic decision that has received the most media attention and the one that Krueger and Zhu claim to be the "most

important" evidence in support of null findings—does not, by itself, provide a basis for their conclusion. Results remain significantly positive for African American students in all three outcome years when these students are added to the study (see Table 2, row 6). Nor do results change materially if one takes a second step upon which Krueger and Zhu place great weight, the reclassification of students as African American when either their mother or their father is African American. When these observations are added to estimations of voucher effects for African-American test scores, they remain significantly positive in all years for students with baseline test scores—and in years one and three, if students without baseline scores are included in the analysis.

Although these methodological innovations do not, by themselves, significantly alter the results, both are problematic. Adding students for whom no test scores are available at baseline raises the risk of introducing bias. Classifying students as African American whenever their fathers are African American, even when he is not the parental caretaker, ignores the fact that the vast majority of children in the sample live with their mothers, only 20 percent of whom are married. Still other questionable steps are taken by Krueger and Zhu. They classify students by ethnicity in a unique way that deviates from federal guidelines. And they include numerous background variables that do not improve the precision of the findings and increase the risk of bias. For these and other reasons, our original findings still provide the best estimate of the effects of private-school attendance in the New York City voucher program on various programmatic outcomes, including test scores.[14]

Issue #1: How Important Are Baseline Test Scores?

In a study of student achievement, of all information to be collected at baseline, the most critical is test scores. As stated in the project proposal prepared before any outcome data had been collected, "The math and reading achievement tests completed by students [at baseline] will provide

a benchmark against which to compare students' future test scores."[15] Such a benchmark is critical. More than any other information collected, baseline test scores have the highest correlations with test score outcomes—0.7, 0.6, and 0.7 for years one, two and three, respectively. None of the correlations logged by demographic variables is even half as large.[16]

Unfortunately, Mathematica Policy Research (MPR), the firm that administered the evaluation, was not able to obtain test-score data for everyone at baseline.[17] Some students in grades 1–4 were sick, others refused to take the test, and some tests were lost in the administrative process.[18] And due to the substantial difficulties of testing students who lacked reading skills, no kindergartners were tested at baseline.[19]

So as to follow the intended research plan and use the highest quality data, the original analyses of test scores were limited to those for whom benchmark data were available. For African American students with available baseline test scores (the Available Tests at Baseline, or the ATB group), one observes moderately large impacts of attending a private school on the combined math and reading portions of the Iowa Test of Basic Skills.[20] As stated previously, effects are 6.1, 4.2, and 8.4 percentile points in years one, two and three—all of which are statistically significant (see Table 2, row 1).[21, 22] The estimated impacts of private-school attendance on test scores remains significantly positive when students without baseline test scores (No Available Tests at Baseline or NATB group) are added to the analysis. The magnitude of the estimations, however, attenuates because the test scores of African American NATBs were affected either trivially or negatively by attending a private school. For African American NATBs, impacts are 0.1, -3.5, and -13.3 NPR points in years one, two, and three respectively, none of which is statistically significant (see Table 2, row 3).

The differences in results for the ATBs and the NATBs are sufficiently striking to raise questions about which set of data has greater credibility. Consider the following thought experiment: two randomized experiments are conducted, one for a larger number of cases with baseline test scores, the other for fewer cases without this crucial baseline information. The two studies yield noticeably different results. Which of the two should be given greater weight by policy analysts? Unless the baseline data reveal a departure from random assignment, we doubt any scientist would give greater credence to the study set lacking such crucial baseline information.

The thought experiment is a useful exercise because it underscores the fact that concerns about bias arise whenever key baseline information is missing. For ATBs, we have solid grounds for concluding that estimations are unbiased, simply because we know the treatment and control groups do not differ significantly in their baseline test scores. Only a minuscule, statistically insignificant 0.4 NPR points differentiate the composite baseline scores of African American students in the treatment and control groups.[23] But if there seems to be little danger of bias among ATBs, the same cannot be said for NATBs, which may have initially been—or subsequently became—significantly unbalanced. Krueger and Zhu argue otherwise, saying that "because of random assignment . . . estimates are unbiased." But estimates are unbiased only if the randomization process worked as well for the NATBs as it did for the ATBs—an outcome that Krueger and Zhu assume but cannot show.[24] And without verification that the randomization process worked, this assumption should be treated with considerable skepticism, especially given certain attributes of the New York experiment.

The administration of the New York experiment was exceedingly complicated, as Krueger and Zhu themselves admit. Half the sample was selected by means of a matching propensity score design, half by stratified sampling that took into account the date students took the test, the quality of the public school they came from, and the size of the family applying for a voucher. Because many

more students and families came to the testing sessions than were eventually included in the control group, lotteries proceeded in two steps: lottery winners first were drawn randomly, and then a second sample was drawn randomly from non-winners for inclusion in the experiment.

For ATBs taken as a whole, we know that administrative complications did not generate significant test-score differences at baseline. Unfortunately, no information on this crucial point is available for the NATBs. We do know, however, that along a variety of other dimensions (whether a student came from an under-performing public school, the student's gender, and whether the mother graduated from college), significant differences between NATBs in the treatment and control groups are observed. Whether these imbalances extend to NATB test scores is impossible to know.

After the initial lotteries, additional administrative errors may have occurred. For one thing, matching student names from one year to the next presented numerous complications. For ATB students, the risk of mismatching was reduced because students put their own names on the baseline test and all subsequent tests they took. But for NATBs, student identification at baseline could be obtained only from parent surveys, which then had to be matched with information the child gave on tests taken in subsequent years. NATB parents, furthermore, were less likely to complete survey questionnaires than ATB parents. Background information is missing for 38 percent of NATBs, as compared to 29 percent of ATBs, a difference that is statistically significant at $p<.01$.[25]

The seemingly mundane job of matching students actually presented multiple challenges. In a low-income, urban, predominantly single-parent population, children's surnames often do not match that of both their parents; children may take their mother's maiden name, their father's name, the name of a stepfather, or of someone else altogether. Also, students may report one or another nickname on follow-up tests, while parents report the student's formal name. Without documentation completed by students at baseline, ample opportunities arise for mismatching parent

survey information at baseline and child self-identification in years one, two, and three—raising further doubts about the reliability of the NATB data.[26]

Finally, attrition from the experiment introduces additional risks of bias.[27] Even though response rates of the ATB and NATB treatment and controls are comparable for each year, this hardly guarantees that the balance observed in the former population extends to the latter. Ultimately, and unfortunately, one cannot rule out the possibility that attrition compromised the baseline test-score balance between NATB treatment and control groups.

For the moment, though, let us set aside the possibilities of bias arising due to administrative error or differential attrition. What, exactly, is to be gained from introducing the NATBs to the analysis? Krueger and Zhu suggest two potential benefits: the ability to generalize findings to another grade level (kindergartners) and the efficiency gains associated with estimating models with larger sample sizes. On the former score, the kindergartners appear to be quite different from their older peers, making any such generalization hazardous. African American students in grades 1-4 posted significant and positive test score gains (whether or not one includes the NATBs in the analysis, and whether or not controls for baseline test scores are included) in all three years. Impacts for kindergartners, meanwhile, appear more erratic, bottoming out at -13.9 in year three.[28]

At first glance, however, Krueger and Zhu appear justified when espousing the benefits of enlarging the number of available observations. All else equal, the precision of estimated impacts increases with sample size. The problem, of course, is that all else is not equal. And the efficiency gains associated with increasing the number of observations do not make up for the losses associated with not being able to control for baseline test scores.[29] Among African American ATBs, the standard errors for impacts in years one, two, and three in test score models that do not include baseline test scores are 2.3, 2.4, and 3.2 (see Table 2, row 2). When controls for baseline test scores

are added, the standard errors drop noticeably to 1.7, 1.9, and 2.5 for the three years (Table 2, row 1). But when expanding the sample to include both ATBs and NATBs and dropping controls for baseline test scores, the standard errors jump back up to 2.0, 2.2, and 3.0 (Table 2, row 4). As the English would put it, what is gained on the straightaway is more than lost on the roundabouts.

To make up for these efficiency losses, Krueger and Zhu employ a hybrid model that controls for baseline test scores whenever possible (see Krueger and Zhu 2003, equation 3). Unfortunately, Krueger and Zhu never estimate this model in a manner that allows for straightforward comparisons with the impacts originally reported. Instead, they estimate the hybrid model only after recoding the ethnic identity of some African Americans and adding numerous other demographic controls and missing-data indicators (on these issues, see below). But when one estimates a simple, transparent hybrid model that just controls for baseline test scores, whenever possible, results are only marginally different from those originally reported (see Table 2, rows 5 and 6).[30]

Recall, Krueger and Zhu insist that the inclusion of NATBs in the analysis yields the "most important" evidence that the originally reported findings are "less robust than commonly acknowledged." Yet when NATBs are included in the sample, but controls for baseline test scores are preserved, one observes statistically significant, positive impacts of private-school attendance on the test scores of African Americans in all three years. To substantiate their objection to our original findings, Krueger and Zhu cannot rely on simple, transparent results that merely add those students without baseline scores to the estimations. In fact, they must make additional methodological moves, the next being the introduction of a flawed ethnic classification scheme.

Issue #2: Who Is African American?

In the New York evaluation, families' ethnic backgrounds were ascertained from information provided in the parent questionnaire.[31] At baseline (and, again, at the year-two and year-three

follow-up sessions), accompanying adults were asked to place the student's mother into one of the following ethnic groups:[32] 1) Black/African American (non-Hispanic); 2) White (non-Hispanic); 3) Puerto Rican; 4) Dominican; 5) Other Hispanic (Cuban, Mexican, Chicano, or other Latin American); 6) American Indian or Alaskan Native; 7) Chinese; 8) Other Asian or Pacific Islander (Japanese, Korean, Filipino, Vietnamese, Cambodian, Indian/Pakistani, or other Asian); 9) Other (Write in: _____).

Students of "other" background. In most instances, one can easily infer each student's ethnicity based upon the parent's; for a limited number of cases, however, judgment is required. Should those classified as "other" be reclassified into one of the listed categories? If so, which category? Much, of course, depends upon whether a parent selected the "other" category intentionally or inadvertently. For example, if respondents checked "other" but then claimed to be "Hispanic," it seems safe to assume that they overlooked the Hispanic category above, making reclassification appropriate. The same applies for anyone who inadvertently checked "other" but listed themselves as "African American" or "black." If, however, the "other" category appears chosen with some clear intention, then the respondent should, in our judgment, be left in that category. To do otherwise is not only to introduce inconsistencies but to display a lack of respect for the respondent's own ethnic self-identification.

At baseline, the ethnic background of 78 mothers and 73 fathers was identified as "other." Among those students for whom test score information is available beyond the baseline year, *none* of these parents can be reclassified as African American simply because a clear mistake was made by those completing the survey.[33] Rather, these parents identified themselves, quite intentionally, as black/Hispanic, black/Haitian, black/Dominican, black/Greek, black/Cuban, black/West Indies, and

so forth. Because none of these parents identified themselves simply as "African American" or "black," the safest classification decision is to preserve their self-identification as "other."

Krueger and Zhu, however, reclassify some of those in the "other" category as "Black/African American (non-Hispanic,)" even when the respondents themselves have rejected that label.[34] But it is very strange—indeed, contrary to the very federal guidelines that Krueger and Zhu use to bolster their case—to classify as "Black/African American (non-Hispanic)" people who openly identify themselves as "Hispanic," "Dominican," or "West Indian."

According to Office of Management and Budget (OMB) Statistical Directive 15, a person is to be defined as "Hispanic" if she is "of Mexican, Puerto Rican, Cuban, Central or South American or other Spanish culture or origin, regardless of race," while "a person is 'black'" if she is from "any of the black racial groups of Africa." The Directive goes on to say that if a "combined format is used to collect racial and ethnic data, the minimum acceptable categories are 'Black, not of Hispanic Origin,' 'Hispanic,' and 'White, not of Hispanic Origin,'" adding further that "any reporting . . . which uses more detail shall be organized in such a way that the additional categories can be aggregated into these basic racial/ethnic categories."[35]

To defend their classification scheme, Krueger and Zhu cite studies that indicate that "society treats individuals with different skin tones differently," a point that Krueger made more starkly when he identified the dark-skinned Dominican baseball player, Sammy Sosa, as "black" when displaying his picture in his National Press Club presentation of the Krueger-Zhu paper.[36] But the point to be taken away from this image is not that Sosa is "black" but that ethnicity does not reduce to "skin tones."[37] The "skin tones" of many Hispanic students in New York City are just as dark as those of many African Americans (just as the "skin tones" of many African Americans are as light as those of other ethnic groups, e.g., Pacific Islanders, Pakistanis, or Indians). Nothing in

OMB's Statistical Directive 15 says that Hispanics should be classified according to their skin color or any other physical attribute. To the contrary, the Directive says that if "race and ethnicity are collected separately, the number of White and Black persons who are Hispanic must be identifiable, and capable of being reported in that category."

Significantly, Krueger and Zhu did not use their new classification system in the paper originally presented at the Yale Conference.[38] In that document, they instead employed a probit model to estimate the percentage of Dominicans thought to be black, and then used the results of the model to recalculate voucher effects. In subsequent versions of the paper, Krueger and Zhu reconsidered their decision to shift of some of the Dominican students to their own definition of the category "Black/African-American (non-Hispanic)"—though they did revive the classification scheme in Krueger's oral presentation to the National Press Club, and then adopted yet another scheme that shifted Hispanic students to the African American category.[39]

Students of Mixed Ethnic Heritage. According to OMB's Statistical Directive 15, persons who are of mixed racial and/or ethnic origins should be placed in the category "which most closely reflects the individual's recognition in his community." The procedure we employed—classifying students by the ethnicity of the mother—is certainly consistent with the guideline, for the simple reason that in the overwhelming percentage of cases the mother is the person with whom the child lives. However, the guidelines might also be interpreted as allowing for the classification of students according to the ethnicity of the mother and father, taken together, or of the primary parental caretaker.

Eschewing these alternatives, Krueger and Zhu employ a unique classification scheme. They do not classify students of mixed heritage by their father's ethnicity (as the Chinese do).[40] Instead, they identify students of mixed heritage as African American on the basis of just one parent, either

the mother or the father. But when this principle is applied, no student can be classified as a member of any other ethnic group on the basis of just one parent's ethnicity (if the other parent is African American). Oddly, Kruger and Zhu defend this classification scheme on the grounds that it is "symmetrical." But symmetry is hardly the word for a scheme that classifies African Americans according to a different principle from that used to identify other ethnic groups.

In *The Education Gap,* we classified all students according to a single principle—students consistently were assigned to their mother's ethnic identification. To ensure that our classification was as inclusive as possible, we obtained information on the mother's ethnic identity in follow-up surveys whenever this information was missing at baseline. In educational research, use of the mother's ethnic identity as the basis for classification is especially appropriate. As we have previously argued, African Americans may benefit more from vouchers because in the public sector they enjoy fewer (and inferior) schooling options. A wide body of research, furthermore, shows that mothers strongly influence the educational outcomes of low-income, inner-city children.[41] And since low-income mothers are most involved in the educational lives of their children, it is the schooling options available to these mothers that matter most.

Several items in the parent questionnaire demonstrate the primary role that mothers played in the lives of the students participating in the study. Of the 792 ATB students with African American mothers who were tested in at least one subsequent year, 67 percent lived with their mother only, as compared to just 1 percent who lived only with their father.[42] The mothers of 74 percent of these students were single, divorced, separated, or widowed; in fact, only 20 percent of the children lived in families where the mother was married. Mothers accompanied 84 percent of children to testing sessions; and in 94 percent of the cases, the accompanying adult claimed to be a caretaker of the child. All of these factors point in the same direction—mothers, as an empirical fact, were most

responsible for the educational setting in which the children in this study were raised. Since the educational choices available to the mother are what matter most for the child, students should be classified according to her ethnicity.[43]

With this in mind, we show results in Table 3 from four classification schemes. The first three represent classification schemes that are consistent with federal guidelines. First, as done originally, we identify the students' ethnic background on the basis of their mothers' ethnic identity.[44] Second, we set aside all students that we know are of mixed background, so that there is almost no doubt that we are ascertaining effects for that group of students whose parents are both African American.[45] Third, we rely upon the ethnic identity of the parental caretaker (most frequently the mother, but occasionally the father). In all three years, and for all three plausible classification schemes, the same results emerge: private-school impacts on the test scores of African Americans, however defined, are positive and significant (see columns 1–3, Table 3).

Nor do the results change materially when students are identified as African American if their father or their mother is African American. Although inconsistent, this decision, by itself, is not sufficient to reach conclusions different from those originally reported. For all students with baseline test scores, statistically significant, positive impacts on African Americans are estimated in all three years; for all students, whether or not baseline scores are collected, significant, positive impacts are estimated in years one and three (see column 4, Table 3).

Krueger and Zhu state that "results are so sensitive" to alternative ways of classifying students as African American 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." But this assertion, far from being based on a robust set of alternative classification systems, depends upon one that not only departs from federal guidelines but partially erases the dividing line between

Hispanic and African American students, thereby obliterating the distinction originally observed to be significant.

Issue #3: Which Covariates Should Be Included in the Analysis?

Using hybrid models that take into account baseline scores, we have shown significantly positive impacts of private schooling on the test scores of all participating African American students (defined in various ways). Krueger and Zhu do not report these simple, transparent estimates. Instead, in the 2002 presentation at a Yale conference, their hybrid models include 12 other regressors (8 family and student characteristics and 4 missing variable indicators). In the 2003 paper presented before the National Press Club, they add 16 more (8 characteristics and 8 missing data indicators).[46]

When introducing these covariates, Krueger and Zhu impute means and include an indicator variable for those cases with missing values. Unfortunately, because they assume that all students for whom information is missing are exactly alike (with respect to the item in question), such methodological fixes introduce new opportunities for estimation bias.[47] Further, because the randomization process for the NATBs may have been compromised by administrative error or attrition from the study, the inclusion of additional background controls in regressions that include these students could exacerbate the bias. As Christopher Achen points out, when working with less-than-perfect randomized experiments, "controlling for additional variables in a regression may worsen the estimate of the treatment effect, even when the additional variables improve the specification."[48]

Given such risks, a good rule of thumb is to avoid adding a covariate unless *ex ante* theory justifies its inclusion, treatment and control groups are shown to be balanced, and significant gains in precision are achieved. As previously shown, inclusion of benchmark test scores passes these three

tests: *ex ante*, baseline scores were scheduled for inclusion because it was known that benchmark performance strongly predicts subsequent performance; baseline test scores of treatment and control groups remained balanced from baseline to the year three study; and the inclusion of baseline test scores as covariates substantially improves the precision of estimated treatment effects. The same, however, cannot be said for the 28 additional covariates that Krueger and Zhu introduce to the analysis.

Elsewhere in their essay, Krueger and Zhu themselves express doubts about models that include background controls. As they put it,

> 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.

This argument suggests that only baseline scores, the one variable identified in the project proposal as theoretically relevant, should be included in statistical models that estimate achievement gains.[49] Inasmuch as additional background controls were not introduced from the beginning of the research project, it is problematic to add them now.

The rules set forth by Krueger and Zhu, of course, apply to secondary analyses as well. Whenever possible, researchers should identify in advance the covariates to be included in their statistical models, especially when these covariates can artificially inflate or deflate the estimates. And when lists of covariates change over time—as Krueger and Zhu's have from the initial conference paper to the one released before the National Press Club—questions naturally arise about the possibility of specification searching.[50]

To show how results change when covariates are added, Table 4 reports third-year private-school impacts that control for different numbers of background control variables, for different classifications of African Americans, and for students with and without baseline test scores.

Columns 1–4 report estimated impacts for ATBs; columns 5–8 report impacts for ATBs and NATBs together.

For those African American students with baseline scores, the results do not change significantly when covariates are added (see columns 1–4). No matter how many additional regressors are successively added to the statistical models, positive and statistically significant impacts emerge.[51]

Inclusion of new covariates changes results only when the NATBs are added to the analysis (see columns 5–8). Even then, in most estimations results remain significantly positive when one adds just the covariates originally identified by Krueger and Zhu to be relevant (2002). Only when still further background characteristics are introduced do the effects of private-school attendance drop below standard thresholds of statistical significance. But with the addition of each new background characteristic, one after another, one repeatedly makes the assumption that all students with missing data are alike with regards to the item in question.

Since the inclusion of additional covariates requires strong assumptions, one should avoid them unless they add materially to the precision of the estimate. In this instance, it is not even a close call. The inclusion of additional covariates never reduces standard errors by more than a minuscule 0.04 NPR percentile points. And even these gains are obtained simply by controlling for the grade the student is in, which by itself does not alter significantly the results we reported. All the other covariates offer no efficiency gains whatsoever.[52] Indeed, for hybrid models that include all those with and without baseline scores, the addition of these covariates actually causes standard errors to increase in three of the four definitions of African American background. Far from providing a more "powerful" estimate, as Krueger and Zhu have claimed, the addition of all these variables usually has the opposite effect.

Concluding Observations

Further analysis of the data from the New York experiment reveals significant and positive private school impacts on the test scores of students with African Americans mothers under all of the following conditions:

1) In all three years, when effects are estimated for all those with baseline test scores, whether or not baseline test scores are controlled.

2) In all three years, when effects are estimated for all those with or without baseline test scores in grades 1–4 at baseline, controlling for baseline scores whenever possible.

3) In all three years, when effects are estimated for all those with or without baseline test scores in grades k–4 at baseline, controlling for baseline scores whenever possible.

4) In addition, significant, positive estimates are obtained under the following conditions:

    a. In all three years, when effects are estimated for all those who have both an African American mother and father, whether or not estimates are made only for those students with baseline test scores.

    b. In all three years, when effects are estimated for all those whose parental caretaker is an African American, whether or not estimates are made only for those students with baseline test scores.

    c. In all three years, when effects are estimated for those with baseline test scores for whom either the mother or the father is African American.

    d. In the first and third years, when effects are estimated for all those with or without baseline test scores for whom either the mother or the father is African American.

Plainly, the findings originally reported are robust to a wide variety of plausible specifications and classifications. In a few models they are positive, but not significant at conventional levels (see Table 1). These models, however, suffer from at least two of the following difficulties: 1) students for whom no baseline data were available were introduced into the analysis; 2) a novel, inconsistent ethnic classification scheme was employed; 3) the analysts, without *ex ante* theoretical justification and after conducting at least two separate specification searches, added to the

model 28 covariates for which much information is missing, a step which may have "worsen[ed] the estimate[s] of the treatment effect." In our view, there is no basis for privileging estimations dependent upon the few statistical models that employ at least two of these questionable approaches over the many others that have a superior scientific foundation.

What, then, can be learned of more general significance from this further analysis of the New York voucher experiment? The following come to mind:

1) Randomized experiments yield data that are less threatened by selection bias than most observational studies, but they are usually difficult undertakings in which administrative error is possible and sample attrition likely. To verify an experiment's integrity, baseline data on the key characteristic one is measuring are vital.

2) A randomized field trial is not strengthened by introducing observations that potentially disrupt the balance between treatment and control groups.

3) When classifying students by ethnicity, equivalent coding rules that follow standard (if possible, federally approved) practice should apply to students of different ethnic backgrounds, unless theory developed *ex ante* demands otherwise.

4) In the context of randomized field trials, simple, transparent models are to be preferred to complex models that contain many covariates for which numerous observations are missing. Covariates should only be added when *ex ante* theory justifies their inclusion, treatment and control groups are shown to be balanced, and significant gains in precision are achieved.

For these reasons, we conclude that the weight of the evidence from the evaluation of the New York voucher intervention lends further support to the finding—found repeatedly in both experimental and observational studies—that poor African American students living in urban environments benefit from private schooling.

ENDNOTES

[1] The many groups and individuals who assisted with the evaluation are acknowledged in William G. Howell, Paul E. Peterson with Patrick J. Wolf and David E. Campbell, *The Education Gap: Vouchers and Urban Schools* (Brookings 2002). Here we wish to thank as well those who have provided comments on this paper, including Alan Altshuler, Christopher Berry, David E. Campbell, Morris Fiorina, Jay Greene, Erik A. Hanushek, Frederick Hess, Caroline Minter Hoxby, Martin West, and Patrick J. Wolf.

[2] For full citation, see note 1. The volume also includes data from voucher experiments in other cities.

[3] In all three years, a small percentage of the control group (less than 5 percent) attended private schools.

[4] The assessment used in this study is Form M of the Iowa Test of Basic Skills, Copyright 1996 by The University of Iowa, published by The Riverside Publishing Company, 425 Spring Lake Drive, Itasca, Illinois 60143-2079. All rights reserved. The producer of the ITBS graded the tests.

[5] For a comprehensive analysis of these data, see Howell, Peterson, Wolf, and Campbell (2002).

[6] Exact procedures for the formation of the control group are described in Jennifer Hill, Donald B. Rubin, and Neal Thomas, "The Design of the New York School Choice Scholarship Program Evaluation." Paper presented at the American Political Science Association annual meeting. (Boston, Massachusetts, August 31, 1998).

[7] In total, 622, 497, and 519 African American ATBs were included in test score models after years one, two, and three, respectively. Model specifications provided in Table 2.

[8] Cecilia Elena Rouse, "School Reform in the 21st Century: A Look at the Effect of Class Size and School Vouchers on the Academic Achievement of Minority Students." Working Paper 440, (Princeton University, 2000), p. 19.

[9] Jeffrey Grogger and Derek Neal, "Further Evidence on the Effects of Catholic Secondary Schooling." In Brookings-Wharton Papers on Urban Affairs: 2000. Brookings Institution Press, 2000, p. 153. As early as 1985, Christopher Jencks determined that "the evidence that Catholic schools are especially helpful for initially disadvantaged students is quite suggestive, though not [at that time] conclusive." Christopher Jencks, "How Much Do High School Students Learn?" *Sociology of Education* Vol. 58 (April, 1985), p. 134. Also, see Derek Neal, "The Effects of Catholic Secondary Schooling on Educational Achievement," *Journal of Labor Economics* (1997) 15(1): 98-123. William N. Evans and Robert M. Schwab, "Who Benefits from Private Education? Evidence from Quantile Regressions," (Department of Economics, University of Maryland, 1993); David N. Figlio and Joe A. Stone, "Are Private Schools Really Better?" *Research in Labor Economics*, (JAI Press, Inc., 1999) 1(18): 115-140. Other studies finding positive educational benefits for African Americans from attending private schools include James S. Coleman, Thomas Hoffer, and Sally Kilgore, *High School Achievement* (New York: Basic Books, 1982); John E. Chubb and Terry M. Moe, *Politics, Markets, and America's Schools* (Washington: Brookings 1990);

William Sander, *Catholic Schools: Private and Social Effects* (Kluwer, 2000). As one reviewer of the Sander's book notes, "When both [experimental and non-experimental] types of studies yield similar conclusions, the results inspire greater confidence." R. Kenneth Godwin, "Choice Words," *Education Next*, Fall, 2002, 2(3): p. 83.

In Milwaukee, positive impacts of vouchers on student test scores were observed, most clearly after three and four years. Jay P. Greene, Paul E. Peterson, and Jiangtao Du, "School Choice in Milwaukee: A Randomized Experiment," in Paul E. Peterson and Bryan C. Hassel, *Learning from School Choice* (Brookings, 1998), pp. 335-56. In this randomized field trial, baseline test scores were available for only 29 percent of the voucher students and 49 percent of the control group—just 83 students after three years and 31 students after four years, making it extremely difficult to detect effects, positive or negative. As a result, the researchers placed greater weight on data from all students (300 in the third year, 112 in the fourth), whether or not baseline information was available (pp. 345-48). All results were positive, though at various levels of significance. Nonetheless wary of the problem missing benchmark scores posed, the authors pointed out that "the conclusions that can be drawn from our study are . . . restricted by limitations of the data. . . . . The percentage of missing cases is especially large when one introduces controls for . . . pre-experimental test scores. But given the consistency and magnitude of the findings . . . they suggest the desirability of further randomized experiments capable of reaching more precise estimates of efficiency gains through privatization. Randomized experiments are underway in New York City, Dayton, and Washington, D.C. If the evaluations of these randomized experiments minimize the number of missing cases and collect pre-experimental data for all subjects. . . , they could . . . provide more precise estimates of potential efficiency gains" (p. 351).

[10] In Washington, D.C., however, no statistically significant effects for African Americans were observed in year three. Howell, Peterson, Wolf, and Campbell (Brookings, 2002); William G. Howell, Patrick J. Wolf, David E. Campbell, and Paul Peterson, 2002, "School Vouchers and Academic Performance: Results from Three Randomized Field Trials." *Journal of Policy Analysis and Management*. 21(2): 191-218; Paul E. Peterson, William G. Howell, Patrick J. Wolf, and David E. Campbell, "School Vouchers: Results from Randomized Field Trials," in Caroline M. Hoxby, ed., *The Economics of School Choice*, (Chicago: University of Chicago Press); Daniel P. Mayer, 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," Program on Education Policy and Governance, Kennedy School of Government, Harvard University, 2002. Report No. 02-01.

[11] All references to Krueger and Zhu, if not otherwise identified, are to the paper released in April 2003 entitled "Another Look at the New York City School Voucher Experiment." (Department of Economics, Princeton University, 2003, unpublished paper). The reference here is to models that estimate impacts on composite test scores (see note 20 below) that do not divide the sample according to propensity score matching.

[12] Krueger and Zhu's essay focuses on a narrow band of the research reported in *The Education Gap.* Krueger and Zhu do not question the results from the parent surveys, which showed that private schools have lower levels of fighting, cheating, property destruction, absenteeism, tardiness, and racial conflict; assign more homework; establish more extensive communications with parents; contain fewer students and smaller classes; and provide fewer resources and more limited facilities. Nor do Krueger and Zhu question

certain null findings we report, namely that the voucher programs did not consistently increase parental involvement with their child's education, that they had little effect on children's self-esteem, and that they did not adversely impact the degree of racial integration in school.

[13] Most analysts are interested in obtaining the *best* estimate of programmatic impacts, not the "safest" ones, a criterion that implicitly favors the status quo. In his book *Education Matters: Selected Essays* (Elgar Publishing, 2001), Krueger clarifies his preferences: "My personal view is that policymakers should be risk-averse when it comes to changing public school systems. To alter the institutional structure of U.S. schools without sufficient evidence that the 'reforms' [elsewhere he indicates that he is referring to "vouchers, magnet schools, and charter schools"] would be successful is to put our children at risk." As quoted in Derek Neal, "Investment Planning," *Education Next*, (Winter, 2003), 3(1): p. 85.

[14] Estimates here differ slightly from those originally reported because MPR, after certifying an original set of weights and lottery indicators in 2002, revised them in 2003.

[15] Corporation for the Advancement of Policy Evaluation with Mathematica Policy Research, Inc., "Evaluation of the New York City Scholarship Program, Technical and Cost Proposal." Proposal submitted to Phoebe Cottingham, Senior Program Officer, Smith Richardson Foundation, November 24, 1997, roughly five months prior to the beginning of the collection of outcome data.

[16] A few other characteristics—mother's education, entry into grade 4, learning disabled student, gifted student, and Protestant religious affiliation—register significant correlations with test score outcomes in all three outcome years. Their correlations, however, never exceed 0.25.

[17] According to the original research proposal, MPR, the firm responsible for data collection, was to include in the lottery only those students in grades 1–4 for whom baseline test score information was available. As stated in the proposal, "The second phase of the application process will include completing a questionnaire with items that ask parents . . . to describe the basic demographic characteristics of the families. In addition, MPR will administer a standardized achievement test to students and ask students to complete a short questionnaire . . . . Children will be excluded from the lottery if they do not complete the . . . application process." Corporation for the Advancement of Policy Evaluation with Mathematic Policy Research, Inc., "Evaluation of the New York City Scholarship Program," Proposal submitted to Phoebe Cottingham, Senior Program Officer, Smith Richardson Foundation, December 11, 1996. After the lottery was held, MPR reported that administrative procedures were not fully executed according to plan, as some students for whom no baseline test scores were available nonetheless were given a chance to win a voucher.

[18] Twenty-four African American students (or 10.6 percent of the sample) in grade 1, 34 (12.9 percent) in grade 2, 21 (8.9 percent) in grade 3, and 25 (13.6 percent) in grade 4 had missing baseline test scores. All 245 African American kindergartners had missing baseline test scores.

[19] Parent surveys and tests were administered to all students in subsequent years; to do otherwise would have drawn distinctions among children and families, inviting suspicion among the participants.

[20] Krueger and Zhu report results for composite scores as well as for the math and reading portions of the test, separately. Composite scores yield more precise estimations, however; their standard errors are 15 to

20 percent lower. Given these efficiency gains, we report only impacts on composite test scores. Krueger (1999) employed this analytical strategy in the Tennessee class size study, even when precision is less of an issue, as the number of cases available for observation totaled around 10,000 students.

[21] Weighted, 2SLS regressions estimated where treatment status is used as an instrument. As covariates, models for the ATB group include private school status, baseline test scores, and lottery indicators. For the NATB group, covariates only include private school status and lottery indicators.

Estimates of private school impacts compare those students who attended a private school for three years to those students who did not. If students benefited from attending a private school for one or two years and then returned to a public school, this approach will overstate the programmatic impacts. On the other hand, if switching back and forth between public and private schools negatively impacts student achievement, then this model will underestimate the true impact of consistent private-school attendance.

When Krueger and Zhu estimate two-stage models, they assume that private school impacts accrue at a linear rate. (Nothing about the models we estimate imposes an assumption that gains must be linear.) Still, whether one estimates impacts one way or another is not particularly consequential. Our third-year estimated impact for ATB students is 8.4 NPR points; Krueger and Zhu's is 6.4 points. Although Krueger and Zhu stress that they show voucher impacts that are 31 percent less than the size of the impacts originally estimated, this appears a rather forced interpretation of the finding. Both estimates are statistically significant, and neither is significantly different from the other.

[22] Krueger and Zhu argue that programmatic effects are best understood by examining the impact of being offered a voucher rather than the impact of actually attending a private school. The first impact, known as intent-to-treat, is estimated by ordinary least squares (OLS); the second by a two-stage model (2SLS), which uses the randomized assignment to treatment and control conditions as an instrument for private school attendance. Almost all of the estimates Krueger and Zhu provide are based on the OLS model.

To ascertain the statistical significance of programmatic effects, it makes no difference which model is estimated. Both yield identical results. If, however, one is interested in the magnitude of an intervention's impact, not just its statistical significance, then the choice of models is critical. The two estimators will yield different results in direct proportion to the percentage of treatment group members who did not attend a private school and control group members who did not return to public school. If only half those offered vouchers use them, and none of the control group attends a private school, then the impact, as estimated by the OLS model, will be exactly one half that of the estimated impact of actually attending a private school. As levels of non-compliance among treatment and control group members were substantial in New York, Krueger and Zhu's OLS estimates are considerably lower than the 2SLS estimates we report above.

Krueger and Zhu provide two justifications for focusing on the effect of a voucher offer. First, they claim that the OLS estimates provide a "cleaner interpretation" of the efficacy of school vouchers. We disagree. It is not at all clear why the act of offering a voucher—as distinct from the act of using a voucher to attend a private school for one, two, or three years—should affect student achievement. Presumably, differences between treatment and control groups derive from the differential attendance patterns at public

and private schools, not from the mere fact that only one group was offered vouchers. Results that isolate the impact of attending a private school provide the "cleaner interpretation" of programmatic impacts.

Krueger and Zhu also argue that the OLS model provides the better estimation of the "societal" effects of school vouchers. Presumably, the effect of an offer establishes some baseline for assessing the average gains that one can expect from a voucher intervention. This claim, however, assumes that voucher usage rates are unrelated to programmatic issues of scale, publicity, and durability. Since the New York voucher program was small, privately funded, initially limited to three years, and given only modest attention by the news media, one must make strong assumptions to infer that the voucher offer provides an accurate estimate of impacts in larger-scale programs.

Other scholars also prefer 2SLS to OLS models when estimating the impacts of an intervention. For example, Alan Gerber and Don Greene employ the two-stage model when reporting results from a randomized experiment designed to ascertain the effects of door-to-door campaigning on voter turnout in New Haven. Alan S. Gerber and Donald P. Green, "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment," *American Political Science Review,* Vol. 94, (September 2000), pp. 653-63. Using the two-stage model, they observed that personal contacts with voters increases turnout by 9 percentage points, on average. Had they followed Krueger and Zhu's recommendation to privilege the OLS estimate, they would have found only a 3 percentage point impact, a finding that would have underestimated the actual impact of door-to-door contacts—for the simple reason that many families assigned to treatment in their experiment were not contacted. In his class-size research, Krueger reports without apology 2SLS estimates of attending small classes (1999).

[23] ATB reading baseline test scores were 25.4 (st. dev.=22.7) for the control group, 23.3 (st. dev.=22.5) for the treatment group. Math scores were 15.4 (st. dev.=18.2) and 15.8 (st. dev.=18.7), respectively. Nor, as a result of attrition, did the ATB group become unbalanced later on. Average composite baseline test scores were 19.3, 19.9, and 20.4 NPR points among African American students who attended the follow-up sessions in years one, two, and three, respectively; among the control group, baseline scores were 20.0, 20.4, and 21.1 NPR points for the three respective years. None of the differences are statistically significant. Point estimates barely change when baseline test scores are included in models estimating the effects of private school attendance (see Table 2).

[24] In his study of class size effects, Krueger argues that randomization worked as planned, yielding balanced treatment and control groups (1999). Unfortunately, no baseline data on student test scores were available in the Tennessee Star Study. Pointing to this fact, other scholars have identified risks that the randomization process was compromised, and, as a consequence, that the reported estimations of class size effects may have been biased. See, for example, Eric A. Hanushek, "The Evidence on Class-Size," in Susan E. Mayer and Paul E. Peterson, eds., *Earning and Learning: How Schools Matter* (Brookings, 1999), pp. 153-61; Eric A. Hanushek, "Evidence, Politics and the Class Size Debate," (Hoover Institution, Stanford University, 2000); Derek Neal, "Review of Alan Krueger's 'Education Matters: Selected Essays,'" *Education Next*, Vol. 2, (Winter, 2003), p. 85.

[25] The percentages are for missing information on one of the 16 demographic variables that Krueger and Zhu introduce (see below).

[26] Mismatches, however, may result from more than just administrative error. Some NATB parents may have brought to follow-up testing sessions children different from those who participated in the initial lotteries. Given that older NATBs apparently refused to take tests at baseline, they may well have resisted attending testing sessions in subsequent years. Families in the control group and decliners in the treatment group, nonetheless, had financial incentives to attend these follow-up testing sessions—families were awarded between $50 and $100 for their continued participation in the study. Because parental surveys provided the only information available to verify the identity of these children, however, parents could have brought a child other than their own. If enough parents in the control group brought a better performing student in their child's place, this by itself could account for negative private-school impacts observed among NATBs. The problem is much less acute for ATBs, who identify themselves on both the baseline and follow-up tests.

[27] William G. Howell and Paul E. Peterson, "The Use of Theory in Randomized Field Trials: Lessons from School Voucher Research on Disaggregation, Missing Data, and the Generalizability of Findings," *The American Behavioral Scientist* (forthcoming).

[28] The possibility that voucher effects varied by grade level has been the subject of a good deal of commentary in *New York Times* coverage of our research. Reporters and columnists have conveyed the impression that impacts for African Americans varied significantly by grade level, sometimes quoting MPR researcher David Myers to this effect. Kate Zernike, "New Doubt Is Cast on Study that Backs Voucher Effects," *New York Times*, September 15, 2000; Richard Rothstein, "Judging Voucher Merits Proves a Difficult Task," *New York Times*, December 13, 2000; Michael Winerip, "What Some Much-Noted Data Really Showed about Vouchers," *New York Times*, May 7, 2003.

Despite these news reports, David Myers has never identified significantly different impacts from one grade level to another in either year two or year three. Zernike quoted Myers at the end of the second year of the study as saying that positive effects were "concentrated" in a particular grade. That information, however, is not to be found in the scientific report issued at the end of the second year, which reveals no statistically significant differences in the effects by grade level (David Myers, Paul E. Peterson, Daniel Mayer, Julia Chou, and William G. Howell, "School Choice in New York City after Two Years: An Evaluation of the School Choice Scholarships Program," Program on Education Policy and Governance, Harvard University, Report 00-17. Available at www.ksg.harvard.edu/pepg/). In the final report issued by MPR, Myers and his co-authors specifically reported no significant differences by grade level, writing the following: "When the impact of attending private school for three years on African American student test scores was examined by grade level, we observed no statistically significant differences in the impact between grade levels (See Appendix D.) The impact for students in the younger grouping was 8.5 percentile points, and in the older grades the average impact was 9.1 points. Both impacts were statistically significant." (Mayer, Peterson, Myers, Tuttle, and Howell, 2002, p. 38) In an April 1, 2003, memorandum, "Comments on 'Another Look at the New York City Voucher Experiment'" (Mathematica Policy Research, Washington, D. C.), Myers modifies his position, saying "the offer of a voucher had a small positive impact on the achievement of African American students no matter which of the black definitions . . . are used; however, the impacts are concentrated among the oldest students (the grade 4 cohort)." But Myers and Daniel Mayer, in this memorandum, once again fail to show statistically significant differences between the grade 4 cohort and cohorts 1-3. Using MPR's revised weights, estimated private-school impacts after three years are 7.4, 3.4, 7.5, and 10.9 NPR points for African Americans in grades 1, 2, 3, and 4, respectively.

None of these estimates differs significantly from the others. When grades 1–2 and 3–4 are combined, the estimates are 7.8 and 8.0 NPR points—both statistically significant at $p<.05$. In total, 127, 156, 130, and 106 African American ATBs are included in the year three test score models for grades 1, 2, 3, and 4, respectively.

The results do not change when the African American NATBs are added to the analysis. The year three impacts from hybrid models are 6.0, 4.8, 4.0, and 11.8 NPR points for grades 1, 2, 3, and 4, respectively. None of these impacts is significantly different from the others. Once again, impacts in grades 1–2 and grades 3–4 combined are 7.9 and 7.1 NPR points—both significant at $p< 0.05$. In total, 139, 177, 139, and 122 African American ATBs and NATBs are included in the year-three test score models for each of the four respective grades.

Disagreeing with Myers, Krueger and Zhu conclude that grade differences are minimal. As they put it, "The grade at which students are offered vouchers is unrelated to the magnitude of the treatment effect in the third year of the experiment . . . although there we find some tendency for older students to have a larger treatment effect when Kindergarten students are included." Indeed, impacts for kindergartners are negative in all three years: -0.7, -2.1, and -13.9 NPR points, respectively. By contrast, impacts for all students in the other grades, regardless of whether baseline scores are available, are significantly positive: 5.7, 4.2, and 7.5 NPR points. Interaction terms between kindergartners and treatment in test-score models come up significant in years one and three. These differences raise the question as to whether the kindergartners are genuinely different from the other cohorts or whether the data on kindergartners are invalid (see discussion in text for ways in which bias may have been introduced).

[29] Controlling for baseline test scores will not bias the estimated treatment effects as long as they are unrelated to students' assignments to treatment and control groups. As previously indicated, after years one, two, and three, the balance of baseline test scores between the treatment and control groups appears intact (see footnote 23).

[30] Among African American NATBs and ATBs in grades 1–4, 695, 562, and 577 observations are available for years one, two, and three, respectively; for grades K–4, 882, 722, and 734 observations are available for the three respective years. In addition to controlling for baseline test scores when possible, hybrid models include missing data indicators, private school status, and lottery indicators.

[31] Inasmuch as demographic information from a parent survey is more reliable than such information collected from young children, the parent, not the student, was the source of this information.

[32] Information on the father's ethnic background was collected only at baseline.

[33] Although one parent inadvertently marked the "other" category, then wrote in "African American," no outcome test scores were available for the children.

[34] "Students . . . were added . . . [to the African American category] 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" (Krueger and Zhu 2003, p. 27, note 25).

[35] Barry Edmonston, Joshua Goldstein, and Juanita Tamayo Lott, eds., *Spotlight on Heterogeneity: The Federal Standards for Racial and Ethnic Classification, Summary of a Workshop* (National Academy Press, 1996). Appendix B: Office of Management and Budget: Statistical Directive No. 15. The Directive also calls for the listing of two other categories: "American Indian or Alaskan Native" and "Asian or Pacific Islander." Krueger and Zhu admonish Mathematica Policy Research for not using a data collection procedure recommended in this Directive, despite the fact that MPR's classification scheme is consistent with one of the options it provides. The admonishment is especially surprising, given that Krueger and Zhu themselves have chosen a classification scheme that fails to conform with those recommended in this very Directive.

[36] National Press Club, Washington, D.C., April 1, 2003.

[37] In *An American Dilemma* (McGraw Hill, 1964), Gunnar Myrdal pointed out that the African American experience, rooted in a history of slavery and intense segregation, is unique in American society. Ethnic classifications based strictly on physical appearances ignore African Americans' distinctive history, culture, and social networks. In *The Education Gap*, for instance, we show that Hispanics, like other immigrant groups, appear to have more educational choice and suffer less from certain kinds of discrimination than African Americans.

[38] Krueger and Zhu 2002.

[39] Krueger and Zhu also look at the public schools from which Hispanic and African American students came, reporting that impacts on African Americans and Hispanics coming from the same public schools differ markedly from one another. Using a system of weights for which insufficient information is available for replication to be possible, they report voucher impacts for African Americans that are a statistically significant 5 NPR points; for Hispanics an insignificant -3 NPR points.

The reported results are quite consistent with our original findings about the differential impacts of the voucher intervention on African Americans and Hispanics. Krueger and Zhu, however, use what might seem as confirmation as evidence to the contrary. More exactly, they insist that we are wrong to argue that gains observed for African Americans are due to inequities within the public sector. As they put it, "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."

In making this suggestion, Krueger and Zhu assume that public schools have uniform impacts on all students within them. A bad school is equally bad for African Americans and Hispanics. But schools can be good for one student without being good for another. Indeed, that is one of the central objectives of school voucher initiatives: by expanding educational options, families are able to search for schools that address the particular needs and interests of their individual child.

Schools are not necessarily poor or excellent, in fixed or absolute terms. Indeed, if teachers at schools with overlapping populations treat non-Hispanic African Americans differently from others; if they communicate differently with the mothers of African American students with the mothers of other students; if the expectations for those from African American households are different from the expectations from

other households, then the quality of public schools is not accurately ascertained when estimating test score impacts for students from different ethnic backgrounds attending overlapping schools.

[40] Krueger and Zhu point out that the Chinese classify people by their fathers' ethnicity. Although that procedure seems inappropriate for the population under consideration in this study, it is at least a consistent coding principle, while Krueger and Zhu's is not.

[41] See, for example, Meredith Phillips, Jeanne Brooks-Gunn, Greg J. Duncan, Pamela Klebanov, Jonathan Crane, "Family Background, Parenting Practices, and the Black-White Test Score Gap," in Christopher Jencks and Meredith Phillips, eds., *The Black-White Test Score Gap* (Brookings, 1998), pp.103-48.

[42] Results are similar when ATB and NATB students are considered together.

[43] A methodological consideration provides a further basis for classifying according to mother's ethnicity. Because mothers usually completed parent surveys, information about them is more likely to be valid than information about fathers, most of whom were not living with either mother or child. Indeed, demographic information is missing for 76 percent of the fathers, as compared to only 21 percent of the mothers. Krueger and Zhu themselves recognize the importance of mothers, noting three exceptional cases where "there is no indication the mother lived at home" and yet the child was classified according to the mother's background. When estimating impacts for parental caretakers, these three cases are included in the estimates (see Table 3).

[44] Though the other two classifications are plausible, to avoid any semblance of classification searching, we place primary weight on the original classification scheme, devised prior to the conduct of the study. Differences in results from the three plausible classification schemes are trivial.

[45] Eighty students had an African American father and a mother from a different ethnic background; 78 students had an African American mother and a father from a different ethnic background.

[46] There are no missing cases for the four grade level indicators.

[47] In all, 32 percent of observations had at least one missing value on the additional covariates Krueger and Zhu introduce to the analysis. For a simple overview of the statistical problems associated with estimating regression models with missing data, see Paul Allison, *Missing Data* (Thousand Oaks: Sage Publications, 2002). For the problems of dummy variable adjustments for missing data, see M.P. Jones, "Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression." *Journal of the American Statistical Association*. 91:222-230.

[48] Christopher Achen, *The Statistical Analysis of Quasi-Experiments*. (Berkeley: University of California Press, 1986), p. 27. For the ATBs, such concerns are alleviated because we know that treatment and control groups are balanced.

[49] Only baseline test scores were mentioned, *a priori*, as a necessary benchmark when estimating achievement effects. See discussion above.

[5050] Krueger and Zhu's paper presented at the original conference on randomized experiments included nine background controls: four indicator variables for student grade level, mother's education, log of family income, mother's employment, and gender. (Alan B. Krueger and Pei Zhu, "Another Look at the New York City School Voucher Experiment." A paper prepared for the Conference on Randomized Experimentation in the Social Sciences, Yale University, August 16, 2002). In the version presented at the National Press Club, Krueger and Zhu dropped marital status while adding controls for gifted, special education, mother born US, English-speaking household, student's age, residential mobility, mother Catholic, and welfare. With the exception of grade cohorts, none of these variables were included in Krueger's original project proposal. See Alan Krueger, "Data License and Confidentiality Agreement Reanalysis of the Data Used in "School Choice in New York City After Two Years: An Evaluation of the School Choice Scholarships Program." Project proposal submitted to Joanne Pfleiderer, Director of Communications, Mathematica Policy Research, May 9, 2001.

[51] Notice also that notable efficiency gains are realized simply by adding baseline test scores to the models. Standard errors drop by between 0.7 and 0.8 NPR points when baseline test scores are added to the models in all four of the ways used to classify ATB students as African American (rows 1 and 2). But no more than the most trivial gains in efficiency are realized by adding additional covariates. Even when all 28 additional covariates are added to the model, reductions in standard errors are all less than 0.1 NPR points.

[52] The primary effect of adding covariates, instead, is to depress the point estimates on private school attendance, which drop between 1.1 and 1.5 NPR points by the time all are added to the model—a revelation that substantiates Krueger and Zhu's point that additional covariates may artificially "sway the estimated treatment effect," just as it reinforces concerns about specification searching.

Table 1: Summary of Estimated Test Score Impacts for African Americans, Variously Defined

		Positive, Significant Test-Score Impacts Observed in:		
		Year One	Year Two	Year Three
	I. Simple, Transparent Models			
	A. Mother Is African American			
1-3	All students for whom baseline test scores are available, controlling for baseline scores	✓	✓	✓
4-6	All students for whom baseline test scores are available, not controlling for baseline scores (imprecise estimation)	✓	✓	✓
7-9	All students in grades 1-4, controlling for baseline scores when possible	✓	✓	✓
10-12	All students in grades K-4, controlling for baseline scores when possible	✓	✓	✓
	B. Both Mother and Father Are African American			
13-15	All students for whom baseline test scores are available, controlling for baseline scores	✓	✓	✓
16-18	All students for whom baseline test scores are available, not controlling for baseline scores (imprecise estimation)	✓	✓	✓
19-21	All students in grades 1-4, controlling for baseline scores when possible	✓	✓	✓
22-24	All students in grades K-4, controlling for baseline scores when possible	✓	✓	✓
	C. Parental Caretaker Is African American			
25-27	All students for whom baseline test scores are available, controlling for baseline scores	✓	✓	✓
28-30	All students for whom baseline test scores are available, not controlling for baseline scores (imprecise estimation)	✓	✓	✓
31-33	All students in grades 1-4, controlling for baseline scores when possible	✓	✓	✓
34-36	All students in grades K-4, controlling for baseline scores when possible	✓	✓	✓
	D. Either Mother or Father Is African American (inconsistent classification scheme)			
37-39	All students for whom baseline test scores are available, controlling for baseline scores	✓	✓	✓
40-42	All students for whom baseline test scores are available, not controlling for baseline scores (imprecise estimation)	✓	✓	✓
43-45	All students in grades 1-4, controlling for baseline scores when possible	✓	✓	✓
46-48	All students in grades K-4, controlling for baseline scores when possible	✓		✓
	II. Models that Include 12 Additional Covariates: Results from the Krueger/Zhu Initial Specification			
	A. Mother Is African American			
49-51	All students for whom baseline test scores are available, controlling for baseline scores	✓	✓	✓
52-54	All students in grades 1-4, controlling for baseline scores when possible	✓	✓	✓
55-57	All students in grades K-4, controlling for baseline scores when possible	✓	✓	✓
	B. Both Mother and Father Are African American			
58-60	All students for whom baseline test scores are available, controlling for baseline scores	✓	✓	✓
61-63	All students in grades 1-4, controlling for baseline scores when possible	✓	✓	✓
64-66	All students in grades K-4, controlling for baseline scores when possible	✓	✓	✓

Table 1 Continued

		Positive, Significant Test-Score Impacts Observed in:		
		Year One	Year Two	Year Three
	C. Parental Caretaker Is African American			
67-69	All students for whom baseline test scores are available, controlling for baseline scores	✓	✓	✓
70-72	All students in grades 1-4, controlling for baseline scores when possible	✓	✓	✓
73-75	All students in grades K-4, controlling for baseline scores when possible	✓	✓	✓
	D. Either Mother or Father Is African American (inconsistent classification scheme)			
76-78	All students for whom baseline test scores are available, controlling for baseline scores	✓	✓	✓
79-81	All students in grades 1-4, controlling for baseline scores when possible	✓		✓
82-84	All students in grades K-4, controlling for baseline scores when possible	✓		
	III. Models that Include 28 Additional Covariates: Results from the Krueger/Zhu Second Specification			
	A. Mother Is African American			
85-87	All students for whom baseline test scores are available, controlling for baseline scores	✓	✓	✓
88-90	All students in grades 1-4, controlling for baseline scores when possible	✓	✓	✓
91-93	All students in grades K-4, controlling for baseline scores when possible	✓		
	B. Both Mother and Father Are African American			
94-96	All students for whom baseline test scores are available, controlling for baseline scores	✓	✓	✓
97-99	All students in grades 1-4, controlling for baseline scores when possible	✓	✓	✓
100-102	All students in grades K-4, controlling for baseline scores when possible	✓	✓	
	C. Parental Caretaker Is African American			
103-105	All students for whom baseline test scores are available, controlling for baseline scores	✓	✓	✓
106-108	All students in grades 1-4, controlling for baseline scores when possible	✓	✓	✓
109-111	All students in grades K-4, controlling for baseline scores when possible	✓		
	D. Either Mother or Father Is African American (inconsistent classification scheme)			
112-114	All students for whom baseline test scores are available, controlling for baseline scores	✓	✓	✓
115-117	All students in grades 1-4, controlling for baseline scores when possible	✓		✓
118-120	All students in grades K-4, controlling for baseline scores when possible	✓		

Effects deemed significant at $p<.10$, two-tailed test.

Table 2: Private-School Impacts on African American Test Scores: Alternative Estimates and Efficiency Losses Resulting from Exclusion of Baseline Test Scores for Various Groups of Students

	Year One		Year Two		Year Three	
Baseline Scores in Model						
1. Students with baseline scores (ATBs, grades 1–4)	6.13***	(1.69)	4.16**	(1.87)	8.43***	(2.46)
No Baseline Scores in Model						
2. Students with baseline scores (ATBs, grades 1–4)	5.67***	(2.31)	4.36**	(2.37)	8.40***	(3.19)
3. Students without baseline scores (NATBs, grades K–4):	0.09	(4.50)	-3.48	(5.27)	-13.25	(8.39)
4. Students with and without baseline scores (ATBs & NATBs, grades K–4)	4.61 **	(2.03)	3.24	(2.16)	4.88	(2.97)
Hybrid Model: Baseline Scores when Possible						
5. Students with and without baseline scores (ATBs & NATBs, grades 1–4)	6.28 ***	(1.75)	3.94 **	(1.91)	5.75***	(2.52)
6. Students with and without baseline scores (ATBs & NATBs, grades K–4)	5.15***	(1.67)	3.21*	(1.86)	5.31**	(2.53)

Impacts of private school attendance on test scores reported. Weighted, two-stage least squares regressions estimated; treatment status used as instrument. Standard errors reported in parentheses. *** significant at p<.01, two-tailed test; ** significant at p<.05; * at p<.10. ATBs consist of students for whom baseline test scores are available; NATBs consist of students for whom no baseline test scores are available. First set of models include as covariates private school status, baseline test scores, and lottery indicators; the second set include only private school status and lottery indicators; the hybrid model includes private school status, baseline test scores (interacted with a dummy variable for students with baseline test scores), the dummy variable for students with baseline test scores, and lottery indicators.

Significance tests in all tables are based upon OLS standard errors rather than the less efficient bootstrapped standard errors, based either on observations or residuals. In *The Education Gap*, we report OLS standard errors for all studies. The argument for bootstrapping rests upon the assumption of correlated observations, correlation that persists even after appropriate covariates are included in the model. Those who would bootstrap either observations or residuals point out that there may be dependencies of scores among family members; in our view, this is much less of a concern when one is estimating changes in scores (as is being done here) rather than estimating simple test score levels. In his study on class size reductions, Krueger similarly finds no need to use bootstraps to account for intra-family correlations, even though no baseline test scores are available for students there (1999).

Table 3: Test Score Impacts for African Americans, Variously Defined

	Mother African American	Mother & Father African American	Parental Caretaker African American[1]	Mother or Father African American
	(1)	(2)	(3)	(3)
Students with Baseline Scores (ATBs)				
Year One	6.13***	5.78***	6.18***	5.29***
	(1.69)	(1.76)	(1.69)	(1.73)
Year Two	4.16**	4.13**	4.17**	3.28*
	(1.88)	(1.92)	(1.88)	(1.90)
Year Three	8.42***	8.05***	8.36***	7.64***
	(2.46)	(2.59)	(2.45)	(2.40)
Students with and without Baseline Scores (ATBs & NATBs)				
Year One	5.15***	5.00***	5.20***	4.00***
	(1.67)	(1.73)	(1.67)	(1.69)
Year Two	3.21*	3.56*	3.24*	2.66
	(1.86)	(1.86)	(1.86)	(1.87)
Year Three	5.31**	5.08*	5.27**	4.45*
	(2.53)	(2.60)	(2.53)	(2.47)

Weighted, two-stage least squares regressions estimated; treatment status used as instrument. Standard errors reported in parentheses. * significant at .10 level, two tailed test conducted; ** significant at .05 level; *** at .01 level. Models for students with baseline test scores control for baseline scores and lottery indicators; models for all students control for test scores when possible, an indicator variable for missing baseline scores, and lottery indicators. Mother's ethnicity determined on the basis of baseline, year two, and year three surveys; father's ethnicity determined on the basis of baseline surveys only. When accounting for the ethnicity of both parents, if missing, mother [father] assumed African American when father [mother] African American. ATBs consist of students for whom baseline test scores are available; NATBs consist of students for whom no baseline test scores are available.

[1] Mother assumed to be the primary caretaker of the child's education except in those cases where the child lives only with the father.

Table 4: Year Three Test Score Impacts for African Americans, Variously Defined, With and Without Baseline Test Scores (Estimates Obtained from Simple/Transparent Models and from Specification Searches)

	Student with Baseline Test Scores (ATBs)				Students with and without Baseline Test Scores (ATBs & NATBs)			
	Mother African American	Both Mother & Father AA	Parental Caretaker African American[1]	Either Mother or Father AA	Mother African American	Both Mother & Father AA	Parental Caretaker African American[1]	Either Mother or Father AA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Transparent Model (no baseline test scores)	8.40*** (3.19)	7.91** (3.36)	8.41*** (3.19)	7.10** (3.14)	4.88 (2.97)	4.69 (3.06)	4.90* (2.97)	3.53 (2.92)
Transparent Model (with baseline test scores)[2]	8.42*** (2.46)	8.05*** (2.59)	8.36*** (2.45)	7.64*** (2.40)	5.31** (2.53)	5.08* (2.60)	5.27** (2.53)	4.45* (2.47)
1st Search, Controls for:								
four grade levels[3]	7.88*** (2.41)	7.18*** (2.52)	7.84*** (2.40)	7.43*** (2.36)	4.79* (2.48)	4.31* (2.55)	4.79* (2.48)	4.23* (2.43)
Plus mother's education	7.80*** (2.40)	7.15*** (2.52)	7.77*** (2.40)	7.32*** (2.35)	4.82* (2.48)	4.39* (2.55)	4.82* (2.48)	4.19* (2.43)
Plus log income	7.79*** (2.40)	7.17*** (2.51)	7.76*** (2.40)	7.35*** (2.35)	4.82* (2.50)	4.38* (2.57)	4.82* (2.49)	4.18* (2.44)
Plus student's gender	7.74*** (2.42)	7.16*** (2.53)	7.71*** (2.41)	7.36*** (2.36)	4.81* (2.53)	4.46* (2.60)	4.80* (2.53)	4.15* (2.46)
Plus employment status	7.85*** (2.43)	7.23*** (2.55)	7.81*** (2.43)	7.79*** (2.39)	4.40* (2.52)	4.44* (2.60)	4.40* (2.52)	3.88 (2.47)
2nd Search, Adds Controls:								
welfare	7.95*** (2.43)	7.36*** (2.54)	7.91*** (2.43)	7.87*** (2.39)	4.40* (2.53)	4.52* (2.61)	4.39* (2.53)	3.84 (2.47)
Plus mother born US	7.80*** (2.42)	7.11*** (2.53)	7.76*** (2.42)	7.74*** (2.38)	4.20* (2.53)	4.26 (2.61)	4.19* (2.52)	3.51 (2.47)
Plus residential mobility	7.98*** (2.39)	7.07*** (2.49)	7.94*** (2.39)	7.70*** (2.34)	4.32* (2.52)	4.14 (2.60)	4.31* (2.52)	3.55 (2.46)
Plus English spoken home	7.50*** (2.39)	6.61*** (2.48)	7.46*** (2.38)	7.23*** (2.34)	3.96 (2.51)	3.81 (2.59)	3.95 (2.51)	3.19 (2.46)
Plus mother Catholic	7.23*** (2.38)	6.40*** (2.48)	7.19*** (2.38)	7.04*** (2.34)	3.82 (2.50)	3.80 (2.58)	3.81 (2.50)	3.11 (2.45)
Plus student's age	7.19*** (2.38)	6.37** (2.48)	7.15*** (2.38)	7.03*** (2.35)	3.83 (2.51)	3.79 (2.59)	3.81 (2.50)	3.11 (2.46)
Plus student gifted	7.10*** (2.39)	6.28** (2.49)	7.06*** (2.39)	6.96*** (2.34)	3.62 (2.50)	3.61 (2.59)	3.60 (2.50)	3.06 (2.43)
Plus student special ed.	7.20*** (2.40)	6.39** (2.50)	7.16*** (2.40)	6.90*** (2.34)	3.55 (2.50)	3.66 (2.58)	3.53 (2.49)	2.91 (2.42)

Weighted, two-stage least squares regressions estimated; treatment status used as instrument. Standard errors reported in parentheses. * significant at .10 level, two tailed test conducted; ** significant at .05 level; *** at .01 level. Mother's ethnicity determined on the basis of baseline, year two, and year three surveys; father's ethnicity determined on the basis of baseline surveys only. When accounting for the ethnicity of both parents, if missing, mother [father] assumed African American when father [mother] African American. All models include as covariates private school status and revised lottery indicators. Covariates then added cumulatively, so that final row includes test scores and all 16 additional demographic controls and all 12 missing value indicators that are used in the Krueger/Zhu estimations. Among African American mothers, 6.5 percent of cases are missing for mother's education, 7.6 percent for income, 3.3 for gender, 2.6 for employment status, 10.9 for welfare, 2.1 for born U.S., 2.9 for residential mobility, 3.0 for English spoken at home, 7.7 for Catholic, 4.4 for age, 3.5 for gifted, and 3.9 for special education. The first five rows of additional controls are those covariates included in the Krueger/Zhu conference paper (2002). The last 8 rows are those covariates included in the version released at the National Press Club (2003). (The conference paper also included marital status, which subsequently was dropped from later analyses.)

[1] Mother assumed to be the primary caretaker of the child's education except in those cases where the child lives only with the father.

[2] Models for students with baseline test scores include as covariates private school status, baseline scores and lottery indicators; models for all students include as covariates private school status, baseline test scores when possible, an indicator variable for missing baseline scores, and lottery indicators.

[3] Three indicator variables are included for models that only include ATBs.