Efficiency, Bias, and Classification Schemes

A Response to Alan B. Krueger and Pei Zhu

PAUL E. PETERSON
WILLIAM G. HOWELL
Harvard University

When estimating voucher impacts on test scores in the New York City randomized field trial (RFT) for African Americans (defined either by mother's ethnicity, parental caretaker, mother and father's ethnicity, or mother or father's ethnicity), results remain significantly positive, even when models include students for whom no baseline test scores are available. These results obtain as long as one estimates impacts precisely by controlling for baseline test scores for those students who have them. Positive impacts fall below conventional levels of significance only when analysts needlessly drop baseline test score information or add numerous covariates that neither singly nor together enhance the precision of the estimates. When results differ for those with and without baseline scores, analysts should give greater weight to those for whom one has stronger evidence that the RFT has not been contaminated.

Due to space limitations imposed by the editors, this article replies to the original Krueger and Zhu (KZ) article but not to their rejoinder. Our reply to the rejoinder may be found at www.ksg.harvard.edu/pepg/. There, we discuss the trivial nature of the differences between findings reported below and those in KZ's replication, our offer to check the replication effort to see how differences could have arisen, the Office of Management and Budget (OMB) definition of "Black (non-Hispanic)," and other issues raised in their rejoinder.

Keywords: *randomized field trials; school vouchers; bias*

The Education Gap: Vouchers and Urban Schools (Howell & Peterson, 2002) reports that private school did not have any discernible impact, positive or negative, on the test scores of non–African Americans in New York City or students taken as a whole.[1] But for African Americans, significant, positive impacts were observed in all 3 years of the experiment. These findings, furthermore, are robust to numerous alternative specifications and classification schemes. As summarized in Table 1, 108 of 144 different statistical models yield positive and significant effects using a two-tailed test; another 29 are significant using a one-tailed test; and the remaining 7 are also positive but fall short of conventional levels of statistical significance.[2]

AMERICAN BEHAVIORAL SCIENTIST, Vol. 47 No. 5, January 2004 699-717
DOI: 10.1177/0002764203260158

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

		Positive, Significant Test-Score Impacts Observed in		
		Year 1	*Year 2*	*Year 3*
	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	**	†	*
13-15	All students in Grades 1-4, not controlling for baseline scores when possible (imprecise estimation)	**	*	**
16-18	All students in Grades K-4, not controlling for baseline scores when possible (imprecise estimation)	**	†	†
	B. Both mother and father are African American			
19-21	All students for whom baseline test scores are available, controlling for baseline scores	**	*	**
22-24	All students for whom baseline test scores are available, not controlling for baseline scores (imprecise estimation)	**	*	**
25-27	All students in Grades 1-4, controlling for baseline scores when possible	**	**	**
28-30	All students in Grades K-4, controlling for baseline scores when possible	**	*	*
31-33	All students in Grades 1-4, not controlling for baseline scores when possible (imprecise estimation)	**	**	**
34-36	All students in Grades K-4, not controlling for baseline scores when possible (imprecise estimation)	**	†	†
	C. Parental caretaker is African American			
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	**	†	*
49-51	All students in Grades 1-4, not controlling for baseline scores when possible (imprecise estimation)	**	*	**
52-54	All students in Grades K-4, not controlling for baseline scores when possible (imprecise estimation)	**	†	†
	D. Either mother or father is African American (inconsistent classification scheme)			
55-57	All students for whom baseline test scores are available, controlling for baseline scores	**	†	**
58-60	All students for whom baseline test scores are available, not controlling for baseline scores (imprecise estimation)	*	—	**
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	**	†	*
67-69	All students in Grades 1-4, not controlling for baseline scores when possible (imprecise estimation)	*	—	*
70-72	All students in Grades K-4, not controlling for baseline scores when possible (imprecise estimation)	†	—	—

	II. Models that include 12 additional covariates: Results from the initial KZ specification			
	A. Mother is African American			
73-75	All students for whom baseline test scores are available, controlling for baseline scores	**	*	**
76-78	All students in Grades 1-4, controlling for baseline scores when possible	**	*	**
79-81	All students in Grades K-4, controlling for baseline scores when possible	**	†	†
	B. Both mother and father are African American			
82-84	All students for whom baseline test scores are available, controlling for baseline scores	**	*	**
85-87	All students in Grades 1-4, controlling for baseline scores when possible	**	**	**
88-90	All students in Grades K-4, controlling for baseline scores when possible	**	*	†
	C. Parental caretaker is African American			
91-93	All students for whom baseline test scores are available, controlling for baseline scores	**	**	**
94-96	All students in Grades 1-4, controlling for baseline scores when possible	**	*	**
97-99	All students in Grades K-4, controlling for baseline scores when possible	**	†	†
	D. Either mother or father is African American (inconsistent classification scheme)			
100-102	All students for whom baseline test scores are available, controlling for baseline scores	**	*	**
103-105	All students in Grades 1-4, controlling for baseline scores when possible	**	†	**
106-108	All students in Grades K-4, controlling for baseline scores when possible	**	†	†
	III. Models that include 28 additional covariates: Results from the second KZ specification			
	A. Mother is African American			
109-111	All students for whom baseline test scores are available, controlling for baseline scores	**	*	**
112-114	All students in Grades 1-4, controlling for baseline scores when possible	**	†	**
115-117	All students in Grades K-4, controlling for baseline scores when possible	**	†	†
	B. Both mother and father are African American			
118-120	All students for whom baseline test scores are available, controlling for baseline scores	**	*	**
121-123	All students in Grades 1-4, controlling for baseline scores when possible	**	**	**
124-126	All students in Grades K-4, controlling for baseline scores when possible	**	†	†
	C. Parental caretaker is African American			
127-129	All students for whom baseline test scores are available, controlling for baseline scores	**	*	**
130-132	All students in Grades 1-4, controlling for baseline scores when possible	**	†	**
133-135	All students in Grades K-4, controlling for baseline scores when possible	**	†	†
	D. Either mother or father is African American (inconsistent classification scheme)			
136-138	All students for whom baseline test scores are available, controlling for baseline scores	**	†	**
139-141	All students in grades 1-4, controlling for baseline scores when possible	**	—	*
142-144	All students in Grades K-4, controlling for baseline scores when possible	**	—	—

NOTE: Significance tests are based on bootstrap standard errors (10,000 reps completed) that are robust to intrafamily correlations.
†$p < .10$, one-tailed. *$p < .10$, two-tailed. **$p < .05$, two-tailed.

The experimental findings from New York are consistent with those of prior studies using observational data.[3] They are also consistent with results from experiments in Washington, DC, and Dayton, Ohio, which found no impacts for White students but, in the 2nd year, positive impacts for African Americans.[4] In the first secondary analysis of the New York experimental data, Barnard, Frangakis, Hill, and Rubin (2003a) also found, after 1 year, "positive effects on math scores for children who applied to the program from . . . schools with average test scores below the citywide median. Among these children, the effects are stronger . . . for African American children" (p. 299).[5] And, in the tables of the later secondary analysis by Alan Krueger and Pei Zhu (2004 [this issue], hereafter KZ), 30 of 51 of the estimations of the voucher impacts on the overall (composite) test scores of African Americans yield significantly positive findings.[6]

Despite the weight of evidence available from the extant literature and from their own estimations, KZ express strong doubts that African Americans benefited from the New York City voucher intervention. At one point in their article, 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" (p. 668). In the end, however, KZ back away from this statement, asserting only that "the safest conclusion is probably that the provision of vouchers did not lower the scores of African American students" (p. 695)—or, equivalently, that African American students who used vouchers to attend private schools performed as well or better than their peers in public school.[7]

How do KZ generate findings that justify their conclusion? Three analytical decisions stand out as most consequential: (a) include students without baseline scores in the analysis, despite the risk of obtaining a biased estimate of the program's effects; (b) employ an unusual, questionable ethnic classification scheme; and (c) 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 KZ claim to be the "most important" evidence in support of null findings—does not, by itself, provide a basis for their conclusions. Results remain significantly positive for African American students in all 3 outcome years when these students are added to the study. In addition, results do not change materially if one takes a second step on which KZ 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 the sample, estimated voucher effects for African American test scores remain significantly positive.

If these methodological innovations do not, by themselves, significantly alter the results, they are nonetheless problematic for reasons discussed below. For these reasons, the evidence continues to support our original conclusions that African Americans, and only African Americans, posted significantly positive

test score gains from attending private schools that in Year 3 ranged from one quarter to two fifths of a standard deviation, depending on the estimation (Howell, Wolf, Campbell, & Peterson, 2002; Peterson, Howell, Wolf, & Campbell, 2003).[8]

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 are test scores. 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 1, 2, and 3, respectively. None of the correlations logged by demographic variables are even half as large.[9]

Unfortunately, Mathematica Policy Research (MPR), the firm that administered the evaluation, was not able to obtain test score data for everyone at baseline. Some students in Grades 1-4 were sick, others refused to take the test, and some tests were lost in the administrative process.[10] In addition, due to the substantial difficulties of testing students who lacked reading skills, no kindergartners were tested at baseline.[11]

To follow the original research plan and use the highest quality data, Howell and Peterson (2002) examined voucher impacts on students for whom benchmark test score data were available. For African American students with available baseline test scores (the Available Tests at Baseline, or the ATBs), one observes moderately large impacts of attending a private school on the combined math and reading portions of the Iowa Test of Basic Skills.[12] Effects are 6.1, 4.2, and 8.4 percentile points in Years 1, 2, and 3—all of which are statistically significant (see Table 2, row 1).[13] The estimated impacts of private school attendance on test scores remain significantly positive when students without baseline test scores (No Available Tests at Baseline, or NATBs) 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 National Percentile Ranking (NPR) points in Years 1, 2, and 3, respectively.

The differences in results for the ATBs and the NATBs are sufficiently striking to raise questions about the credibility of the data for the latter group. Consider the following thought experiment: two randomized experiments are conducted, one for a larger number of cases with baseline test scores and 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? If the experiments were of equal quality in other respects, we doubt any scientist would give greater credence to the study lacking such crucial baseline information.

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 1			*Year 2*			*Year 3*		
	Impact	SE	N	*Impact*	SE	N	*Impact*	SE	N
Baseline scores in model									
1. Students with baseline scores (ATBs, Grades 1-4)[a]	6.13**	(1.74)	622	4.16*	(2.22)	497	8.43**	(2.86)	519
No baseline scores in model									
2. Students with baseline scores (ATBs, Grades 1-4)	5.67**	(2.32)	622	4.36*	(2.41)	497	8.40**	(3.32)	519
3. Students with and without baseline scores (ATBs & NATBs, Grades 1-4)	5.65**	(2.32)	695	4.24*	(2.33)	562	7.49**	(3.21)	577
4. Students with and without baseline scores (ATBs & NATBs, Grades K-4)	4.61**	(2.07)	882	3.24†	(2.24)	722	4.88†	(3.02)	734
Hybrid model: Baseline scores when possible									
5. Students with and without baseline scores (ATBs & NATBs, Grades 1-4)	6.28**	(1.86)	695	3.94*	(2.21)	562	7.75**	(2.81)	577
6. Students with and without baseline scores (ATBs & NATBs, Grades K-4)	5.15**	(1.71)	882	3.21†	(2.07)	722	5.31**	(2.70)	734

NOTE: The impacts of private school attendance on test scores are reported. Weighted, two-stage least squares regressions are estimated; treatment status was used as an instrument. Bootstrap standard errors (10,000 reps completed) that are robust to intrafamily correlations are reported in parentheses. Available Tests at Baseline (ATBs) consist of students for whom baseline test scores are available; No Available Tests at Baseline (NATBs) consist of students for whom no baseline test scores are available. The 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.

a. Estimates here differ slightly from those originally reported because MPR, after issuing a final report, revised their weights and strata in 2003 in response to recommendations by Alan Krueger.

†$p < .10$, one-tailed. *$p < .10$, two-tailed. **$p < .05$, two-tailed.

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. 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. KZ argue otherwise, saying that "because of random assignment . . . estimates are unbiased" (p. 660). But estimates are unbiased only if the randomization process worked as well for the NATBs as it did for the ATBs—an outcome that KZ assume but cannot show (Peterson & Howell, 2003, note 19). In the words of the statisticians who first conducted a secondary analysis of the New York experiment, KZ's "assertion that 'because assignment to treatment status was random . . . a simple comparison of means between treatments and controls without conditioning on baseline scores provides an unbiased estimate of the average treatment effect' is simply false, because there are missing outcomes" (Barnard, Frangakis, Hill, & Rubin, 2003b, p. 321).

There are a variety of attributes of the New York experiment that make the KZ claim if not false then at least exceedingly problematic. The administration of the New York experiment was quite complicated, as KZ themselves lament. Half of 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 nonwinners 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 underperforming 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 baseline scores is impossible to know.

Baseline test score imbalances among NATBs may be especially likely among those students in the experiment who were assigned to treatment and control conditions using the matched propensity design, which relied on baseline test scores whenever they were available. Among the NATBs, student assignments were made only on the basis of available demographic data, and because these data are weakly correlated with outcome test scores, they make

for fragile indicators when constructing adequate treatment and control group (Barnard et al., 2003a, p. 300).

Beyond the creation of the treatment and control groups, 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% of NATBs, as compared to 29% of ATBs.[14]

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, whereas parents may 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 1, 2, and 3, raising further doubts about the reliability of the NATB data.[15]

Finally, attrition from the experiment introduces additional risks of bias, risks that lead Barnard et al. (2003a) to characterize the experiment as "broken."[16] When baseline scores are not available, one simply does not know whether this attrition compromised the baseline test score balance between the two groups. For all of these reasons, estimates are best made for students for whom baseline test scores are available.

For the moment, though, let us set aside the possibilities of bias arising due to problems encountered during sample construction, administrative error, or differential attrition. What, exactly, is to be gained from introducing the NATBs to the analysis? KZ 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 through 4 posted significant and positive test score gains (regardless of whether one includes the NATBs in the analysis and whether controls for baseline test scores are included) in all 3 years. Impacts for kindergartners, meanwhile, were more erratic, bottoming out at –13.9 in Year 3.[17]

At first glance, however, KZ 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. Among African American ATBs, the standard errors for impacts in Years 1, 2, and 3 in test score models that do not include baseline test scores are 2.3, 2.4, and 3.3 (see Table 2, row 2).[18] When controls for baseline test scores are added, the standard errors drop noticeably to 1.7, 2.2, and 2.9 for the 3 years (Table 2, row 1). When expanding the sample to include both ATBs and NATBs and dropping controls for baseline test scores, the standard errors jump up to 2.1, 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.

When including students without baseline scores, KZ (2003) report only the imprecise model.[19] By contrast, the initial secondary analysis (Barnard et al., 2003a) included baseline test scores, whenever possible, to obtain as precise an estimate as possible. In a series of estimations, KZ (2004) follow suit and control for baseline scores, although without estimating the model in a transparent manner that allows for straightforward comparisons with the impacts originally reported. Instead, the hybrid model is estimated only after recoding the ethnic identity of some students and adding numerous other demographic controls and missing-data indicators (on these issues, see below). When one does estimate 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). To generate findings that justify their conclusion that vouchers had insignificant effects on African American students, KZ cannot simply add students without baseline scores to the estimations. Instead, 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. At baseline (and, again, at the Year 2 and Year 3 follow-up sessions), accompanying adults were asked to place the student's mother and, separately, the student's father into one of the following groups: (a) Black/African American (non-Hispanic), (b) White (non-Hispanic), (c) Puerto Rican, (d) Dominican, (e) other Hispanic (Cuban, Mexican, Chicano, or other Latin American), (f) American Indian or Alaskan Native, (g) Chinese, (h) other Asian or Pacific Islander (Japanese, Korean, Filipino, Vietnamese, Cambodian, Indian/Pakistani, or other Asian), and (i) other (write in: _____).

Students of other background. In most instances, one can easily infer each student's ethnicity simply based on the ethnicity of the parents, as indicated by the responses to this question on the survey. For some 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 on 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 was left in that category.

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.[20] Rather, these parents identified themselves, quite intentionally, as Black-Haitian, Puerto Rican/Black, Black–West Indies, Black–Cuban American, and Black/Jamaica. 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."

KZ (2003, p. 317, Table 2) nonetheless reclassify parents of those in the "other" category as "Black, non-Hispanic" even when the respondents themselves have rejected that label.[21] But it is misleading—and contrary to the very federal guidelines that KZ use to bolster their case—to classify as "Black, non-Hispanic" people who openly identify themselves as Hispanic, Dominican, or West Indian.

According to the Office of Management and Budget (OMB) guidelines KZ cite, 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," whereas "a person is 'Black'" if she is from "any of the Black racial groups of Africa." The guidelines go 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 . . . that uses more detail shall be organized in such a way that the additional categories can be aggregated into these basic racial/ethnic categories."[22]

To defend their classification of some Hispanics as "Black non-Hispanic," KZ (2004) cite studies that indicate that "society treats individuals with different skin tones differently" (p. 686), 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 KZ (2004) article.[23] 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.[24] 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."

KZ (2002) 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, which were not significant when these estimated Black Dominicans were included in the model. Actual results from these models were dropped in KZ (2003) and KZ (2004), but the basic idea of reclassifying Hispanic students as Black, non-Hispanic persists (see, e.g., KZ, 2004). We are unaware of scholarly precedents for this classification system.

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 also might 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, KZ employ a unique classification scheme. They identify students of mixed heritage as African American, as long as either the mother or the father is African American. If a child has a mother who is Hispanic but a father who is African American, KZ classify the child as "Black, non-Hispanic."[25] As a consequence, students cannot be classified as Hispanic (while maintaining mutually exclusive categories) unless neither parent is African American. KZ defend this classification scheme on the grounds that it is "symmetrical." But symmetry is hardly the word for a scheme that classifies Hispanics and African Americans according to different principles.

Howell and Peterson (2002) classify all students according to a single principle—students consistently were assigned to their mother's ethnic identification, a procedure also used by Barnard et al. (2003a, p. 305). Because it is a child's mother who strongly influences the educational outcomes of most low-income, inner-city 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% lived with their mother only, as compared to just 2% who lived only with their father.[26] The mothers of 74% of these students were single, divorced, separated, or widowed; in fact, only 20% of the children lived in families where the mother was married. Mothers accompanied 84% of children to testing sessions, and in 94% 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. Because the educational choices available to the mother are what matter most for the child, students in this study should be classified according to her ethnicity.[27]

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, the student's ethnic background is defined by the mother's ethnic background. Second, students are identified as African American if both parents are. Third, the child's ethnicity is identified by the ethnicity of the parental caretaker (most frequently the mother but occasionally the father). In all 3 years, and for all three of these 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).

In addition, the results do not 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 and without baseline test scores, statistically significant, positive impacts on African Americans are estimated in all 3 years (see column 4, Table 3).

ISSUE 3: WHICH COVARIATES SHOULD BE INCLUDED IN THE ANALYSIS?

Using hybrid models that take into account baseline scores whenever possible, we have shown significantly positive impacts of private schooling on the test scores of all participating African American students (defined in various ways). KZ do not report these simple, transparent estimates. Instead, in KZ (2002), hybrid models include 12 other regressors (8 family and student characteristics and 4 missing variable indicators). KZ (2004) adds 16 more (8 characteristics and 8 missing data indicators).[28]

The decision to add all of these covariates obviously forsakes the values of simplicity and parsimony (see, e.g., Zellner, 1984, p. 31). Unfortunately, it also provides little gain in the precision of the estimates obtained (see below). Equally important, it increases the chances of introducing bias. First, when adding covariates, KZ impute means and include indicator variables to denote cases with missing values. In doing so, KZ must make the highly restrictive assumption that neither the background variables nor missing-value indicators correlate with treatment; if they do, then the estimated treatment effects may be biased.[29] As Achen (1986) points out, when working with less-than-perfect randomized experiments, "controlling for additional variables in a regression may worsen

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

	Mother African American			*Mother and Father African American*			*Parental Caretaker African American*[a]			*Mother or Father African American*		
	Impact	SE	N	*Impact*	SE	N	*Impact*	SE	N	*Impact*	SE	N
Students with baseline scores (ATBs)												
Year 1	6.13**	(1.72)	622	5.78**	(1.84)	587	6.18**	(1.74)	624	5.29**	(1.75)	667
Year 2	4.16*	(2.23)	497	4.13*	(2.29)	469	4.17*	(2.24)	498	3.28†	(2.18)	533
Year 3	8.43**	(2.86)	519	8.05**	(2.93)	485	8.36**	(2.87)	520	7.64**	(2.83)	554
Students with and without baseline scores (ATBs & NATBs)												
Year 1	5.15**	(1.76)	882	5.00**	(1.81)	839	5.20**	(1.74)	884	4.00**	(1.72)	946
Year 2	3.21†	(2.08)	722	3.56†	(2.14)	683	3.24†	(2.07)	723	2.66†	(2.03)	771
Year 3	5.31**	(2.71)	734	5.08*	(2.82)	687	5.27**	(2.70)	735	4.45*	(2.65)	785

NOTE: Impacts of private school attendance on test scores are reported. Weighted, two-stage least squares regressions are estimated; treatment status was used as an instrument. Bootstrap standard errors (10,000 reps completed) that are robust to intrafamily correlations are reported in parentheses. 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 was determined on the basis of baseline, Year 2, and Year 3 surveys; father's ethnicity was 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] was African American. Available Tests at Baseline (ATBs) consist of students for whom baseline test scores are available; No Available Tests at Baseline (NATBs) consist of students for whom no baseline test scores are available.

a. The mother was assumed to be the primary caretaker of the child's education except in those cases where the child lives only with the father.

†$p < .10$, one-tailed. *$p < .10$, two-tailed. **$p < .05$, two-tailed.

the estimate of the treatment effect, even when the additional variables improve the specification" (p. 27),[30] a problem KZ themselves admit:

> If there is a chance difference in a baseline characteristic between treatments and controls, there also could be an erroneous correlation (due to chance or misspecification) between the baseline characteristic and the outcome variable that would sway the estimated treatment effect if covariates are included.[31] (p. 696)

Given such risks, a good rule of thumb is to avoid adding a covariate unless 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 both of these tests: Baseline test scores of treatment and control groups remained balanced from baseline to the Year 3 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 KZ introduce to the analysis.

Elsewhere in their article, KZ 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. (p. 681)

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. 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 KZ, 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—compare KZ (2002) with KZ (2004)—questions naturally arise about the possibility of specification searching.

To show how results change when covariates are added, Table 4 reports 3rd-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 through 4 report estimated impacts for ATBs; columns 5 through 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.

TABLE 4: Year 3 Test Score Impacts for African Americans, Variously Defined, With and Without Baseline Test Scores (Estimates Obtained From Simple/Transparent Models and From Specification Searches)

	Students With Baseline Test Scores (ATBs)				*Students With and Without Baseline Test Scores (ATBs & NATBs)*			
	Mother AA	*Both Mother and Father AA*	*Parental Caretaker AA*[a]	*Either Mother or Father AA*	*Mother AA*	*Both Mother and Father AA*	*Parental Caretaker AA*[a]	*Either Mother or Father AA*
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Transparent model (no baseline test scores)	8.40* (3.35)	7.91** (3.44)	8.41** (3.31)	7.10** (3.24)	4.88† (3.00)	4.69† (3.16)	4.90† (3.04)	3.53 (2.96)
Transparent model (with baseline test scores)[b]	8.42** (2.82)	8.05** (2.91)	8.36** (2.84)	7.64** (2.80)	5.31** (2.70)	5.08* (2.79)	5.27* (2.72)	4.45* (2.64)
First search, controls for								
Four grade levels[c]	7.88** (2.80)	7.18** (2.89)	7.84** (2.83)	7.43** (2.76)	4.79* (2.74)	4.31† (2.77)	4.79* (2.70)	4.23† (2.61)
Plus mother's education	7.80** (2.82)	7.15** (2.86)	7.77** (2.79)	7.32** (2.73)	4.82* (2.69)	4.39† (2.74)	4.82* (2.68)	4.19† (2.59)
Plus log income	7.79** (2.80)	7.17** (2.85)	7.76** (2.83)	7.35** (2.69)	4.82* (2.75)	4.38† (2.77)	4.82* (2.73)	4.18† (2.63)
Plus student's gender	7.74** (2.85)	7.16** (2.89)	7.71** (2.83)	7.36** (2.73)	4.81* (2.78)	4.46† (2.84)	4.80* (2.79)	4.15† (2.64)
Plus employment status	7.85** (2.85)	7.23** (2.89)	7.81** (2.82)	7.79** (2.71)	4.40† (2.86)	4.44† (2.80)	4.40† (2.83)	3.88† (2.65)
2nd search, adds controls								
Welfare	7.95** (2.88)	7.36** (2.94)	7.91** (2.90)	7.87** (2.81)	4.40† (2.83)	4.52† (2.87)	4.39† (2.86)	3.84† (2.76)
Plus mother born in United States	7.80** (2.82)	7.11** (2.83)	7.76** (2.85)	7.74** (2.72)	4.20† (2.81)	4.26† (2.82)	4.19† (2.78)	3.51† (2.68)
Plus residential mobility	7.98** (2.79)	7.07** (2.77)	7.94** (2.82)	7.70** (2.68)	4.32† (2.82)	4.15† (2.78)	4.31† (2.79)	3.55† (2.66)
Plus English spoken in home	7.50** (2.77)	6.61** (2.75)	7.46** (2.79)	7.23** (2.67)	3.96† (2.74)	3.81† (2.74)	3.95† (2.73)	3.19 (2.64)
Plus mother is Catholic	7.23** (2.79)	6.40** (2.75)	7.19** (2.76)	7.04** (2.70)	3.82† (2.73)	3.80† (2.73)	3.81† (2.70)	3.11 (2.63)
Plus student's age	7.19** (2.77)	6.37** (2.74)	7.15** (2.76)	7.03** (2.69)	3.83† (2.75)	3.79† (2.74)	3.81† (2.75)	3.11 (2.64)
Plus student is gifted	7.10** (2.85)	6.28** (2.82)	7.06** (2.80)	6.96** (2.71)	3.62† (2.71)	3.61† (2.75)	3.60† (2.75)	3.06 (2.62)
Plus student is special ed.	7.20** (2.77)	6.39** (2.83)	7.16** (2.76)	6.90** (2.66)	3.55† (2.72)	3.66† (2.74)	3.53† (2.74)	2.91 (2.60)
N	519	485	520	554	734	687	735	785

NOTE: AA = African American, ATB = Available Tests at Baseline, NATB = No Available Tests at Baseline. Impacts of private school attendance on test scores. Weighted, two-stage least squares regressions are estimated; treatment status was used as an instrument. Bootstrap standard errors (10,000 reps completed) that are robust to intrafamily correlations are reported in parentheses. Mother's ethnicity was determined on the basis of baseline, Year 2, and Year 3 surveys; father's ethnicity was determined on the basis of baseline surveys only. When accounting for the ethnicity of both parents, if missing, mother [father] was assumed African American when father [mother] was African American. All models include as covariates private school status and revised lottery indicators. Covariates were then added cumulatively so that the 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% of cases are missing for mother's education, 7.6% for income, 3.3% for gender, 2.6% for employment status, 10.9% for welfare, 2.1% for born in the United States, 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 KZ (2002). The last eight rows are those covariates included in KZ (2004). KZ (2002) also included marital status but it was then dropped from later analyses.

a. The mother was assumed to be the primary caretaker of the child's education except in those cases where the child lives only with the father.

b. 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.

c. Three grade-level indicator variables are included for models that only include ATBs.

†$p < .10$, one-tailed. *$p < .10$, two-tailed. **$p < .05$, two-tailed.

Inclusion of new covariates changes results only when the NATBs are added to the analysis (see columns 5-8). Even then, estimated impacts for two of the four definitions of ethnicity remain significant on a two-tailed test when the first seven background variables are included, and all estimations remain significant on a one-tailed test when one adds the covariates originally identified by KZ (2002) to be relevant. Only when still further background characteristics are introduced do the effects of private school attendance attenuate, although on a one-tailed test, estimates still are significant for every definition of African American except the novel one proposed by KZ. Unfortunately, with the addition of each new background characteristic, one after another, one repeatedly makes the restrictive assumption that none of the covariates correlate with treatment.

Because 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. Among the ATBs and NATBs combined, the inclusion of additional covariates never reduces standard errors by more than a minuscule 0.05 NPR percentile points. Indeed, the addition of these covariates actually causes standard errors to increase in two of the four definitions of African American background. Far from providing a more powerful estimate, as KZ have claimed, the addition of all these variables frequently has the opposite effect.[32]

CONCLUDING OBSERVATIONS

The findings reported by Howell and Peterson (2002) are robust to a wide variety of alternative specifications and classifications. Only in very few models do the results fall short of significance at conventional levels (see Table 1). Of importance, the few models that do not yield statistically significant results are the most restrictive in that they suffer from at least two of the following difficulties: (a) Large numbers of students for whom no baseline data were available were introduced into the analysis; (b) a novel, inconsistent ethnic classification scheme was employed; and (c) 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. In our view, there is no basis for privileging these estimations 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, mutually exclusive categories should be employed and equivalent coding rules that follow standard practice should apply to students of different ethnic backgrounds; and
4. in randomized field trials, covariates should only be added when 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.

NOTES

1. See Howell and Peterson (2004 [this issue]), Note 1.
2. Prior research suggests a one-tailed test; see Howell and Peterson (2004), p. 7.
3. See Notes 3 and 4 of Peterson and Howell (2003).
4. Howell and Peterson (2002).
5. Despite the differences in sample composition and methodological approach, its findings resemble those that we have reported; see Peterson and Howell (2003).
6. If not otherwise identified, all references in this article are to Krueger and Zhu (2004).
7. A preference for safe estimates implicitly favors the status quo. In Krueger's view, "policymakers should be risk-averse when it comes to changing public school systems" (as quoted in Neal, 2003).
8. See Howell and Peterson (2004), Note 16. The sample standard deviation for African Americans is 18.1 NPR points.
9. See Howell and Peterson (2004), Note 5.
10. See Howell and Peterson (2004), Note 21.
11. See Howell and Peterson (2004), Note 21.
12. See Howell and Peterson (2004), Note 13.
13. See Howell and Peterson (2004), Note 22.
14. The difference is statistically significant at $p < .01$. These percentages refer to missing information on one of the 16 demographic variables that KZ introduce (see note 29 below).
15. See Note 26 in Peterson and Howell (2003).
16. For our discussion of this issue, see Howell and Peterson (2004).
17. KZ 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 3rd year of the experiment . . . although there we find some tendency for older students to have a larger treatment effect when kindergarten students are included"(p. 682). Impacts for kindergartners are negative in all 3 years: –0.7, –2.1, and –13.9 National Percentile Ranking (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 are significant in Years 1 and 3. Kindergartners may differ from the other cohorts or, as discussed elsewhere, the data on kindergartners may be invalid.

18. Because bootstrapped standard errors can vary from iteration to iteration, estimates presented in the tables of this article may differ slightly.

19. Earlier, Hill, Rubin and Thomas (2002) stated that such inclusion would be important in any outcome analysis: "The high correlation commonly seen between pre-and posttest scores makes this variable a prime candidate for covariance adjustments within a linear model to take care of the remaining differences between groups" (p. 171).

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

21. Although the key finding under discussion is whether vouchers affect the performance of African American students as distinct from others, KZ (2004) do not consistently employ a mutually exclusive classification scheme. They say, "[our analysis] treats race and Hispanic origin as mutually exclusive unless such a response was written in" (p. 686). When looking at the public schools from which Hispanic and African American students came, meanwhile, KZ (2004) do treat African Americans and Hispanics as mutually exclusive. Substantively, the problem with this particular analysis is that schools are not necessarily poor or excellent, in fixed or absolute terms, but may be appropriate or inappropriate for specific students. For further discussion, see Peterson and Howell (2003).

22. Edmonston, Goldstein, and Lott (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. The U.S. Census does not always use the combined format. When reporting results only by race, the census includes all those who say their race is Black regardless of their nationality, Hispanic or otherwise. But when reporting results within a combined table, it classifies as Hispanic all those who identify themselves as such, regardless of their response to a separate question on race. Whites and Blacks are then identified as White, non-Hispanic, and Black, non-Hispanic.

23. National Press Club, Washington, DC, April 1, 2003.

24. Myrdal (1964) explains why 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. Also, see Howell and Peterson (2004).

25. KZ (2003, p. 317, Table 2). For further discussion of the discrepancy, see Note 21 above.

26. Results are similar when Available Tests at Baseline (ATB) and No Available Tests at Baseline (NATB) students are considered together.

27. Because fathers were often not present in the household, their demographic information was missing in many cases, providing further reason for classifying according to mother's ethnicity.

28. There are no missing cases for the four grade-level indicators.

29. In all, 32% of observations had at least one missing value on the additional covariates KZ introduce to the analysis.

30. For the ATBs, such concerns are alleviated because we know that the baseline test scores of treatment and control groups are balanced.

32. The last sentence of this quote is incorrect. The possibility of swaying the estimated treatment effect is not due to chance correlations between the baseline characteristic and the outcome variable but rather between the baseline characteristic and treatment status.

33. 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 KZ's point that additional covariates may artificially "sway the estimated treatment effect" (p. 696), just as it reinforces concerns about specification searching.

REFERENCES

Achen, C. (1986). *The statistical analysis of quasi-experiments*. Berkeley: University of California Press.

Barnard, J., Frangakis, C. E., Hill, J. L., & Rubin, D. B. (2003a, June). Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City. *Journal of the American Statistical Association, 98*, 299-311.

Barnard, J., Frangakis, C. E., Hill, J. L., & Rubin, D. B. (2003b, June). Rejoinder. *Journal of the American Statistical Association*, *98*, 320-323.

Edmonston, B., Goldstein, J., & Lott, J. T. (Eds.) (1996). *Spotlight on heterogeneity: The federal standards for racial and ethnic classification, summary of a workshop*. Washington, DC: National Academy Press.

Hill, J., Rubin, D., & Thomas, N. (2002). The design of the New York school choice scholarship program evaluation. In L. Bickman (Ed.), *Donald Campbell's legacy*. Thousand Oaks, CA: Sage.

Howell, W. G., & Peterson, P. E. (2004). Uses of theory in randomized field trials: Lessons from school voucher research on disaggregation, missing data, and the generalizability of findings. *American Behavioral Scientist*, *47*, 634-657.

Howell, W. G., & Peterson, P. E. (with Wolf, P. J., & Campbell, D. E.). (2002). *The education gap: Vouchers and urban schools*. Washington, DC: Brookings Institution.

Howell, W. G., Wolf, P., Campbell, D., & Peterson, P. E. (2002). School vouchers and academic performance: Results from three randomized field trials. *Journal of Policy Analysis and Management*, *21*(2), 191-218.

Krueger, A. B., & Zhu, P. (2002, August 16). *Another look at the New York City school voucher experiment*. Paper prepared for the Conference on Randomized Experimentation in the Social Sciences, Yale University.

Krueger, A. B., & Zhu, P. (2003, June). Comment [on Barnard, Frangakis, Hill and Rubin, 2003]. *Journal of the American Statistical Association, 98*, 314-318.

Krueger, A., & Zhu, P. (2004). Another look at the New York City school voucher experiment. *American Behavioral Scientist, 47,* 658-698.

Myrdal, G. (1964). *An American dilemma*. New York: McGraw-Hill.

Neal, D. (2003, Winter). Investment planning. *Education Next*, *3*, 85.

Peterson, P. E., & Howell, W. G. (2003). *Efficiency, bias and classification schemes: Estimating private-school impacts on test scores in the New York City voucher experiments* (Occasional Paper, Program on Education Policy and Governance, Kennedy School of Government, Harvard University, 2003). Available at www.ksg.harvard.edu/pepg/

Peterson, P. E., Howell, W. G., Wolf, P. J., & Campbell, D. E. (2003). School vouchers: Results from randomized experiments. In C. M. Hoxby (Ed.), *The economics of school choice*. Chicago: University of Chicago Press.

Zellner, A. (1984). *Basic issues in econometrics*. Chicago: University of Chicago Press.

***PAUL E. PETERSON**, Henry Lee Shattuck Professor of Government at Harvard University, is the director of the Program on Education Policy and Governance and the editor-in-chief of* Education Next*. With Dr. Howell, he is a principal author of* The Education Gap: Vouchers and Urban School *(2002, Brookings Institution Press).*

***WILLIAM G. HOWELL**, assistant professor of government at Harvard University, is the author of* Power Without Persuasion: The Politics of Direct Presidential Action *(2003, Princeton University Press). With Dr. Peterson, he is a principal author of* The Education Gap: Vouchers and Urban School *(2002, Brookings Institution Press).*